

大規模言語・視覚モデルを用いたロボティクス基盤モデル

LLM • LVM for Robotics Foundation Models

牧原 昂志*, 山田 亮佑**, 花井 亮*, 中條 亨一*, 板寺 駿輝*,
元田 智大*, 篠田 理沙**, 中原 龍一#, 片岡 裕雄**, 堂前 幸康*

産総研 {オートメーション*, コンピュータビジョン**} 研究チーム
岡山大学#

cvpaper.challenge: <http://xpaperchallenge.org/cv>
robotpaper.challenge: <https://sites.google.com/view/robotpaperchallenge>

本資料について

■ {cv/robotics}paper.challenge メタサーベイの一環

- 2023/03/23～2023/06/09に実施した調査
- 論文調査はもちろん, 背景・メタ知識も含め収集
- 今回は「ロボティクス基盤モデル」が主題
- さらにLLM/LVMをいかにロボット適用できるかを検討

産総研 Robotics x Computer Vision 研究者が大型連携



本資料について

■ 主な参考資料

- Transformer メタサーベイ
 - <https://www.slideshare.net/cvpaperchallenge/transformer-247407256>
- 基盤モデル メタサーベイ
 - <https://www.slideshare.net/cvpaperchallenge/foundation-models>
- NVIDIA GTC 2022 「2022年 コンピュータビジョン分野のトレンド」
 - <https://hirokatsukataoka.net/temp/presen/220324GTC2022.pdf>
- 講義資料
 - [https://hirokatsukataoka.net/temp/presen/230119Juntendo_ImageRecognition\(2\).pdf](https://hirokatsukataoka.net/temp/presen/230119Juntendo_ImageRecognition(2).pdf)
- 基盤モデルとロボティクス
 - <https://speakerdeck.com/mertcooking/foundation-model-and-robotics-ii-pan-moderutoroboteikusu?slide=17>



本資料の構成

■ イントロダクション

- 基盤モデルまでの深層学習の歴史について簡単に振り返り

■ 論文紹介(N+論文)

- CV/NLP/Roboticsの分野の論文を選出して紹介

■ メタサーベイ

- 論文サマリーからメタな知識を収集

■ 著者紹介・研究協力者募集



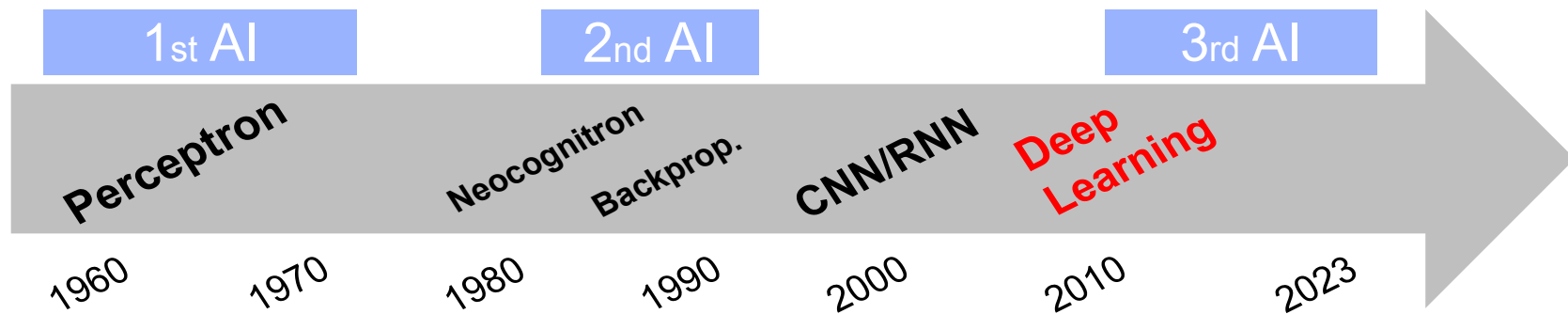
イントロダクション

深層学習の歴史

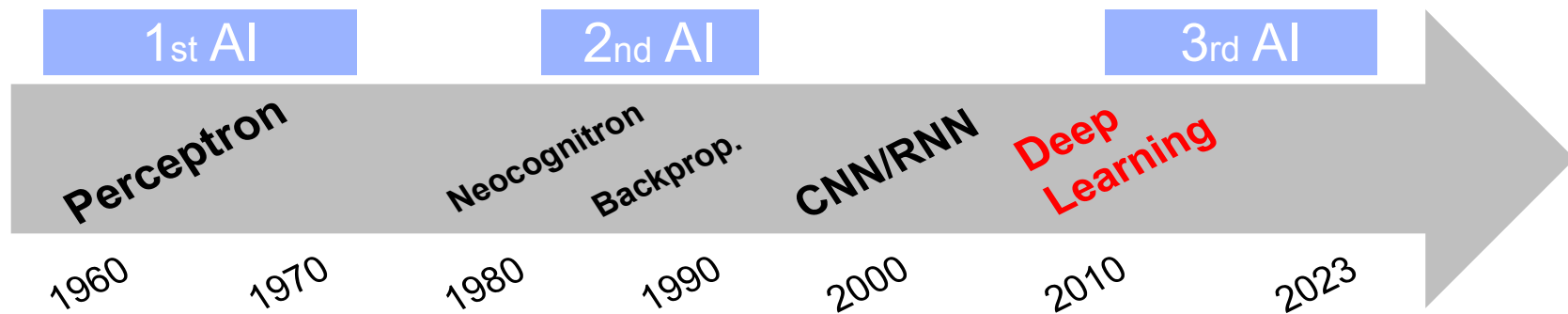
DNN時代以前の動向

- Perceptron, MLP, Neocognitron, BackProp, CNN, RNN
- DNNが流行る直前の画像認識では局所特徴(e.g., SIFT, HOG)を使用
- 深層学習 (Deep Learning)の隆盛期にある現在は第3次AIブーム

1st - 3rd AIまでの流れ



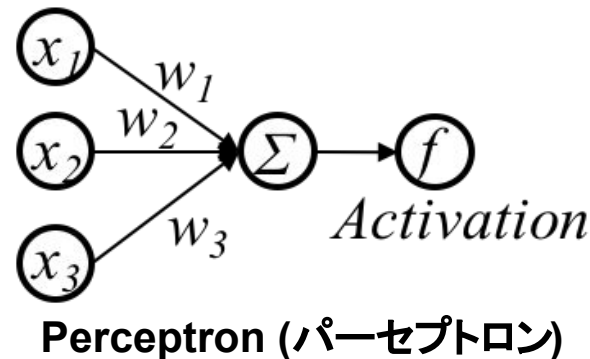
深層学習の歴史



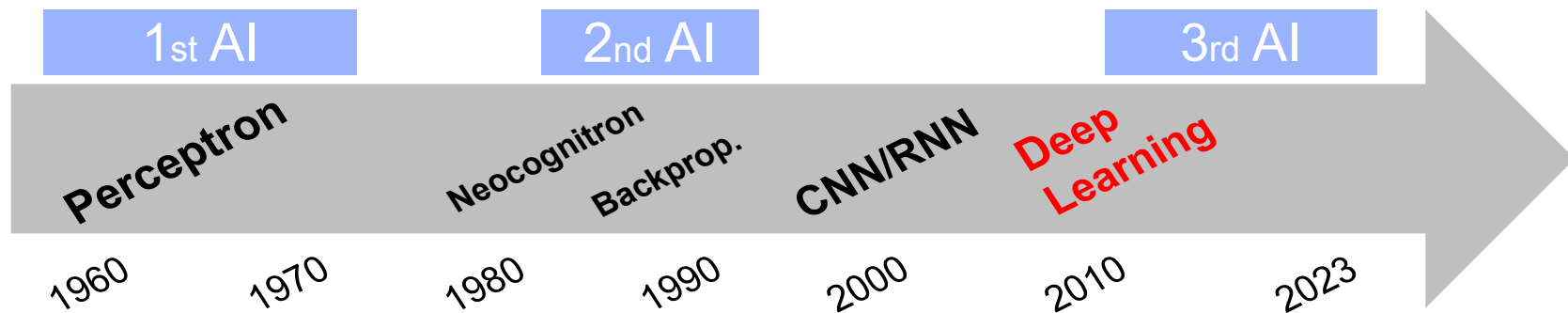
1st AIの重要技術

Perceptron

- 入力とコネクション(重み)の線形和, 活性化関数により構成
- ニューラルネットワーク構造の土台



深層学習の歴史



2nd AIの重要技術

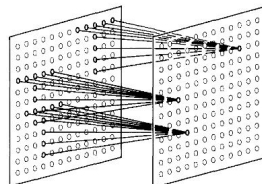
MLP: Multi-layer Perceptron

- Perceptronの多層化

Neocognitron/ConvNet

- 畳込みの概念を導入

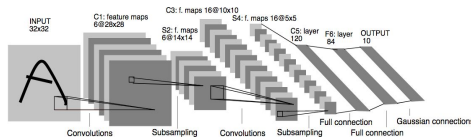
特に隣接ピクセルに類似関係のある画像処理に有効



Neocognitron

K. Fukushima, "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," Biol. Cybernetics 36, pp.193-202, 1980.

<https://www.rctn.org/bruno/public/papers/Fukushima1980.pdf>



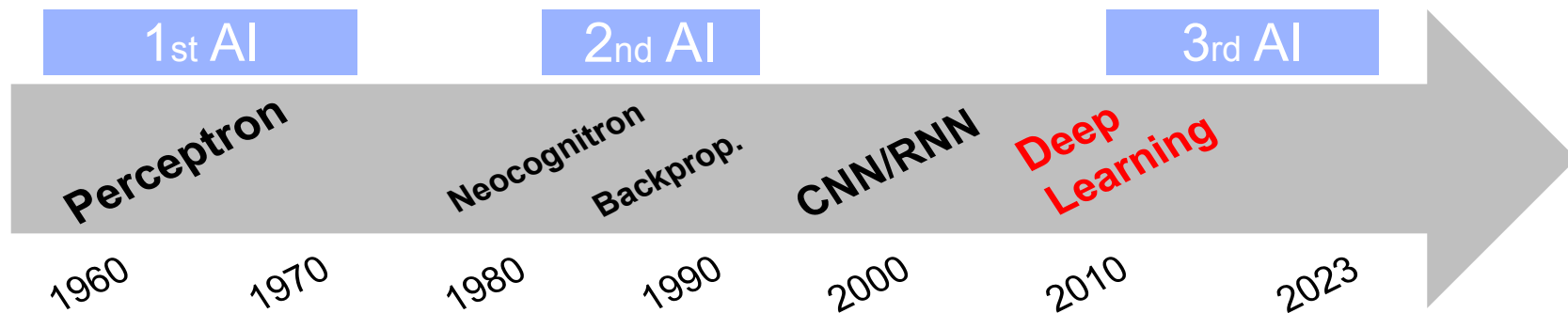
Convolutional Neural Net

Y. LeCun et al. "Gradient-Based Learning Applied to Document Recognition," IEEE, 1998.

<http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf>



深層学習の歴史



3rd AI (特に初期) の重要技術

CV: AlexNet, ResNet

- 画像認識コンペILSVRC2012で圧勝し、深層学習ブームが再来
- 深層化/データ量/計算力により目覚ましい精度向上

NLP: Word2Vec, RNN

- 自然言語を深層学習で利用可能な表現にした

AlexNet

引用: A. Krizhevsky et al. "ImageNet Classification with Deep Convolutional Neural Networks", NIPS 2012, https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

ResNet

引用: K. He et al., "Deep Residual Learning for Image Recognition", CVPR 2016, <https://arxiv.org/pdf/1512.03385.pdf>

Word2Vec

引用: T. Mikolov et al, "Efficient Estimation of Word Representations in Vector Space", ICLR 2013, <https://arxiv.org/pdf/1301.3781>

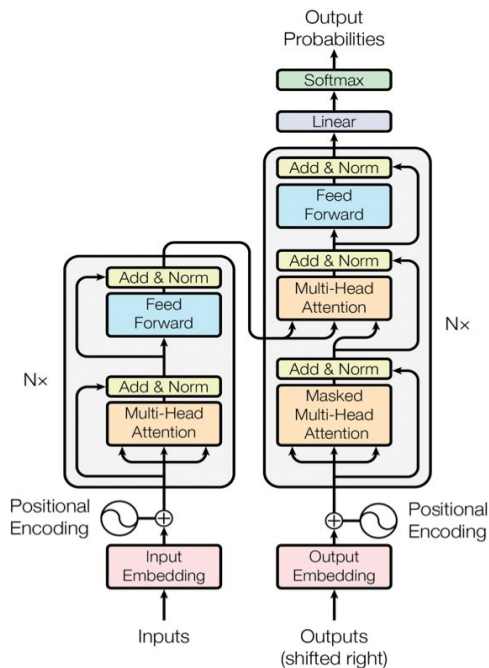
Seq2Seq (RNN)

引用: I. Sutskever et al, "Sequence to Sequence Learning with Neural Networks", NIPS 2014, [f https://arxiv.org/pdf/1409.3215.pdf](https://arxiv.org/pdf/1409.3215.pdf)



Transformerから基盤モデルへ (NLP編)

Transformerが流れを変えた！



Self-Attention(自己注意)機構により
系列データを一括同時処理可能に
学習時間短縮/性能向上を同時に実現
ブレークスルーを引き起こした
(2017年時の翻訳タスクでSoTAを達成)



急発達するHPCによるGPU/TPU並列計算技術と相性抜群
前処理でトークン化できればマルチモーダルも処理可能,
近代のほぼ全ての大規模基盤モデルは, Transformerがベース

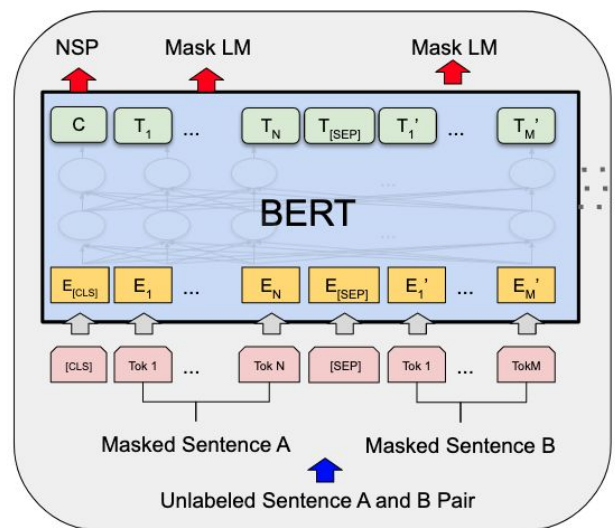
引用: A. Vaswani et al. "Attention Is All You Need", arXiv pre-print:1706.03762, 2017. <https://arxiv.org/abs/1706.03762>



Transformerから基盤モデルへ (NLP編)

BERT / GPTにより大規模化が始まる

BERTの学習方法



Pre-training

BERTでは多くのタスクを単一モデルで解くことができるが、「文章のマスクと復元(MLM)」と「連続する2つの文章が意味的に同じか異なるか推論(NSP)」により自己教師あり事前学習

MLMの例

Attention is All You Need. (元データ)

↓ 意図的に欠損作成

Attention is All ___ Need. (復元前)

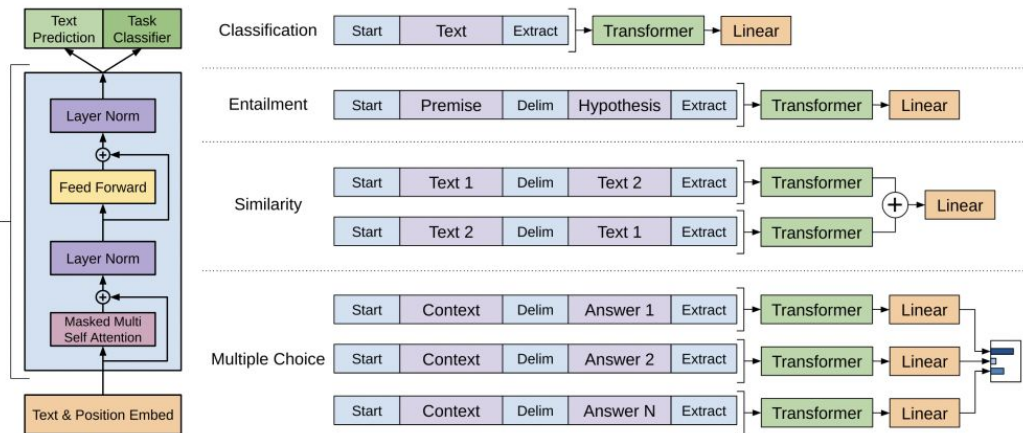
↓ BERTにより推定

Attention is All You Need. (復元後)

Transformerから基盤モデルへ (NLP編)

BERT / GPTにより大規模化が始まる

GPTの学習方法



事前学習の後に各タスクごとに
Fine-tuningを行う

タスクを意識したアーキテクチャの
構成でFine-tuningを行う

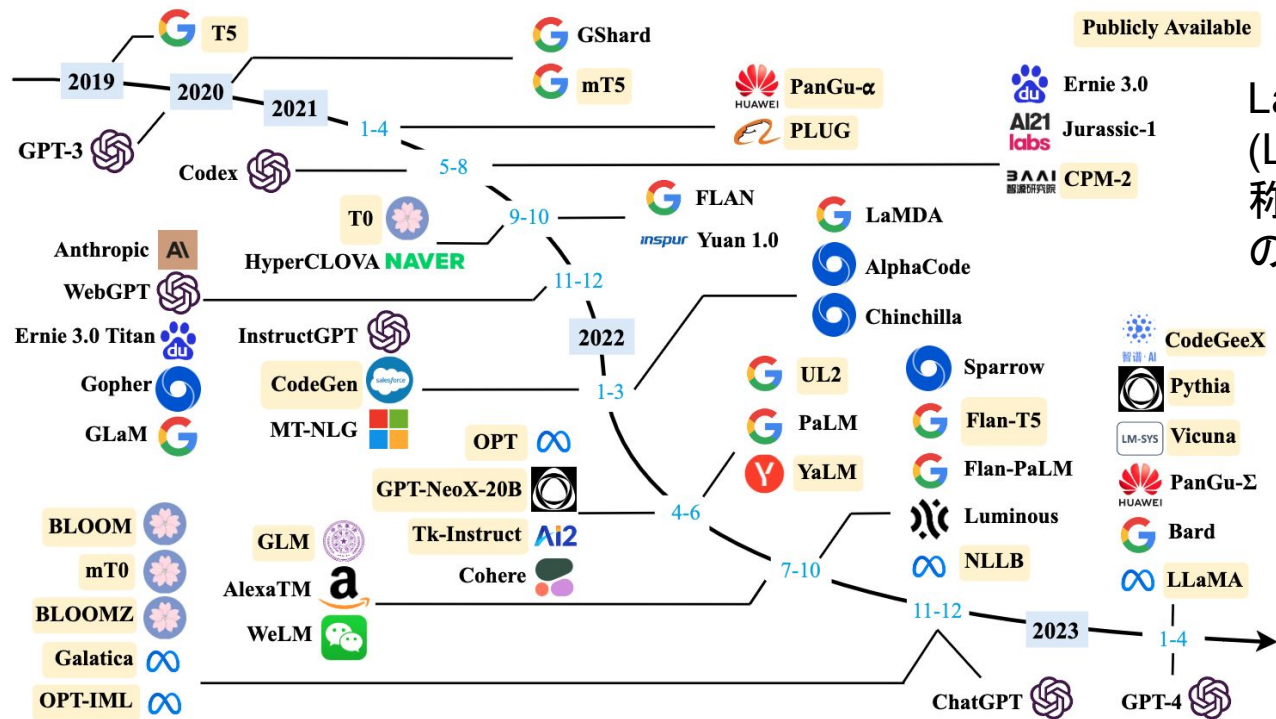
アーキテクチャの構造の変更を最小限に抑
えつつ、様々なタスクで効果的に学習でき
ている

引用: A. Redford et al. "Improving Language Understanding by Generative Pre-training", 2018.
https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf



Transformerから基盤モデルへ (NLP編)

2019年以降の大規模言語モデル

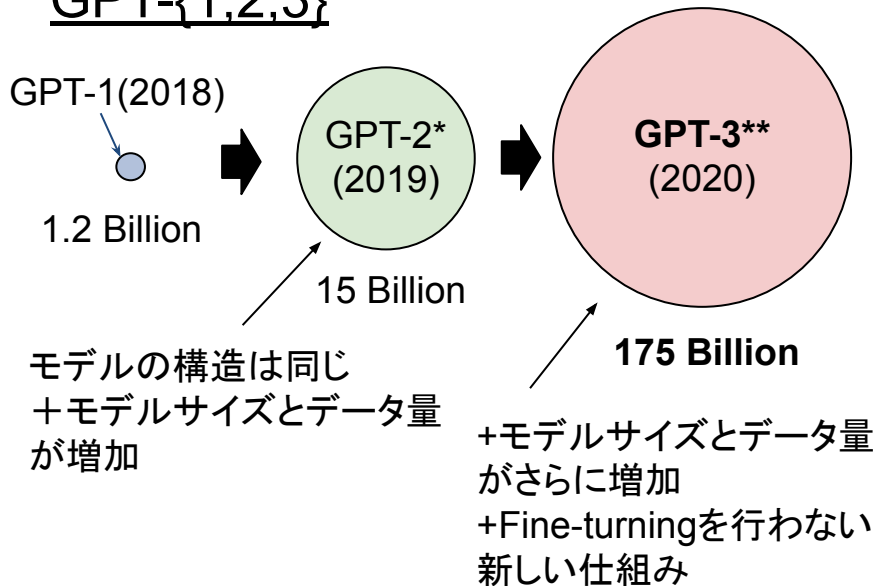


Large Language Models (LLM)は大規模言語モデルと総称, サーベイ論文では10B以上のモデルをLLMと定義

Transformerから基盤モデルへ (NLP編)

OpenAIのLLM (GPTシリーズ)

GPT-{1,2,3}



*引用: I. Solaiman et al. "Release Strategies and the Social Impacts of Language Models" (GTP-2), arXiv, 2019 <https://arxiv.org/ftp/arxiv/papers/1908/1908.09203.pdf>

**引用: Tom B. Brown et al. "Language Models are Few-Shot Learners" (GTP-3), arXiv, 2020 <https://arxiv.org/pdf/2005.14165.pdf>

段階的にモデルサイズとデータ量を増やしつつ、人間に匹敵するような文章生成を行う

Text predictionを行うモデルで, in-context learningによって目的のタスクに迅速に対応できる {Few, One, Zero}-shotにおいて, Fine-tuning時に匹敵する性能を出す

**モデルサイズとパラメータを増やせば
精度が上がっていくことがわかっている**



Transformerから基盤モデルへ (NLP編)

OpenAIのLLM (GPTシリーズ)

{Instruct*, Chat*} GPT

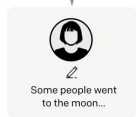
Step 1

Collect demonstration data, and train a supervised policy.

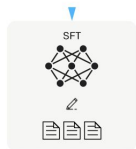
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

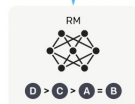
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

RLHF (Reinforcement Learning from Human feedback) で人間にとって望ましい出力を得る

Step1: Supervised Finetuningによってベースモデルを得る

Step2: 望ましい出力かどうかを判断する報酬モデルを人間が判断したデータから学習

Step3: ベースモデル, 報酬モデルを使ってPPO(強化学習手法)で学習する

真実性の向上と有害な出力生成を削減し

, ChatGPT(これをチャット形式に拡張)が波及



*引用: L. Ouyang et al. "Training language models to follow instructions with human feedback", arXiv, 2022.

<https://arxiv.org/abs/2203.02155>

**引用: C. Leiter et al. "ChatGPT: A Meta-Analysis after 2.5 Months", arXiv, 2023.

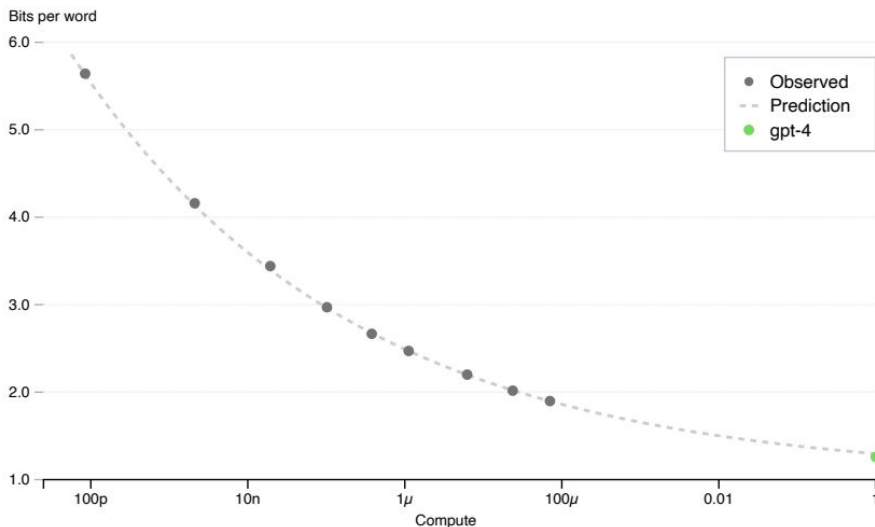
<https://arxiv.org/abs/2302.13795>

Transformerから基盤モデルへ (NLP編)

OpenAIのLLM (GPTシリーズ)

GPT-4

OpenAI codebase next word prediction



GPT-3に対してモデルとデータサイズを格段にスケールアップさせることでより正確な文章を生成できるようになった

モデルサイズを大きくしていくと最終的な損失も小さくなる(=性能が良くなる)ことを小さいモデルから検証し, GPT-4まで大きくしたときの性能の予測を正確に行った

HumanEvalデータに対しても同様の効果を確認

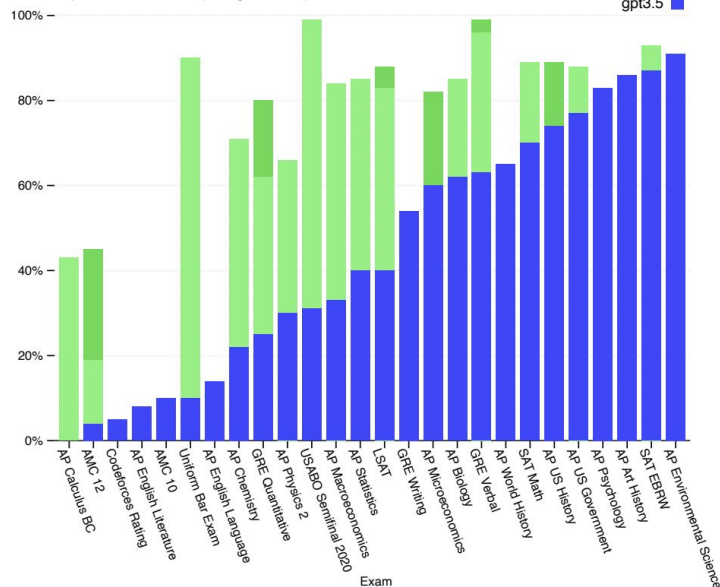
Transformerから基盤モデルへ (NLP編)

OpenAIのLLM (GPTシリーズ)

GPT-4

Exam results (ordered by GPT-3.5 performance)

Estimated percentile lower bound (among test takers)



専門的, 学術的なベンチマークで人間レベルの性能を発揮できることを確認

さまざまな分野の試験において,
GPT-3.5よりも全体的に性能が向上.

これ以外の様々なベンチマークタスクの
性能が向上していることを確認しており,
言語の違いに対しても高い性能.

Transformerから基盤モデルへ (NLP編)

GoogleのLLM

- PaLM

- Googleの540BパラメータLLM
- LLMの異なるタスクを分割・結合させスケール学習
- Wikipedia/CommonCrawl/BookCorpus/OpenWebTextから収集された大規模なデータセットや6144 TPUv4を用いて学習

引用:A. Chowdhery et al. “PaLM: Scaling Language Modeling with Pathways”, arXiv 2022, <https://arxiv.org/pdf/2204.02311>

- Bard

- Googleの対話型AI
- PaLM2をベースに構築したLLMにより実装
- PaLMよりもさらに高い性能を誇る生成AIとして機能

引用:W. X. Zhao et al. “A Survey of Large Language Models”, arXiv 2023, <https://arxiv.org/pdf/2303.18223>



Transformerから基盤モデルへ (NLP編)

METAのLLM

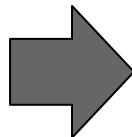
LLaMA (Alpaca, Vicuna)

トレーニングデータ量(数兆個のトークン)を増やす

7B/65Bのモデルサイズ
比較的小さなモデルで高い性能

LLaMa-13BはGPT-3(175B)に
、LLaMa-65BはChinchilla-70Bや
PaLM-5の性能に匹敵する性能

オープンソースで公開



データのリークによって、
様々な派生手法が出てきた

Alpaca: ChatGPTの生成データで
LLaMAをfine-tuning

Vicuna: ユーザとChatGPTの会話ログで
LLaMAをfine-tuning

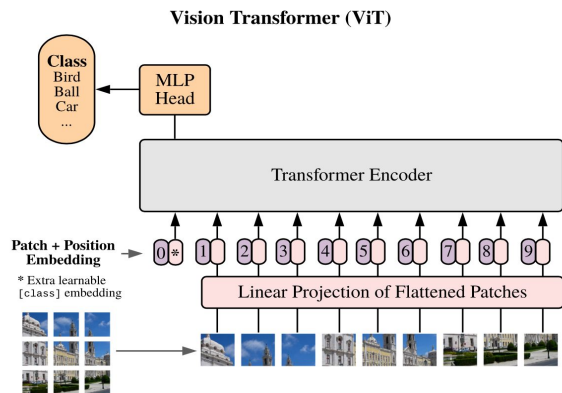
引用: R. Taori et al. "Stanford Alpaca: An Instruction-following LLaMA model",
Github, 2023. https://github.com/tatsu-lab/stanford_alpaca

引用: W.-L. Chiang et al. "Vicuna: An Open-Source Chatbot Impressing GPT-4 with
90%* ChatGPT Quality", 2023.
<https://lmsys.org/blog/2023-03-30-vicuna/>

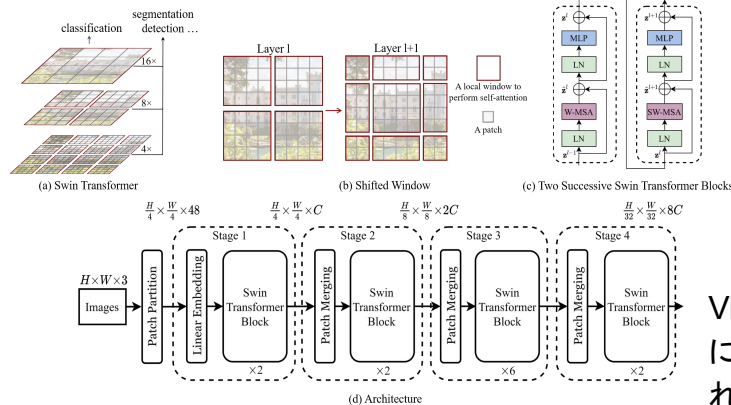
Transformerから基盤モデルへ (CV編)

Vision Transformer (ViT)の登場

- 画像分割した画像パッチを単語(トークン)と見做して入力
- ほとんどのCNNによる画像認識精度を超越
- 事前学習データの増加に伴い精度向上することを発見

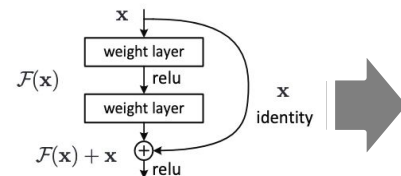


ViT [Dosovitskiy+, ICLR21]

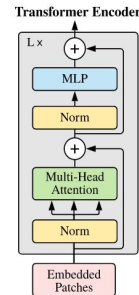


Swin Transformer [Liu+, ICCV21]

CNNからViTへ(2020年10月)



[He al. CVPR16]



Attention is all you need [Vaswani+, NIPS17]
Figure from [Dosovitskiy+, ICLR21]

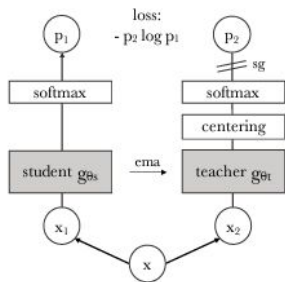
ViT / Swin Transformerは2023年現在においてもベーシックな手法として扱われている



Transformerから基盤モデルへ(CV編)

ViT向けの自己教師あり学習

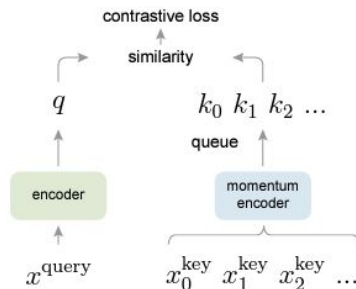
- 実画像に対して自動で一貫した教師を付与
 - DINO: ラベルなしでの自己蒸留
 - MoCoV3: 対照学習MoCoと同様だが, ViT向けに改善
 - MAE: パッチの復元(BERTのMLMを画像版タスク)



PT: DINO + ViT-B/16

FT: 78.2 @ ImageNet-1k val.

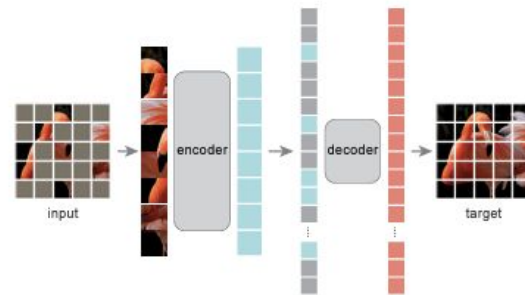
DINO [M. Caron+, ICCV2021]
<https://arxiv.org/abs/2104.14294>



PT: MoCoV3 + ViT-B/16

FT: 83.2 @ ImageNet-1k val.

MoCoV3 [X. Chen+, arXiv2021]
<https://arxiv.org/abs/2104.02057>



PT: MAE + ViT-B/16

FT: 83.6 @ ImageNet-1k val.

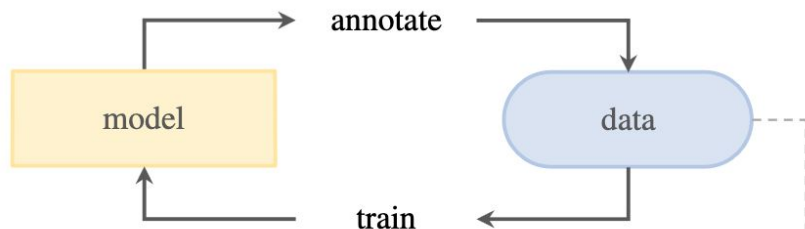
(ViT-H₄₄₈では87.8まで向上)
MAE [K.He+, CVPR2022]
<https://arxiv.org/abs/2111.06377>



Transformerから基盤モデルへ (CV編)

大規模視覚モデル: Large Vision Models (LVM)

ViT-22B (画像認識), Segment Anything (セグメンテーション)



Segment Anything 1B (SA-1B):

- 1+ billion masks
- 11 million images
- privacy respecting
- licensed images



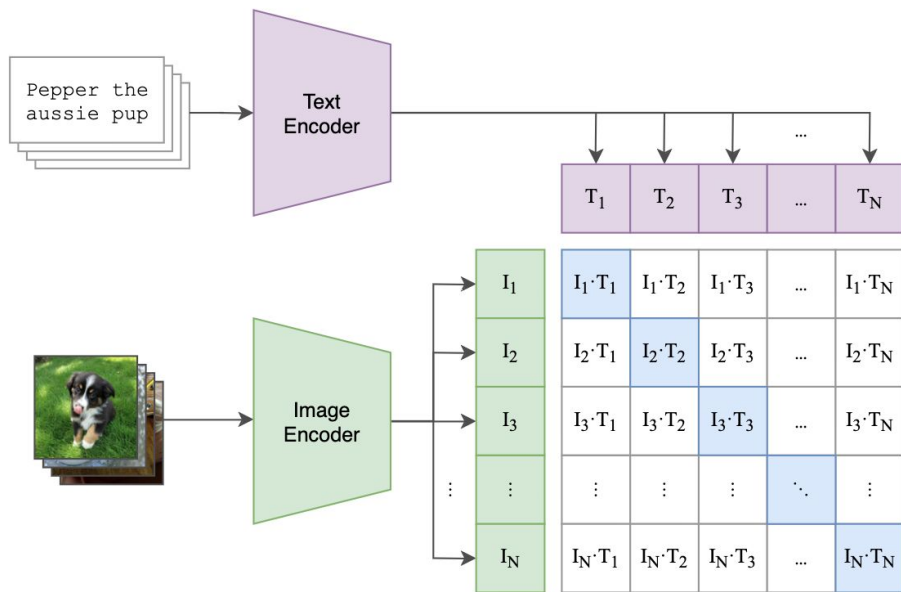
Segment Anything Model (SAM)

Data/ModelのスケールはCVでも有効

- ViT-22Bではゼロショット認識, セグメンテーション, 距離画像推定, 動画認識などで最高水準もしくはそれに準ずる精度まで到達
- Segment Anything Model (SAM) は10億以上のセグメントで学習, データ収集の半自動化手法も考案して SA-1Bデータセット構築(左図)

Transformerから基盤モデルへの流れ(V&L編)

言語と視覚の関係を紐付けるCLIPの登場



言語・視覚タスクをゼロショット認識

- 言語・視覚の特徴空間を画像・テキストでペアリングされたデータセットで対照学習(画像とテキストのエンコーダ特徴を対応づけることで学習)
- 画像に対しテキストが埋め込まれているので、ゼロショット画像認識を始め、画像キャプションやテキストからの画像生成など広く応用される

引用: A. Radford et al. "Learning Transferable Visual Models From Natural Language Supervision", arXiv pre-print:2103.00020, 2021. <https://arxiv.org/abs/2103.00020>

参考: Zero-shot Learning 網羅的サーベイ: CLIPが切り開いた Vision & Language の新しい世界 <https://techblog.exawizards.com/entry/2023/05/10/055218>

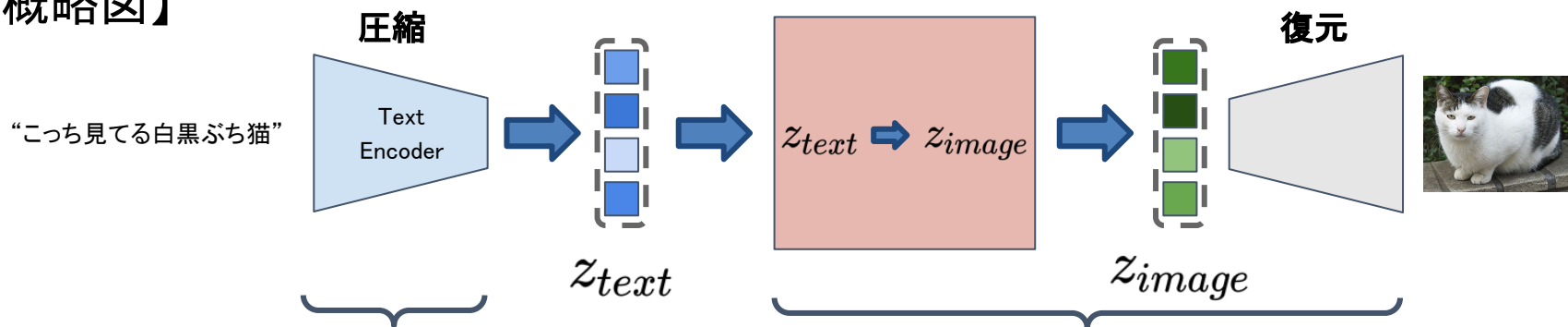


Transformerから基盤モデルへの流れ(V&L編)

テキストからの画像生成(Text-to-Image)

代表例: DALL·E1/2, Stable Diffusion, Midjourney

【概略図】



テキストから潜在表現を獲得するブロック
(Text Encoder)

テキストの潜在表現から画像を生成するブロック
(Image Generator)

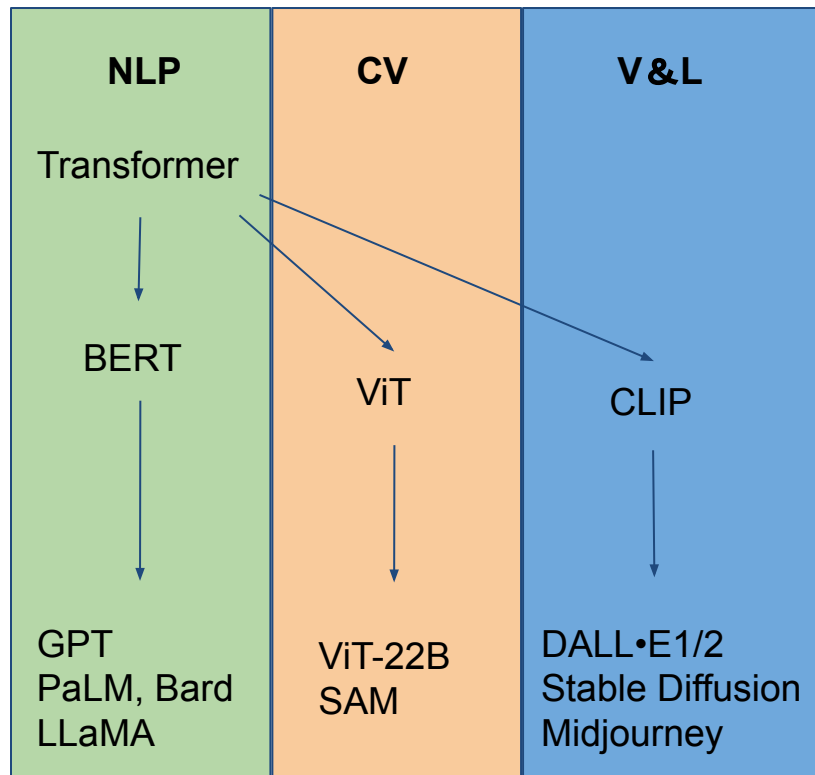
基盤モデル間で異なる主な要素は,

- Text Encoder に何を使うか (CLIPが多い)
- Image Generator に何を使うか (VAE/Diffusion Models)



ここまでのおさらい

Transformer登場以降のNLP, CV, V&L



大規模モデル時代へ突入し, Zero-shotでも高い性能を出す

Billion規模のモデルサイズ, データを扱い, 計算リソースをフル活用して精度をさらに上げる競争が激化

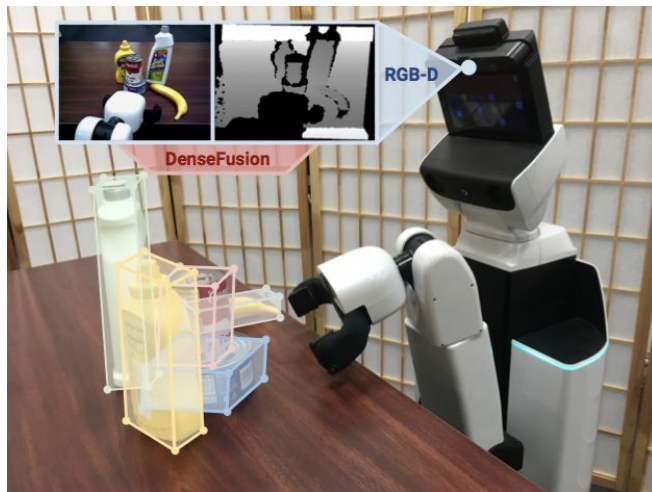
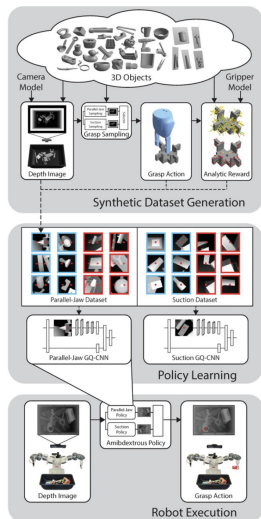
次は**実世界へのGrounding**をいかに行うか
=ロボティクスへの応用



ロボティクスの歴史: 自動運転とマニピュレーション

深層学習登場以降の動向

- 自動運転: SLAMや深層強化学習が中心
- マニピュレーション: タスク特化の学習モデルやEnd-to-Endの動作生成



本資料では, ロボティクスのタスクとして
マニピュレーションを扱うことが多い。

引用: J. Mahler et al. "Learning ambidextrous robot grasping policies," Science Robotics 2019
<https://www.science.org/doi/10.1126/scirobotics.aau4984>

引用: C. Wang et al. "DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion," in CVPR 2019
http://openaccess.thecvf.com/content_CVPR_2019/papers/Wang_DenseFusion_6D_Object_Pose_Estimation_by_Iterative_Dense_Fusion_CVPR_2019_paper.pdf



ロボット学習の経緯: End-to-End学習

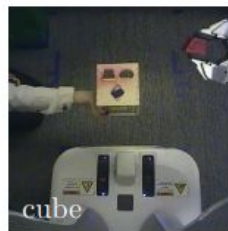
End-to-end training

- 前設計が必要だった画像特徴量を深層学習に置き換えて、画像から行動までをひとつのパスで決定可能に
- 画像特徴量をもとに、報酬に則って方策を更新する
- 実環境のオンライン学習は相当な時間・設備が必要
 - Google や U.C. Berkeley が共同で実験
 - Sim上の学習・Sim2Real にシフト

引用: S. Levine et al., "End-to-End Training of Deep Visuomotor Policies", 2016, <https://arxiv.org/pdf/1504.00702.pdf>



Arm Farm [Google, 2016]

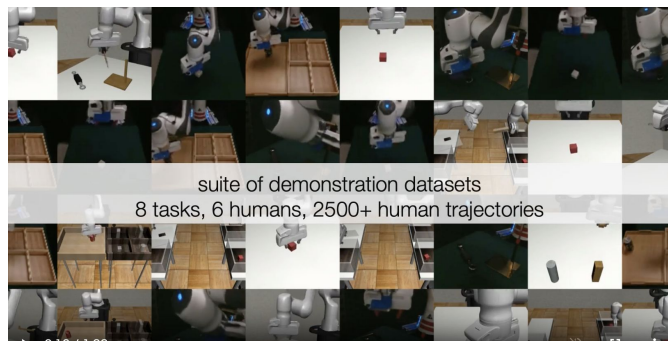


ロボット学習の経緯: 学習データ

Simulationのロボットデータ

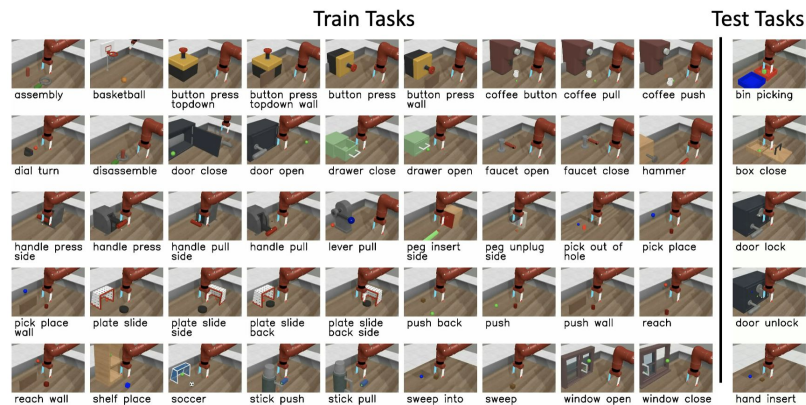
- 収集コストが低い, 安全に収集可能
- 環境の再現が容易なため学習理論の検証・ベンチマーク利用に適する
- 現実に関するリッチな情報(ダイナミクスなど)には限界がある

→ Sim2Real 転移の技術が必要



robomimic: 8タスク2500+のシミュレータ上でのhuman demonstration.

[Stanford, Univ. of Texas, '21]



MetaWorld: 強化学習のベンチマーク. 50タスクの作業環境. GATOの学習でも利用された.

[Stanford他4組織, '21]



robotpaper.challenge

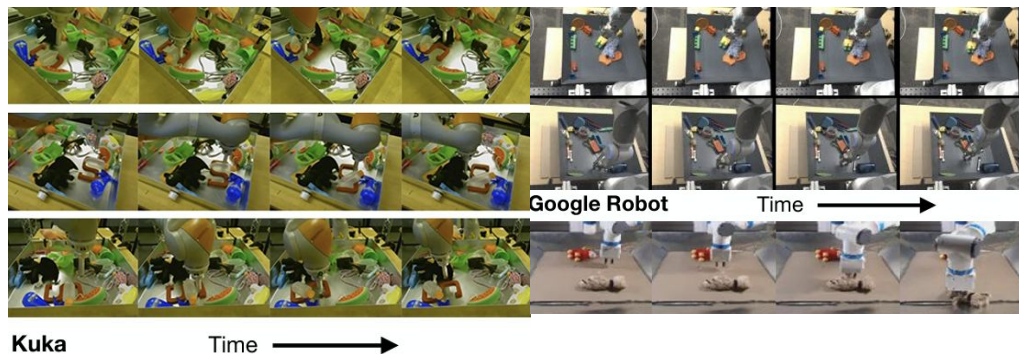


cvpaper.challenge

ロボット学習の経緯: 学習データ

Realのロボットデータ

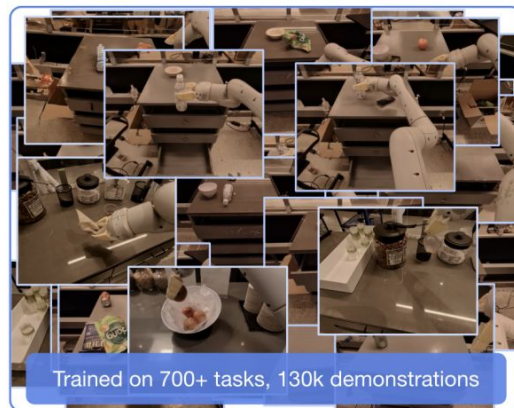
- 時間, リソースを含め収集コストが高い
- データの多様性には限りがある
- 現実に関するリッチな情報(ダイナミクスなど)が取れる



Kuka

Time →

RoboNet: 7種類のロボットを利用, 15M+フレームの多視点から撮影した作業動画
[UC Berkeley他3大学, 2020]



Trained on 700+ tasks, 130k demonstrations

RT-1: 人による操作で, 700以上のタスク130kの demonstration データを17ヶ月かけて収集
[Google, 2022]



ロボット学習の経緯: Sim2real

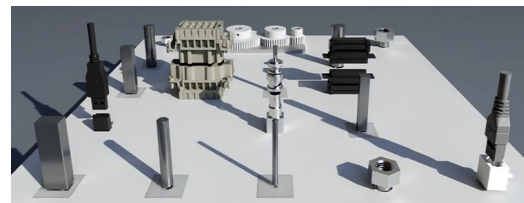
シミュレーション技術の発展

- 視覚の質向上
 - Physically based rendering
- ダイナミクスの質向上
 - 接触モデルの改良
 - 干渉判定の高精度化
 - FEMによる変形物表現
- 高速化
 - GPUを利用した高速な物理計算

- 個別タスクでは、高難度タスクの学習に利用されている



Drake [TRI, MIT]



工場における組立作業
Isaac Sim を利用 [Nvidia, RSS2022]

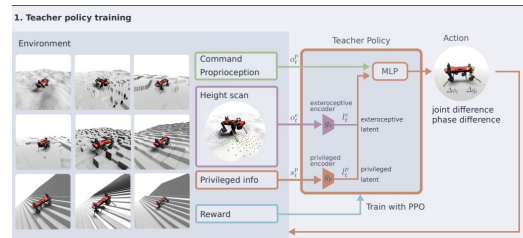
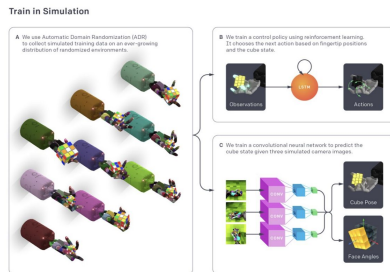
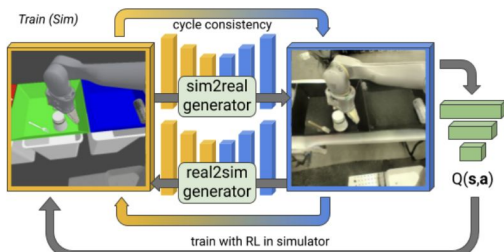
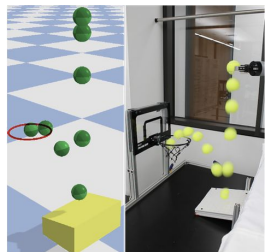
ロボット学習の経緯: 学習データ

Sim2Realの転移のアプローチ

詳細はLil'logを参照:

<https://lilianweng.github.io/posts/2019-05-05-domain-randomization/>

1. シミュレーション自体を限りなく現実近づける(システム同定)
2. 現実からシミュレーション間の変換を学習する(ドメイン適応)
3. シミュレーション内で幅広い範囲を学習する(ドメイン乱択化)
4. シミュレーションでのみ得られる情報を利用して学習を促進する



TuneNet [Allevato+, PMLR2020]

1. 現実に合うようにシミュレーションパラメータを調整

RL-CycleGAN [Rao+, CVPR2020]

2. 現実とシミュレーション間の変換を学習. 変換方向によって sim2real, real2sim と呼ばれる

[OpenAI, arXiv2019]

3. 色やテクスチャ等の視覚, 質量や摩擦等の物理, 遅延時間等のシステムのパラメータをランダムに与えデータ生成

[ETH, Science Robotics2022]

4. 摩擦や真の環境形状などシミュレーションでのみ得られる情報を使って学習したteacherの挙動を模倣するようにstudentを学習

やはり現実世界で大量にデータ収集する必要性に気づいた → ロボティクス基盤モデルの時代へ



robotpaper.challenge



cvpaper.challenge

ロボット学習の経緯: 学習モデル

主流は模倣学習とオフライン強化学習

- 模倣学習 (imitation learning, behavioral cloning)
 - 事前に集めたデータを再現するよう学習
- 強化学習 (reinforcement learning)
 - 観測した状態から報酬を大きくする行動を選ぶよう学習
- 世界モデル (world model)
 - 行動 + 抽象化した環境情報を予測するよう学習する
 - 基盤モデルではナビゲーションへの応用が見られる

参考: 東大松尾研のAdvent Calendar URL: <https://qiita.com/advent-calendar/2022/robot-ai>

>> 世界モデルに関する資料

- <https://qiita.com/MENDY/items/c70741665356cd27449c>
- <https://qiita.com/Jumpei-Arima/items/9348ab187852e200a287>



ロボット学習の経緯: 模倣学習

収集したデータから行動を再現する学習

- ユーザによる直接教示や遠隔操作で画像や言語, 運動データを集める
- 2018年ごろから事例が増え, 収集データのバリエーションと共に汎化性・頑健性が向上

Domain-Adaptive Meta-Learning
provide demonstration data

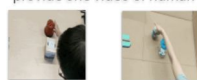


human demos robot demos

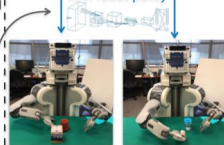
learn how to infer a policy
from one human demo

Deployment

provide one video of human



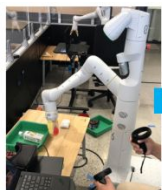
infer robot policy



One-shot imitation

[Yu+, 2018]

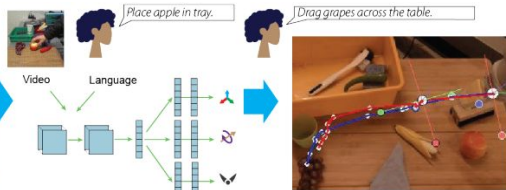
約600 ~ 1300エピソード



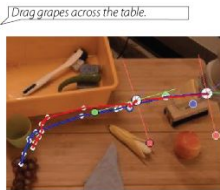
Shared Autonomy Data Collection



Diverse Multi-Task Dataset

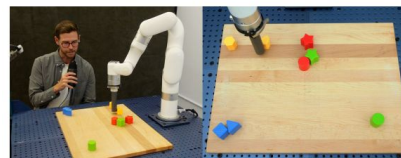


Continual Policy Training



Generalization to Unseen Tasks

a) Interactive, real-time natural language guidance

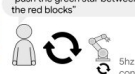


(example) human + robot goal:

make a green-eyed
smiley face



"push the green star between
the red blocks"

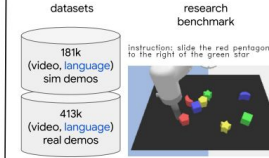


b) One robot, 87,000 diverse behaviors

- push the green star between the red blocks
- place the triangle on the top right side of the heart
- slide the blue cube down a bit
- move red star into the red circle

- nudge the red star slightly right
- separate the green circle from the red star
- put the heart in the center of the board
- point your arm at the blue triangle

c) Language-Table: data + benchmark



Interactive language [Google, 2022]

87kエピソード

ロボット学習の経緯: 強化学習

観測した状態から報酬を増やす行動を選ぶ学習

- ロボットが観測した画像から報酬(タスクの成否・障害物への衝突など)をもとに次の行動する方策を獲得する
- 2016年当初は行動決定に現在の方策を利用するon-policyな学習
 - タスクの複雑化によってon-policyな学習は次第に困難に
 - Offline dataとの併用やoff-policyな学習にシフト



Arm Farm [2016]
on-policy RL



QT-opt [2018]
offline data & on-policy RL

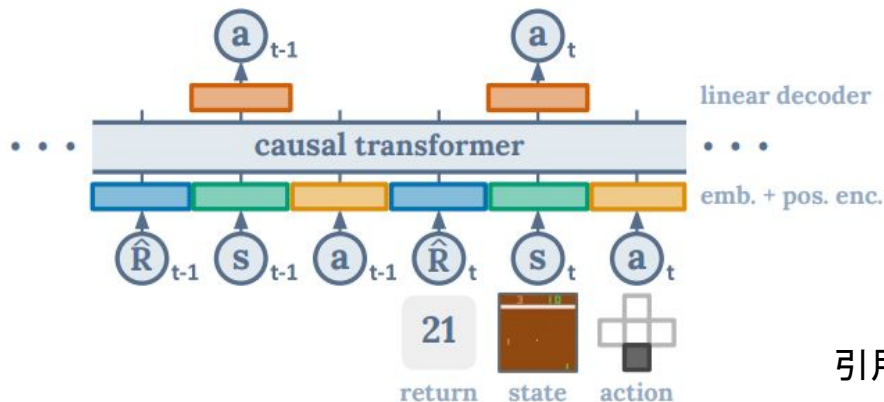


MT-opt [2021]
off-policy RL

ロボット学習の経緯: オフライン強化学習

Decision Transformer [Chen+, arXiv2021]

- Transformer を強化学習で利用するためのパラダイム
 - 状態の入力 $s(t)$ に対して行動 $a(t)$ を予測する
- 将来の報酬和 $R(t)$ で条件付け
 - 大きい報酬和 \rightarrow エキスパート / 小さい報酬和 \rightarrow ビギナー
 - 報酬和で条件付けした Behavioral cloning とも解釈可能



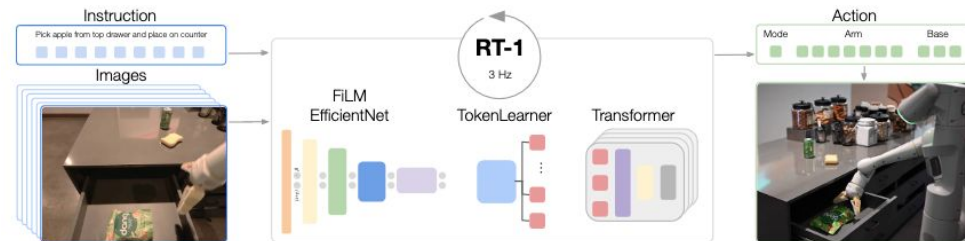
引用: <https://arxiv.org/pdf/2106.01345.pdf>



ロボティクス基盤モデル

RT-1の登場

ロボット向けの実環境の大規模データ(言語, 画像, 行動)とモデル(Transformerベース)によって様々なタスクをこなす



言語, 画像の基盤モデルが実環境への Groundingを可能にした

- 700以上のタスク処理
- 高い汎化能力
- 環境に対するロバスト性
- 多様なタスクへの適用
- 新規データへの対応 など

ロボットのためのデータとモデルが初めて大規模になった



RT-1の主要技術: 大規模データ

実環境の大量の行動データを収集

長期間・大量のデータを様々な環境やタスク設定において、人間のデモンストレーション(テレオペレーション)から収集

→ 言語指示のアノテーション/ ロボットの行動 / 視点画像, のデモデータ)



(a)

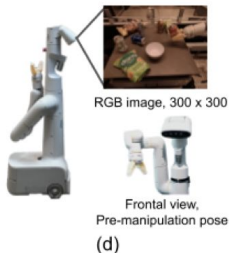


(b)



(c)

| Skill | Count | Description | Example Instruction |
|--|-------|--|--|
| Pick Object | 130 | Lift the object off the surface | pick iced tea can |
| Move Object Near Object | 337 | Move the first object near the second | move pepsi can near rxbar blueberry |
| Place Object Upright | 8 | Place an elongated object upright | place water bottle upright |
| Knock Object Over | 8 | Knock an elongated object over | knock redbull can over |
| Open Drawer | 3 | Open any of the cabinet drawers | open the top drawer |
| Close Drawer | 3 | Close any of the cabinet drawers | close the middle drawer |
| Place Object into Receptacle | 84 | Place an object into a receptacle | place brown chip bag into white bowl |
| Pick Object from Receptacle and Place on the Counter | 162 | Pick an object up from a location and then place it on the counter | pick green jalapeno chip bag from paper bowl and place on counter |
| Section 6.3 and 6.4 tasks | 9 | Skills trained for realistic, long instructions | open the large glass jar of pistachios pull napkin out of dispenser grab scoop |
| Total | 744 | | |



(d)



(e)



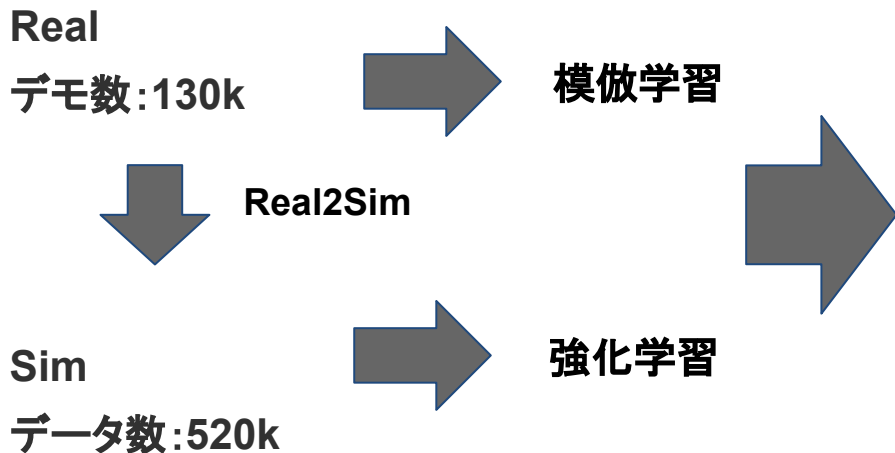
(f)

- デモ : 130K
- タスク : 744
- ロボット: 13台 (EBR)
- 収集期間: 17ヶ月



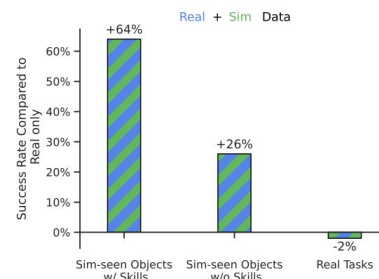
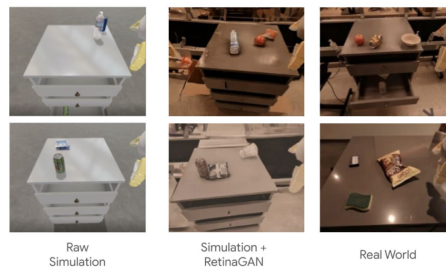
RT-1の主要技術: 大規模データ

SimulationとRealのデータを活用



これまでの学習戦略(効果的な強化学習),
Sim2Realの技術が統合されている

| Models | Training Data | Real Objects | | |
|--------|---------------|-----------------------|---|-------------------------|
| | | Seen Skill w/ Objects | Sim Objects (not seen in real) Seen Skill w/ Objects | Unseen Skill w/ Objects |
| RT-1 | Real Only | 92 | 23 | 7 |
| RT-1 | Real + Sim | 90(-2) | 87(+64) | 33(+26) |



RealのみよりもSimで経験データを拡張することで高い汎化性能を持たせる

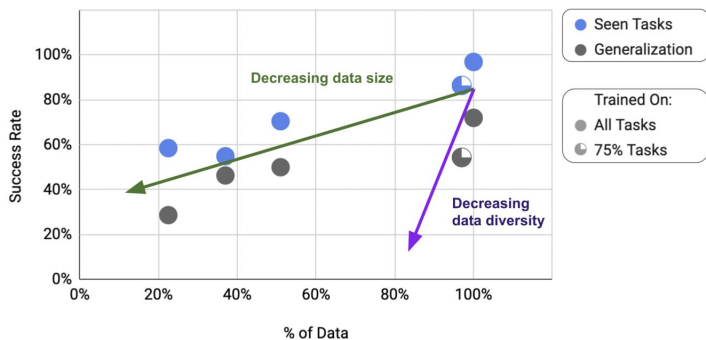


RT-1の主要技術: 大規模データ

データの量に関するスケーリング則

Transformerベースのモデルが言語, 画像に置いてデータの量が増えると性能も上がるという報告がRobotics分野においても確認

| Models | % Tasks | % Data | Seen Tasks | Generalization | | | |
|---------------|---------|--------|------------|----------------|--------------|-------------|-------------|
| | | | | All | Unseen Tasks | Distractors | Backgrounds |
| Smaller Data | | | | | | | |
| RT-1 (ours) | 100 | 100 | 97 | 73 | 76 | 83 | 59 |
| RT-1 | 100 | 51 | 71 | 50 | 52 | 39 | 59 |
| RT-1 | 100 | 37 | 55 | 46 | 57 | 35 | 47 |
| RT-1 | 100 | 22 | 59 | 29 | 14 | 31 | 41 |
| Narrower Data | | | | | | | |
| RT-1 (ours) | 100 | 100 | 97 | 73 | 76 | 83 | 59 |
| RT-1 | 75 | 97 | 86 | 54 | 67 | 42 | 53 |



実環境での動作を考えるとモデルサイズよりもデータを重視

データの量を増やすと、汎化性能も向上するが、データの多様性もかなり重要

→ データ量の49%削減と多様性(タスクの種類)の25%削減が同じくらいの性能

ロボットのデータ収集はコストが大きいですが、
闇雲に集めるだけではいいデータにはならない!



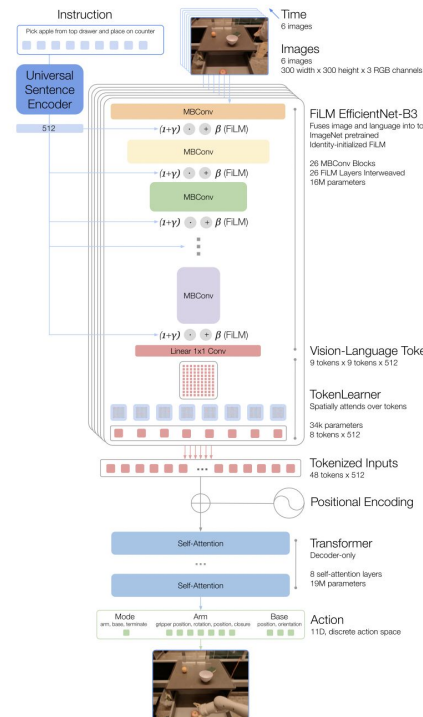
RT-1の主要技術: 学習モデル

Transformerをベースとした非自己回帰モデル

- 多様なタスクをこなせるモデルサイズ
- ロボットがスムーズに動けるくらい
速い推論の双方を満たすことが必須



- Tokenlearner
- Transformerでは自己回帰をしないことにより
言語・画像入力からの動作出力を高速化 (約3Hz)



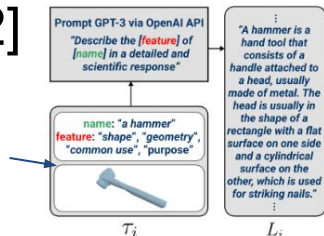
大規模言語モデルの活用: ロボティクス

ロボティクスへの展開

● ATLA [Ren+, 2022]

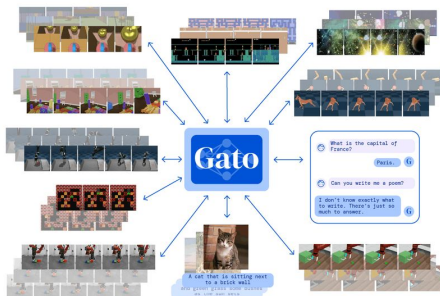
道具の形状と役割を言語から解釈
ロボットの操作に活用

例: 'ハンマー'の知識を
言語モデルがから得ている



● Gato [Reed+, 2022]

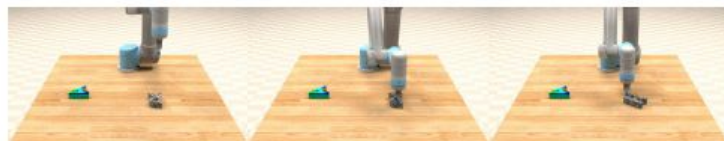
任意の入力に対応するエージェントの操作を含むLLM
→ Atari (ゲーム) ~ Pick-and-Placeを実現



あらゆるのタスクに人間レベルで対応できるモデルを目指している

● VIMA [Jiang+, 2023]

画像(イメージ)と言語(命令)を組み合わせたプロンプトマニピュレーションタスクの表現に適することを証明



テキストと図を指示に使う!

Rotate the  120 degrees.

● SayCan [Ahn+, 2022]

抽象的な指示からロボットが
実行可能な動作を選択
→ LLMが現実世界から文脈を理解

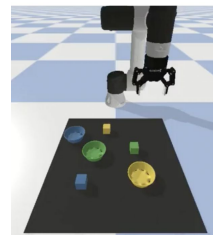
ロボットが理解・実行できるレベルに調整



Robotics at Google



Everyday Robots



Task: Move all the blocks to different corners.

SayCan:

1. Pick up the green block and place it in the lower left corner.
2. Pick up the blue block and place it in the lower right corner.
3. Pick up the yellow block and place it in the upper right corner.



robotpaper.challenge



cvpaper.challenge

大規模言語モデルの活用: LLM/LVM

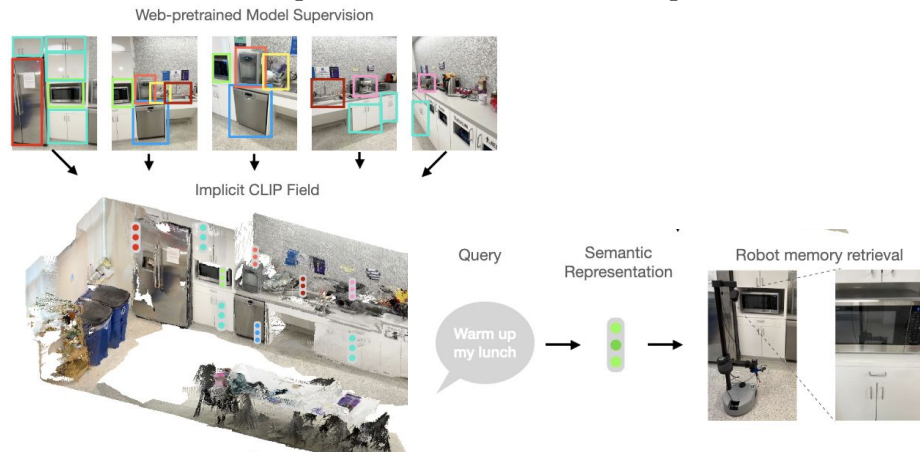
CLIPの活用

- **CLIPort** [Shridhar+, CoRL2021]



- CLIPの言語・画像対応を活用してロボットタスクのためのアフォーダンス予測
- ゼロショット認識の能力を活用してタスク特有のエンジニアリングが不要に

- **CLIP-Field** [Shafiullah+, RSS2023]

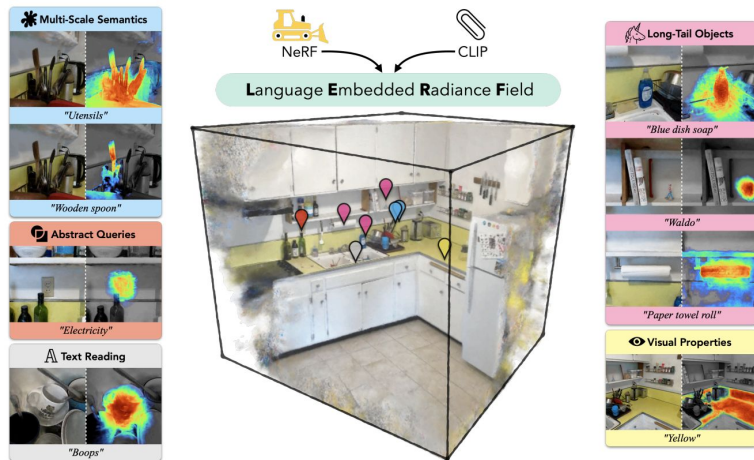


- V&Lのモダリティから3D空間のセマンティック表現を獲得

大規模言語モデルの活用: 3D (NeRF)

ロボットが人間との会話を通じて三次元世界の理解

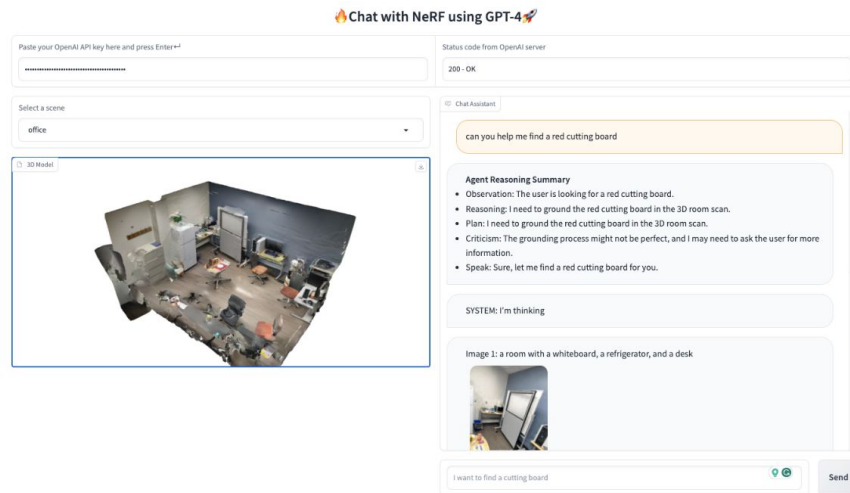
LERF: Language Embedded Radiance Fields
[Kerr+, 2023]



LLMとNeRFにより再構成されたデジタル3D世界の架け橋に
CLIPからNeRFへ言語埋め込みを実現
抽象的なクエリ:「Electricity」、視覚的なクエリ:「Yellow」、long-tail
な物体:「Waldo」、文章読み上げ:「Boops」等の幅広いクエリに対
応可能

引用:<https://www.lerf.io/>

Chat with NeRF



GPT-4/LLaVA/BLIP-2/NeRF Studio/LERFの技術を組み合わせる
ことで、3Dシーン環境/物体についての対話が可能

デモ必見!! <http://sled-whistler.eecs.umich.edu:7777/>



robotpaper.challenge



cvpaper.challenge

ロボティクスにおける重要技術と今後

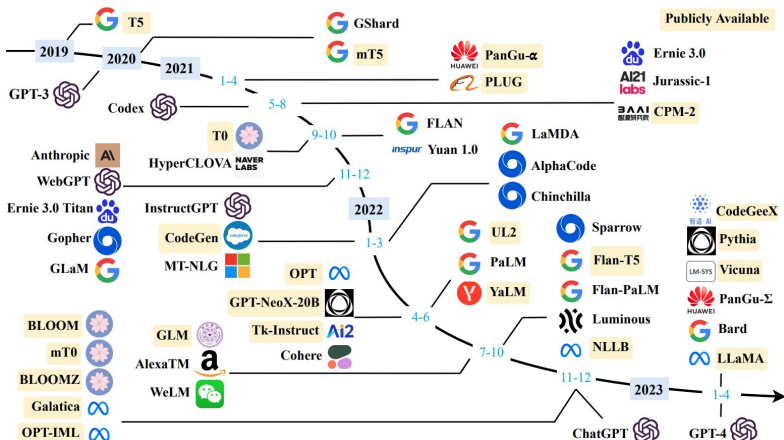
- 学習データ (Sim2Real)
 - 現状: シミュレーションデータとSim2Real or 実環境データ
 - **今後:** RT-1は実環境を重視していたが, シミュレータもまだ可能性はある(Differentiable simulationなど), 継続学習
- 継続学習
 - シミュレーションでのデータスケージング(無限に生成)
 - 比較的少量ながら現実世界で調整
- 学習モデル
 - 現状: CNNからTransformerへ
 - **今後:** クロスモーダル, 推論速度
- 学習方法 (模倣学習&強化学習)
 - 現状: **On-policy**から**Off-policy**へ,
 - **今後:** 模倣学習と強化学習は組み合わせられる
- 大規模言語モデルの活用
 - 現状: 言語と画像, 3Dの関係を利用してロボットができることと対応づける
 - **今後:** 言語をフルに活用して常識を引き出しロボットの行動を生成

論文紹介

※全ての論文を網羅していないことにご注意ください

*Transformer*から基盤モデルへ (NLP 編)

【概要】LLMのサーベイ論文



主要なLLMの流れ: T5, GPT以降の10Bパラメータ以上のモデルの発展がひと目でわかるようにまとめられている。

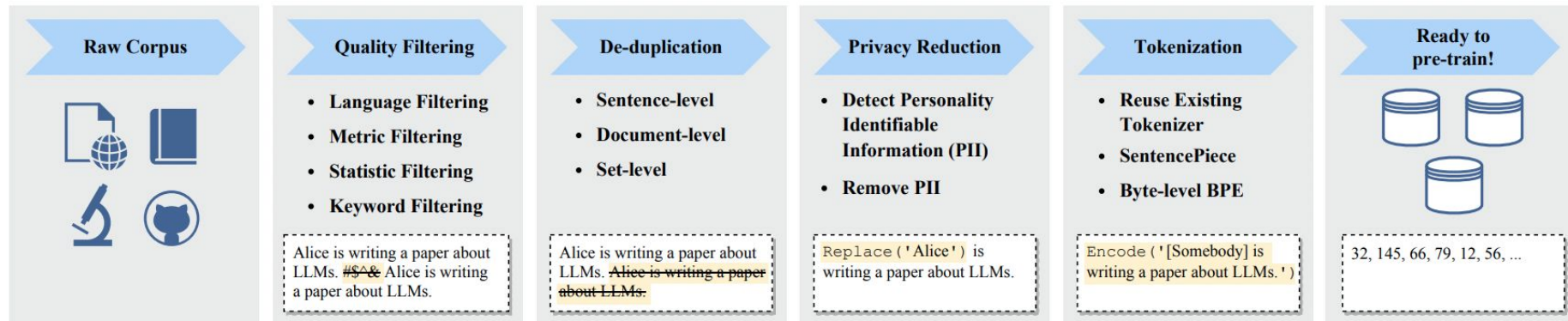
言語モデル一覧: 公開・非公開と基本情報がまとめられている

| Model | Release Time | Size (B) | Base Model | Adaptation IT | RLHF | Pre-train Data Scale | Latest Data Timestamp | Hardware (GPUs / TPUs) | Training Time | ICL | CoT |
|----------------------|--------------|----------|------------|---------------|------|----------------------|-----------------------|------------------------|---------------|-----|-----|
| T5 [73] | Oct-2019 | 11 | - | - | - | 1T tokens | Apr-2019 | 1024 TPU v3 | - | ✓ | - |
| mT5 [74] | Oct-2020 | 13 | - | - | - | 1T tokens | - | - | - | ✓ | - |
| PanGu-α [75] | Apr-2021 | 13* | - | - | - | 1.1TB | - | 2048 Ascend 910 | - | ✓ | - |
| CPM-2 [76] | Jun-2021 | 198 | - | - | - | 2.6TB | - | - | - | ✓ | - |
| T0 [28] | Oct-2021 | 11 | T5 | ✓ | - | - | - | 512 TPU v3 | 27 h | ✓ | - |
| CodeGen [77] | Mar-2022 | 16 | - | - | - | 577B tokens | - | - | - | ✓ | - |
| GPT-NeoX-20B [78] | Apr-2022 | 20 | - | - | - | 825GB | - | 96 40G A100 | - | ✓ | - |
| Tk-Instruct [79] | Apr-2022 | 11 | T5 | ✓ | - | - | - | 256 TPU v3 | 4 h | ✓ | - |
| UL2 [80] | May-2022 | 20 | - | - | - | 1T tokens | Apr-2019 | 512 TPU v4 | - | ✓ | - |
| OPT [81] | May-2022 | 175 | - | - | - | 180B tokens | - | 992 80G A100 | - | ✓ | - |
| NLLB [82] | Jul-2022 | 54.5 | - | - | - | - | - | - | - | ✓ | - |
| GLM [83] | Oct-2022 | 130 | - | - | - | 400B tokens | - | 768 40G A100 | 60 d | ✓ | - |
| Flan-T5 [64] | Oct-2022 | 11 | T5 | ✓ | - | - | - | - | - | ✓ | - |
| BLOOM [69] | Nov-2022 | 176 | - | - | - | 366B tokens | - | 384 80G A100 | 105 d | ✓ | - |
| mT0 [84] | Nov-2022 | 13 | mT5 | ✓ | - | - | - | - | - | ✓ | - |
| Galactica [35] | Nov-2022 | 120 | - | - | - | 106B tokens | - | - | - | ✓ | - |
| BLOOMZ [84] | Nov-2022 | 176 | BLOOM | ✓ | - | - | - | - | - | ✓ | - |
| OPT-IML [85] | Dec-2022 | 175 | OPT | ✓ | - | - | - | 128 40G A100 | - | ✓ | - |
| LLaMA [57] | Feb-2023 | 65 | - | - | - | 1.4T tokens | - | 2048 80G A100 | 21 d | ✓ | - |
| CodeGeex [86] | Sep-2023 | 13 | - | - | - | 850B tokens | - | 1536 Ascend 910 | 60 d | ✓ | - |
| Pythia [87] | Apr-2023 | 12 | - | - | - | 300B tokens | - | 256 40G A100 | - | ✓ | - |
| GPT-3 [55] | May-2020 | 175 | - | - | - | 300B tokens | - | - | - | ✓ | - |
| GShard [88] | Jun-2020 | 600 | - | - | - | 1T tokens | - | 2048 TPU v3 | 4 d | ✓ | - |
| Codex [89] | Jul-2021 | 12 | GPT-3 | - | - | 100B tokens | May-2020 | - | - | ✓ | - |
| ERNIE 3.0 [90] | Jul-2021 | 10 | - | - | - | 375B tokens | - | 384 V100 | - | ✓ | - |
| Jurassic-1 [91] | Aug-2021 | 178 | - | - | - | 300B tokens | - | 800 GPU | - | ✓ | - |
| HyperCLOVA [92] | Sep-2021 | 82 | - | - | - | 300B tokens | - | 1024 A100 | 13.4 d | ✓ | - |
| FLAN [62] | Sep-2021 | 137 | LaMDA-PT | ✓ | - | - | - | 128 TPU v3 | 60 h | ✓ | - |
| Yuan 1.0 [93] | Oct-2021 | 245 | - | - | - | 180B tokens | - | 2128 GPU | - | ✓ | - |
| Anthropic [94] | Dec-2021 | 52 | - | - | - | 400B tokens | - | - | - | ✓ | - |
| WebGPT [72] | Dec-2021 | 175 | GPT-3 | ✓ | - | - | - | - | - | ✓ | - |
| Gopher [59] | Dec-2021 | 280 | - | - | - | 300B tokens | - | 4096 TPU v3 | 920 h | ✓ | - |
| ERNIE 3.0 Titan [95] | Dec-2021 | 260 | - | - | - | 300B tokens | - | 2048 V100 | 28 d | ✓ | - |
| GLM [96] | Dec-2021 | 1200 | - | - | - | 280B tokens | - | 1024 TPU v4 | 574 h | ✓ | - |
| LaMDA [63] | Jan-2022 | 137 | - | - | - | 2.81T tokens | - | 1024 TPU v3 | 57.7 d | ✓ | - |
| MT-NLG [97] | Jan-2022 | 530 | - | - | - | 270B tokens | - | 4480 80G A100 | - | ✓ | - |
| AlphaCode [98] | Feb-2022 | 41 | - | - | - | 967B tokens | Jul-2021 | - | - | ✓ | - |
| InstructGPT [61] | Mar-2022 | 175 | GPT-3 | ✓ | ✓ | - | - | - | - | ✓ | - |
| Chinchilla [34] | Mar-2022 | 70 | - | - | - | 1.4T tokens | - | - | - | ✓ | - |
| PaLM [56] | Apr-2022 | 540 | - | - | - | 780B tokens | - | 6144 TPU v4 | - | ✓ | - |
| AlexaTM [99] | Aug-2022 | 20 | - | - | - | 1.3T tokens | - | 128 A100 | 120 d | ✓ | - |
| Sparrow [100] | Sep-2022 | 70 | - | ✓ | - | - | - | 64 TPU v3 | - | ✓ | - |
| WeLM [101] | Sep-2022 | 10 | - | - | - | 300B tokens | - | 128 A100 40G | 24 d | ✓ | - |
| U-PaLM [102] | Oct-2022 | 540 | PaLM | - | - | - | - | 512 TPU v4 | 5 d | ✓ | - |
| Flan-PaLM [64] | Oct-2022 | 540 | PaLM | ✓ | - | - | - | 512 TPU v4 | 37 h | ✓ | - |
| Flan-U-PaLM [64] | Oct-2022 | 540 | U-PaLM | ✓ | - | - | - | - | - | ✓ | - |
| GPT-4 [46] | Mar-2023 | - | - | ✓ | ✓ | - | - | - | - | ✓ | - |
| PanGu-Σ [103] | Mar-2023 | 1085 | PanGu-α | - | - | 329B tokens | - | 512 Ascend 910 | 100 d | ✓ | - |



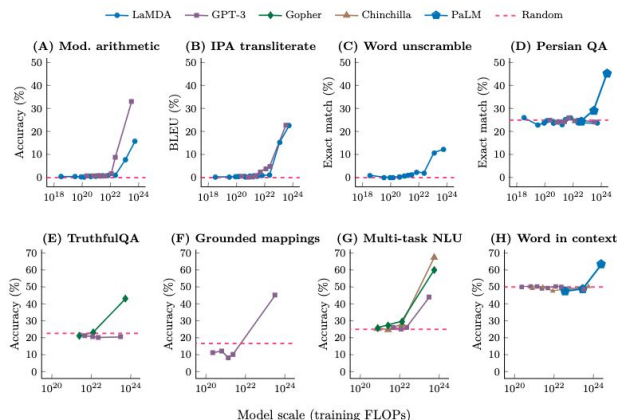
● 【概要】LLMのサーベイ論文

- LLMを包括的に知ろうとしたらまずはこのサーベイを読むべき！LLMについての基礎情報・この数年の流れ・具体的な言語モデル・言語データセットの構築方法・利用とリスクなど、50ページ以上に渡りまとめられている。



典型的なデータ前処理の流れ: クローリング, フィルタリング, 重複除外, プライバシ対策, トークナイズを経て事前学習ができる形に加工される

- 【概要】LLMにおける、言語モデル規模と言語能力の関係性に関する調査を行った論文。
 - 【新規性】大規模モデルは小規模モデルより高度な言語処理能力を獲得できることを示した(下図参照)。
 - 【実験結果】テキスト分類・質問応答タスクにより小規模と大規模な言語モデルでは挙動が異なることが明らかになった
 - 【議論】議論として、LLMの学習・運用は高コストであることが指摘されている。今後、より効率的な学習方法やリソース削減する手法の開発が求められる。また、多言語処理やマルチモーダル処理など、より高度なAI学習の応用も期待される。

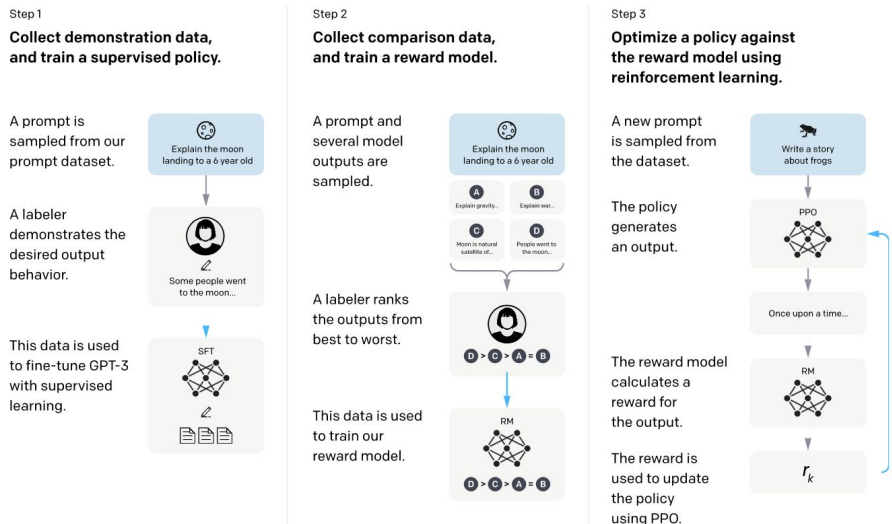


論文中のFigure 2: LLMの性能とモデルスケールの関係。横軸には、モデルのモデルスケール(学習のFLOPs), 縦軸には、言語処理タスクの性能がプロットされている。図中には、GPT-3やLaMDAなど、従来から知られているLLMの性能とスケールが示されている。図からは、パラメータ数が増えるにつれて言語処理タスクの性能が向上することがわかる。ただし、一定以上のパラメータ数を超えると性能向上が鈍化する傾向がある。



● 【概要】InstructGPTの提案論文

- 【新規性】人間のフィードバックによる教示により言語モデルをFine-tuningして出力結果を改善.
- 【手法】下図参照
- 【実験結果】COCO Caption, Visual Genome, CLEVR, GQAにより有効性を確認.



InstructGPTの概要:

LLMに対して人間が望ましい出力結果に関する教師あり学習のラベルを与える (Step 1). 出力結果を比較して, 順位付けを行うことで報酬を獲得 (Step 2). 報酬を元にして強化学習 (Step 3).



- **【概要】**OpenAIによる対話型AIサービスChatGPTのリリース・紹介である。2022/11/30に公表され、瞬間に世界中に広がり最短で100万・1億ユーザを獲得した。GPT-3やInstructGPTの後継モデル(さらにGPT-4の前身)として紹介されたChatGPT。同時にWEBベースのアプリケーションもリリースされた。
 - **【手法】**RLHFを用いたInstructGPTモデルをベースとして、学習をしているが学習データセットが拡張され、Supervised Fine-tuningにより学習されている。
 - **【実験結果】**ChatGPTは言語に対する受け答えに対して汎用的に機能するモデルとなり、言語モデルの有効性を世界中に示す結果となった。右図に示すようにコードに対する質問や回答もできるようになっている。

ChatGPTによるユーザのコードデバッグの様子。エラー出力からユーザとChatGPTがインタラクティブにエラー解消に取り組んでいる。画像は一部を転載。

```
User this code is not working like i expect — how do i fix it?
```

```
resultWorkerErr := make(chan error)
defer close(resultWorkerErr)
go func() {
    defer cancel()
    resultWorkerErr <- b.resultWorker(ctx)
}()

err := b.worker(ctx)
cancel()
if err == nil {
    return <-resultWorkerErr
}
return multierror.Append(err, <-resultWorkerErr)
```

```
ChatGPT It's difficult to say what's wrong with the code without more context. Can you provide more information about what the code is supposed to do and what isn't working as expected? Also, is this the entire code or just a part of it?
```

```
User just part of it — the error never surfaces. I think there is something wrong with the channel
```

```
ChatGPT It's difficult to say without more information about what the code is supposed to do and what's happening when it's executed. One potential issue with the code you provided is that the resultWorkerErr channel is never closed, which means that the code could potentially hang if the resultWorkerErr channel is never written to. This could happen if b.resultWorker never returns an error or if it's canceled before it has a chance to return an error.

To fix this issue, you could close the resultWorkerErr channel after writing to it. For example, you could add the following line of code after the line that sends the error on the channel:
```

```
close(resultWorkerErr)
```

```
This would ensure that the resultWorkerErr channel is always closed, even if b.resultWorker never returns an error. Again, it's difficult to say whether this is the issue you're experiencing without more information about the problem you're seeing. Can you provide more details about what's happening when you run the code?
```

- 【概要】2022/11のリリース後, ChatGPTがどのように使われているのかをメタ解析. ChatGPTに関する30万Tweet, 150論文をダウンロードして調査を行った. 次第に喜びよりも驚き・恐怖が多くなってきた. 医療分野においても用いられるようになった. 英語以外のツイートも増えている. 研究的な要素というよりもOpenAIによるメタな解析である.

| Attribute | Detail |
|------------------|--------------------------|
| date range | 2022-11-30 to 2023-02-09 |
| number of tweets | 334,808 |
| language counts | 61 |
| English tweets | 228127 |
| number of users | 168,111 |

Table 1: Information of the collected Dataset

| Sentiment | Number of tweets |
|-----------|------------------|
| Positive | 100,163 |
| Neutral | 174,684 |
| Negative | 59,961 |

Table 2: Sentiment Distribution of all tweets.

- **【概要】**Googleの提案する540BパラメータのLLM
 - **【新規性】**LLMの異なるタスクに対応するPathwayに分割してスケーリング. 自然言語のみならず, コード生成にも長けている.
 - **【手法】**基本構造はTransformerであり, Wikipedia/CommonCrawl/BookCorpus/OpenWebTextから収集された大規模なデータセットや, 計算リソースとして6144枚のTPUv4を用いて学習されている.
 - **【実験結果】**少量サンプル学習の文脈で最高精度の言語理解タスクを記録している. さらに, BIG-Benchにおいては, 平均的な人間の性能を超える性能を示した.
 - **【議論】**過去最高パラメータのPaLMモデルであるが, まだ性能は止まるところを知らない. さらに, モデル・計算リソース・データをスケーリングさせること, データセットの品質改善が課題と位置付けられている.

- **【概要】**Metaが提案したLLMであるLLaMA論文
 - **【新規性】**7~65Bのモデルを学習, LLaMA-13Bでは少ないパラメータながらGPT-3(175B)の性能を上回る. LLaMA-65BはChinchilla-70B/PaLM-540Bに匹敵.
 - **【手法】**基本はTransformer構造によりモデルが形成されており, 一兆を超える単語のデータセットで学習された.
 - **【議論】**比較的少ないパラメータで高性能なLLaMAモデルを公開することにより, LLM開発が加速する. また, 今後はより大規模なデータセットを使用して学習することで, さらに高度な性能を発揮することができる可能性も示唆される.

| | | | |
|---------------|-------|------|--------|
| CommonCrawl | 67.0% | 1.10 | 3.3 TB |
| C4 | 15.0% | 1.06 | 783 GB |
| Github | 4.5% | 0.64 | 328 GB |
| Wikipedia | 4.5% | 2.45 | 83 GB |
| Books | 4.5% | 2.23 | 85 GB |
| ArXiv | 2.5% | 1.06 | 92 GB |
| StackExchange | 2.0% | 1.03 | 78 GB |

言語データセットをどこから取得したのかが示されている.

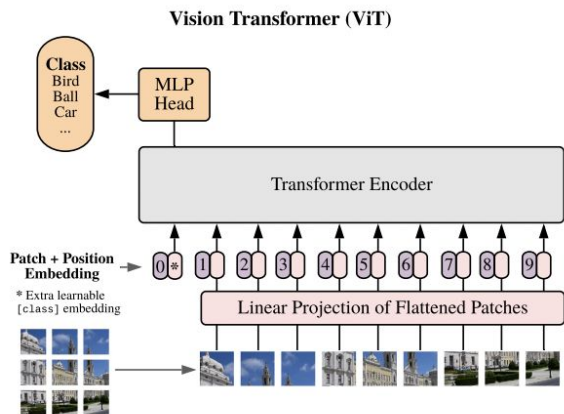
CommonCrawlが最も多く, GitHub/Wikipedia/arXivなどからも言語情報が抽出されている.



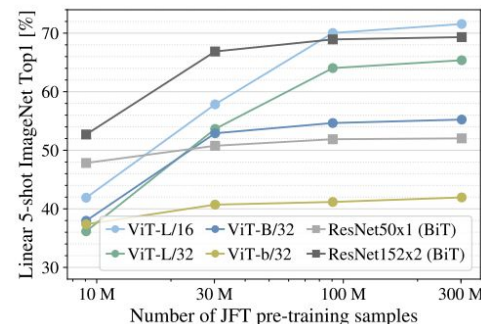
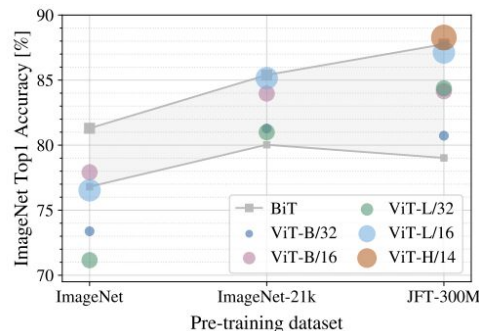
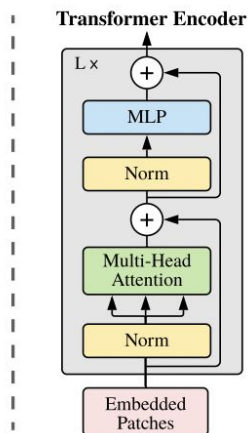
*Transformer*から基盤モデルへ (CV編)

【概要】TransformerアーキテクチャをComputer Visionタスクに適用

- 【新規性】従来畳み込みが主流だったCVのモデルに対して、計算効率が高く精度も良い
- 【手法】画像をパッチに分割してPositional embeddingと共にTransformer Encoderに入力する仕組みで画像を扱えるようにした
- 【実験結果】多くの画像分類データに対して高い性能を出し、データ量が大きくすれば畳み込みベースよりも高精度になる



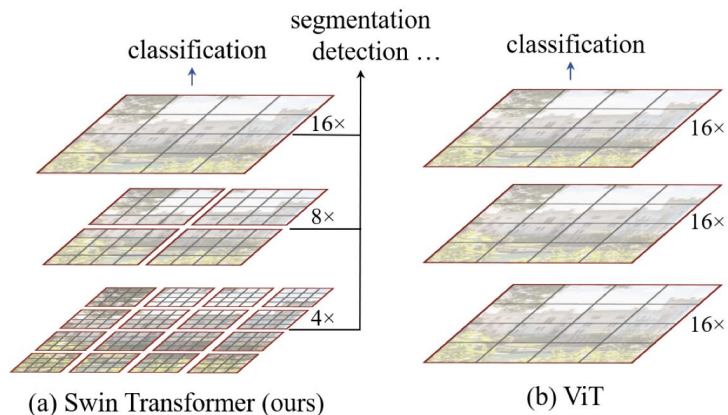
【16x16に画像をパッチ分割して入力する】



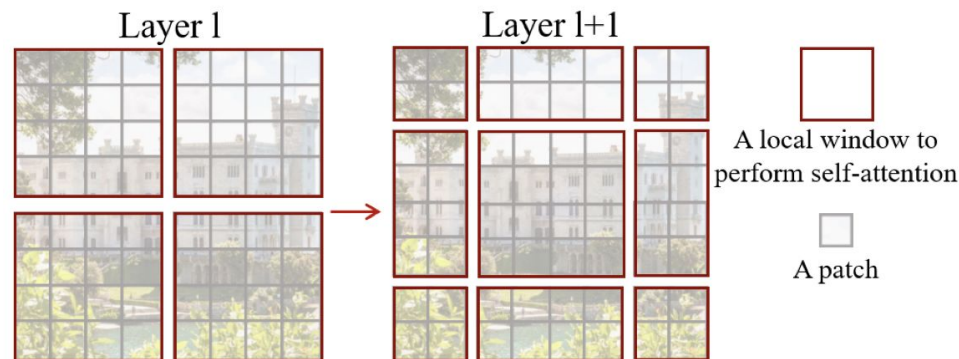
【同一データ量で比較した時、データ量が多ければさらに性能が向上し、畳み込みベースより高止まりしにくい】



- **【概要】**物体のスケールと画像の解像度に対応できるtransformerのモデルの提案
 - **【新規性】**Window sizeを変化させられる階層構造によって計算効率を上げる
 - **【手法】**階層的に分割されるWindowの量を減らしてマージすること, Shifted windowによってWindow間の接続を行う.
 - **【実験結果】**画像分類, 物体検出, セグメンテーションにおいてSOTA

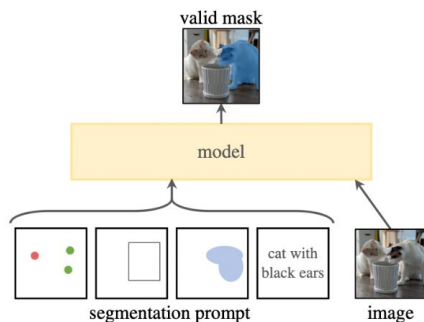


【そうごとと同じWindowサイズを使っていたViTに対して, 階層的に分解してその数を減らしていく】

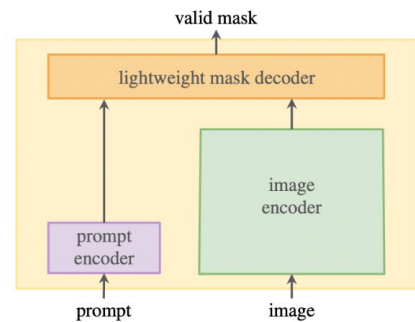


【各Window内でSelf-attentionが計算され, 次の層でWindowの分割がシフトされ, 新しいWindowが生成される】

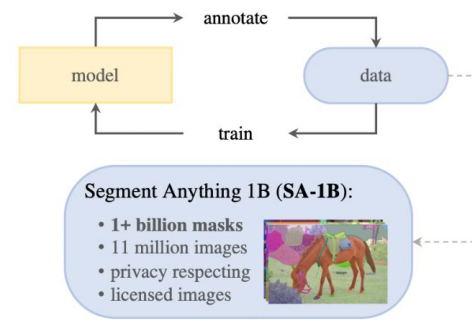
- 【概要】Zero-shotでさまざまなシーンでセグメンテーションが可能な基盤モデル。
 - 【新規性】著作権問題に配慮した11Mの画像と1Bのマスクのデータセットとプロンプト(言語)を対応付けた学習モデル
 - 【手法】セグメンテーションに加えて、プロンプトとセグメントの対応関係を学習し、1つのプロンプトに対して複数のマスクを予測する。アシストマニュアル、セミオート、フルオートの3つのステージで、人とSAMでデータを生成する
 - 【実験結果】Zero-shotで既存の教師ありモデルと同等またはそれ以上の精度



(a) Task: promptable segmentation



(b) Model: Segment Anything Model (SAM)

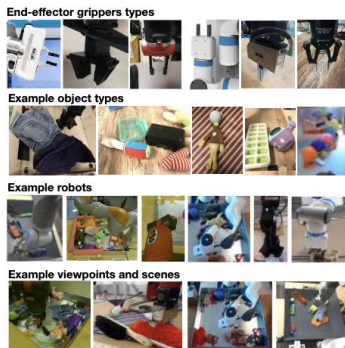


(c) Data: data engine (top) & dataset (bottom)

【プロンプトとマスクを対応づけることでOpenなカテゴリでタスクを可能にし、言語からもマスクを集めることができる。また、人間によるアノテーションを参加することでデータを大量に増やしていく、】

ロボット学習の経緯: 学習データ

- **【概要】**ロボット学習のための大規模データセットRoboNetの提案論文. 異なる研究室で収集された多様なデータで構成, ロボットの相互作用, 環境, カメラの視点などを含む. RoboNet事前学習により, 新規環境での性能が大幅向上し, 学習効率が向上.
 - **【新規性】**ロボット学習の大規模事前学習を可能とした
 - **【実験結果】**RoboNet事前学習したモデルが, 新規環境でのタスクに対して従来の約4倍高いロボットピッキング成功率を示した. また, 少量データ学習の文脈でも5-20分の1の学習で従来法のモデルと同等の精度を実現.
 - **【その他】**RoboNetは拡張可能であり, 今後もデータセットサイズをスケールできる. タスク・環境・ロボット種類・実験設定に対する一般化が進むとされる.

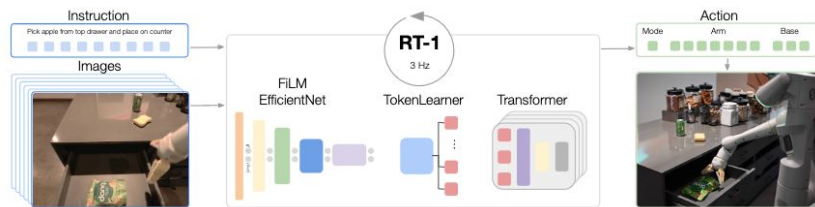


| | |
|--|---|
| Robot type (number of trajectories) | Sawyer (68k), Baxter (18k), WidowX (5k), Franka (7.9k), Kuka (1.8k), Fetch (5k) [6], GoogleRobot (56k) [5] |
| Gripper type | Weiss Robotics WSG-50, Robotiq, WidowX, Baxter, Franka, Kuka |
| Arena types | 7 |
| Arena inserts | 10 |
| Gripper configurations | 10 |
| Camera configuration | 113 |
| Lab environments | 4 |

RoboNetの環境一覧. グリッパー, 物体, ロボット種類, 視点やシーンの例を示しており, さらに表では統計的な数値についても説明されている.

- 【概要】ロボティクス分野における基盤モデル構築に挑戦！大規模かつ多様、タスク非依存のデータから知識転移させることで、ゼロ/少量教師で個別タスクに高い汎化性能を達成
 - 【新規性】効率的な推論によるリアルタイム制御、構造化されたデータセットによる学習やロボティクスタスクに特化したモデル(RT-1: Robotics Transformer)の提案
 - 【実験結果】700以上のロボティクスタスクにおいて、97%以上の成功率を達成

コメント: データとモデルをこれまでより大きくスケールアップさせたRobotics向けの基盤モデル



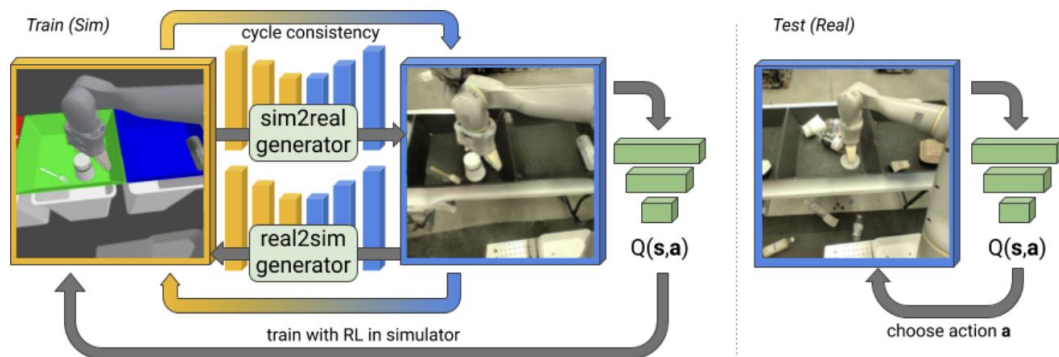
RT-1は画像や自然言語の教示からロボットの移動やアームの行動を促す。35Mパラメータであるにも関わらず、1秒に3FPSで動作するなど効率的である。モデルはFiLMにより条件付けられたEfficientNet, TokenLearner, Transformerにより構成される。



RT-1の大規模かつ実世界での学習により、700以上のタスク処理、高い汎化能力、環境に対するロバスト性、多様なタスクへの適用、新規データへの対応、などが実現できた。

ロボット学習の経緯: *Sim2Real*

- **【概要】**sim-to-realの手法としてドメイン適応を用いた強化学習。
 - **【新規性】**強化学習側の情報をドメイン変換に持ち込むことで、タスクに適したfeatureを保存するドメイン変換を提案
 - **【手法】**CycleGANを用いた画像のドメイン変換において、シミュレーション画像と実画像に対するQ関数が一致するという制約(loss)を追加
 - **【実験結果】**把持タスクにおいてドメイン乱択化やCycleGANによるドメイン変換を用いた先行研究より高い成功率を達成



シミュレータから実画像への変換と強化学習タスクを一緒に学習させることでタスクに有用な情報が変換先に含まれるようにする。

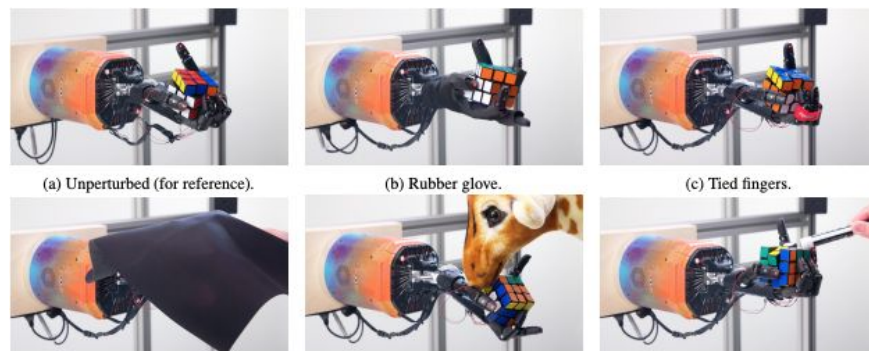
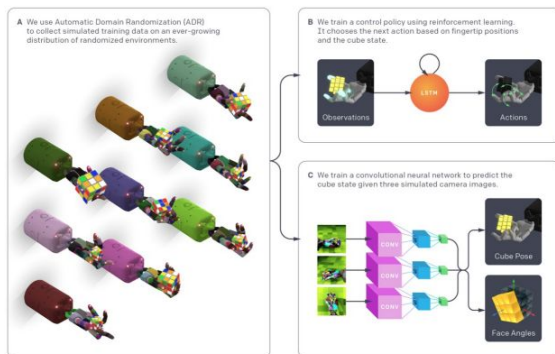
把持タスクにおいてCycleGANを用いた一般的なドメイン変換より高い成功率を実現

| Simulation-to-Real Model | Robot 1 Grasp Success |
|--------------------------|-----------------------|
| Sim-Only [19] | 21% |
| Randomized Sim [19] | 37% |
| GAN | 29% |
| CycleGAN | 61% |
| GraspGAN | 63% |
| RL-CycleGAN | 70% |

- 【概要】多指ハンドによるin-handマニピュレーションタスクの学習

- 【新規性】ルービックキューブを揃えるという非常に複雑度が高いマニピュレーションタスクをsimulation上の学習とドメイン乱択化を用いることで実環境で実現したこと
- 【手法】ドメイン乱択化の自動調整, 動作学習はシミュレーション上での強化学習, ロボットプラットフォームの工夫
- 【実験結果】平均26.8回の操作(回転, フリップ)に成功. 43回の操作が必要な状態からキューブを完全に揃えるタスクの成功率20%, 半分揃えるタスクの成功率60%.

Train in Simulation

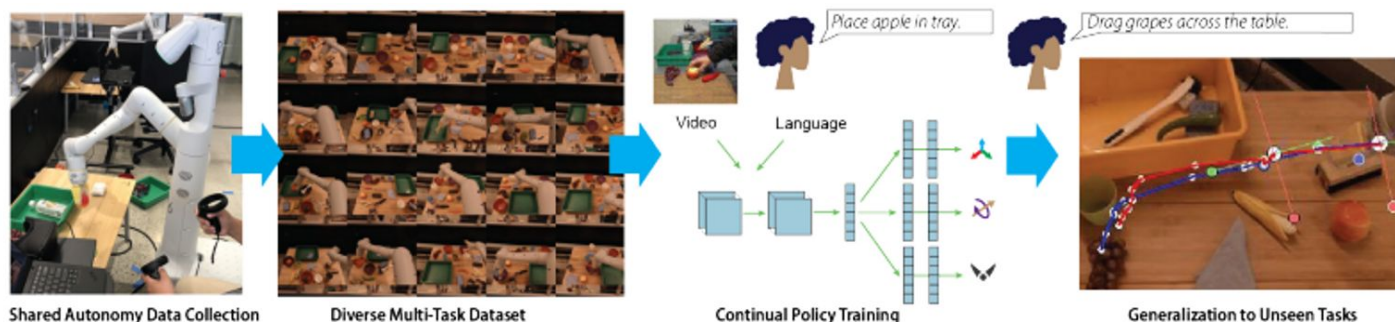


【ドメイン乱択化を用いてシミュレーションで学習データを生成. 強化学習で制御方策を学習. 画像からキューブの状態を予測するモデルをCNNで学習】

【学習した方策は, 学習時に存在しない多様な外乱に対してもロバスト】

ロボット学習の経緯: 学習モデル

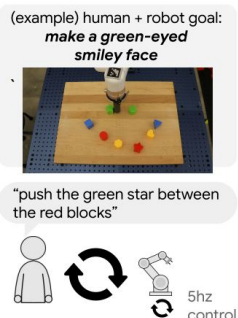
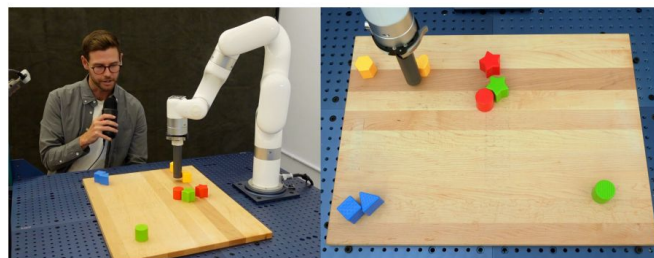
- **【概要】** Video/language で条件付け可能な zero- or few-shot 模倣学習モデルの提案
 - **【新規性】** 大規模かつインタラクティブに多種マニピュレーションを学習するシステム
 - **【手法①】** 学習済みモデルによるタスクを人が修正可能な shared-autonomy で学習データセットを増強する。
 - **【手法②】** Video/language の特徴でロボットの視覚処理ネットワークを条件付け. 同じ video/language の特徴をコサイン類似度で近づけるよう学習する。
 - **【実験結果】** 100種のマニピュレーションを学習し, 未経験の24種のタスクにも約44%で成功



模倣学習システムの概要. データ収集ではまず人が遠隔操作した11,108回の学習データを収集して事前学習する. 学習済みネットワークがタスクを実施し, 修正が必要であれば人が介入してデータを増やす. 合計25,877回・125時間のマニピュレーションデータで学習している.

- 【概要】リアルタイムにopen vocabulary対応ロボットの模倣学習を実現。
 - 【新規性】咄嗟の指示(「e.g. 少し右へ」)にリアルタイムで対応するLLM
 - 【手法】ロボットの軌道(教示用遠隔操作デモ)に対して人がラベル化情報を与え、大規模モデルを構成→言語志向のBehaviour cloning.
 - 【実験結果】リアルタイムの指示にも93.5%(成功率)で正しく対応できた
 - 【その他】87,000のユニークな教示を含むモデルで検証. 言語ラベル付き軌道データ

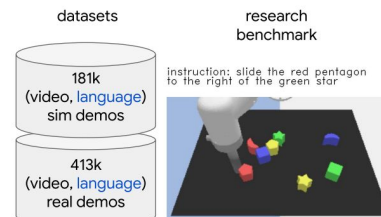
a) Interactive, real-time natural language guidance



b) One robot, 87,000 diverse behaviors

- push the green star between the red blocks
- nudge the red star slightly right
- place the triangle on the top right side of the heart
- separate the green circle from the red star
- slide the blue cube down a bit
- put the heart in the center of the board
- move red star into the red circle
- point your arm at the blue triangle

c) Language-Table: data + benchmark

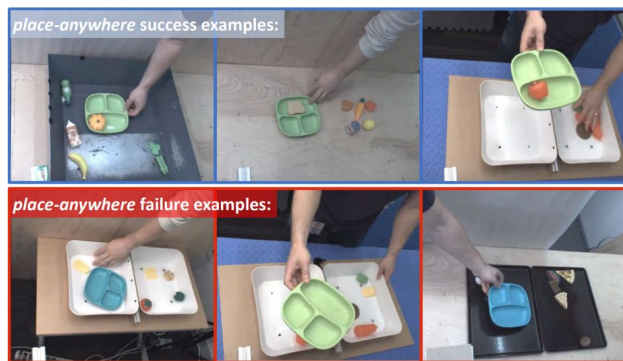


【人のリアルタイムの指示に対応した模倣学習. 人の教示は5Hz間隔で送信可能.】

【既存のビデオデータに再ラベル化を与えることでデータを構成. ここでは一般被験者の説明(ラベル化)に基づく.】

ロボット学習の経緯: 強化学習

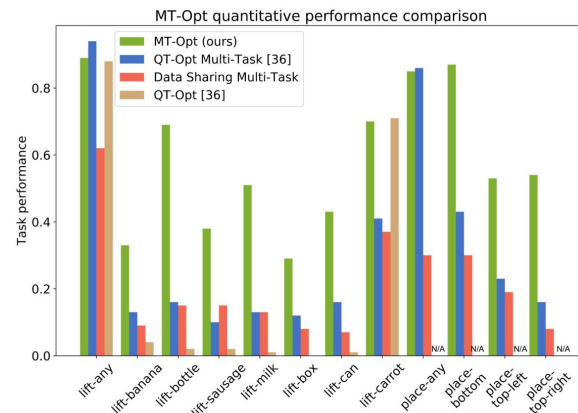
- 【概要】大規模データセットを効果的に学習するロボットシステムおよび強化学習手法**
 - 【新規性】**デモ動画からデータの成否を事前に管理. データセットの効率を調整.
 - 【手法】**多数のロボットが経験を共有しながら, 複数タスクを学習する深層強化学習
 - 【実験結果】**既知のタスクは既存手法同等だが, データセットに多く含まれない(<50%)タスクやほとんど未知のタスク(<1%)にも対応できることを示した.
 - 【その他】**7台のロボットによる9,600時間分のデータ. 成否の数のバランス調整などによって, 「質」を改善し「量」以上の重要性を指摘.



【簡単なインターフェイスにより人のデモにラベルをつけて成否に関する情報を付加する】



【MT-Optの概要】



【タスク別成功率比較. データの偏りに影響されない.】

ロボット学習の経緯: オフライン強化学 習

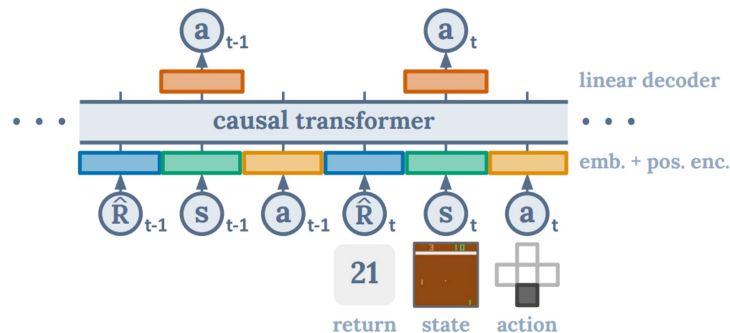
- 【概要】Transformer で逐次予測可能なオフライン強化学習モデル

- 【新規性】強化学習で Transformer を利用できる学習モデル

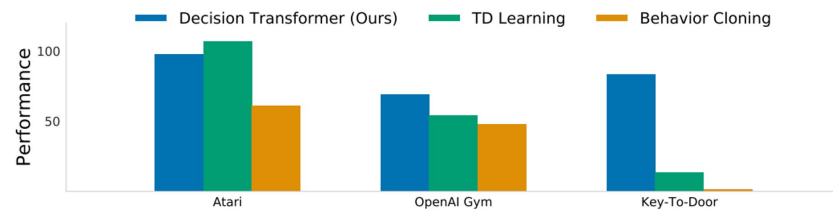
- 【手法】(報酬・状態・行動)を1組としたトークン列を Transformer に入力, 状態から行動を予測する. 報酬には逐次報酬ではなく期待できる将来の報酬和を入力する.

- 【実験結果】複数の行動決定タスクでベースラインを上回る性能. 特に長期タスクで高いパフォーマンスを示した.

- 【その他】指定する目標の報酬和で習熟度が変わる. 大きい報酬和でエキスパートに, 小さい報酬和でビギナーのように振る舞う.



将来の報酬和・状態・行動を入力し, 状態から行動を予測する Transformer モデル.
Causal transformer は GPT と同じアーキテクチャを利用できる.

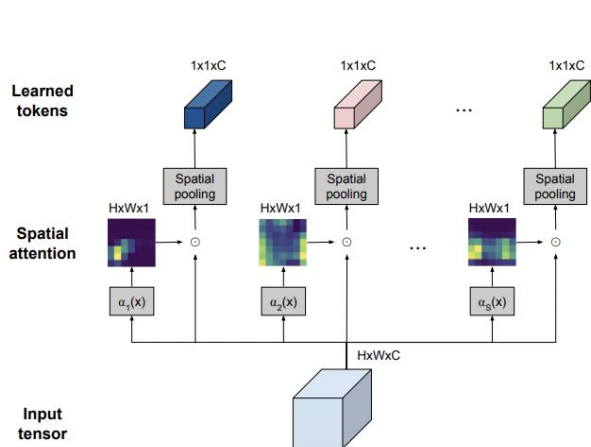


学習結果の性能比較. 将来の報酬和を入力することでパフォーマンスに重要でないデータも方策の学習に適用できる. 長期タスク (Key-To-Door) では状態や報酬の予測を後半のタスクに引き継ぐことが重要となる.

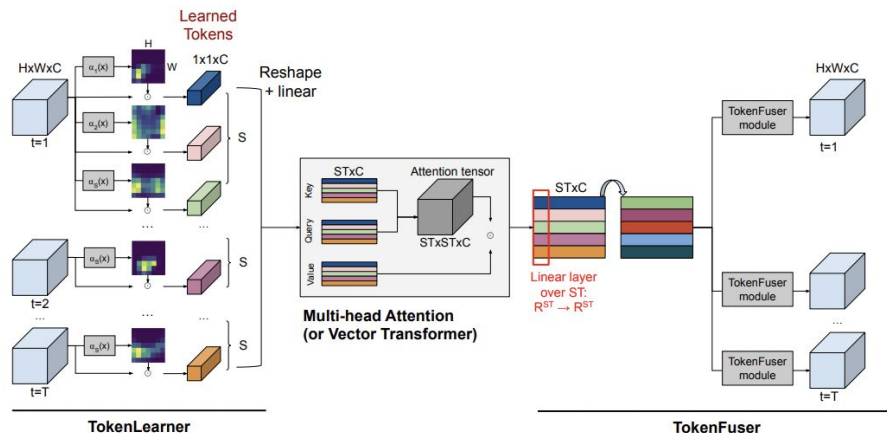


RT-1の主要技術: 学習モデル

- 【概要】視覚データから適応的に少数のトークンを見つけ出す表現学習手法
 - 【新規性】画像と動画において、パッチや解像度によらないトークン数で効率的かつ効果的に得ることができる
 - 【手法】チャンネルと時間方向に対してそれぞれ分割してトークンの集合を生成する
 - 【実験結果】動画認識に対して、計算量を抑えつつ、高い性能を発揮した



【入力を分割して複数のトークンベクトルを生成する】




【トークンを統合して再構成するTokenFuserと一緒に用いて表現学習を行う】

大規模言語モデルの活用: ロボティクス

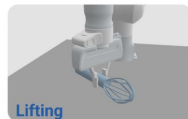
【概要】道具のアフォーダンスの認識に向けた言語付きメタ学習フレームワークモデル

- 【新規性】言語が道具の共有する構造やアフォーダンス理解の補助に活用。
- 【手法】道具の情報を示すGPT-3のプロンプトと画像を入力したOff-policy RL. 収集されたエピソードは学習済みBERTに入力され効果的にサンプリング。
- 【実験結果】メタ学習の効果により4つの異なるタスクについて性能の向上。ゼロショット性能についても未知の道具に対するアフォーダンスの獲得が確認。
- 【その他】道具の初期姿勢(見え方)によってはアフォーダンスが得られにくい場合がある。基本的にはシンプルなタスクに限定→Keypointの動作認識などを考慮した既存アプローチとの統合が今後の解決策。

Phase 1: Meta-training with tool descriptions

 "A mallet typically has a long, cylindrical shape with a flat head on one end. Mallets are used to strike or push objects, or to hammering in nails"

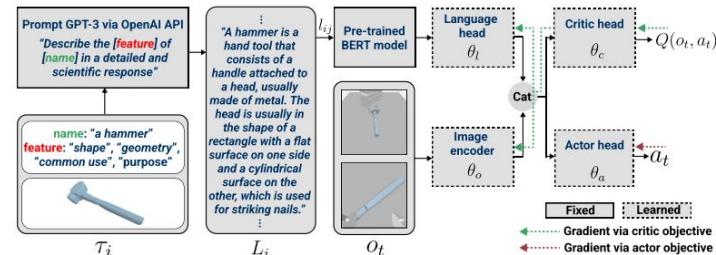
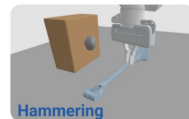
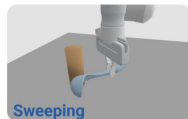
 "The handle of a plier is a tool used to grip objects. It is designed to fit comfortably in the hand. The handle is typically long and thin"



Phase 2: Adaptation to unseen tools

 "Most paint rollers are cylindrical in shape, with a rounded end and a flatter end. The rounded end is used to apply paint to the surface"

 "A faucet generally has a cylindrical shape, with a hole near the top for the water to come out of. The hole is usually surrounded by a ring"

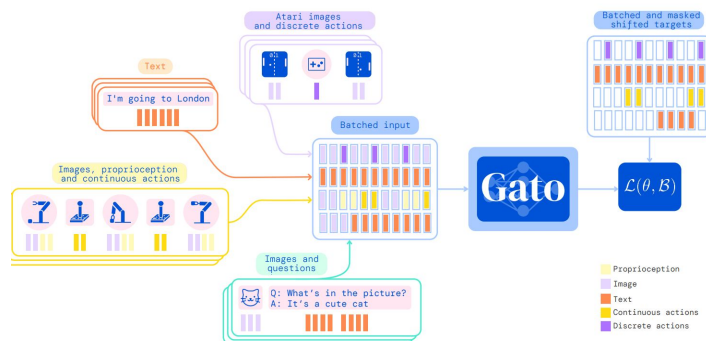


GPT-3が説明する各種の道具の情報が
視覚認識とマッチングすることで、
自動的にアフォーダンスを獲得することができる

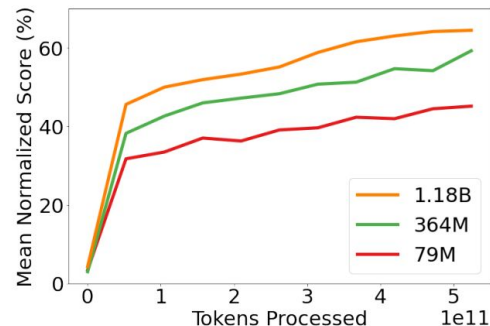
- 【概要】単一のモデルでのマルチモーダル、マルチタスクを実行「万能型AIモデル」
 - 【新規性】大規模言語モデルを他のタスク(制御など)での利用のために拡張.
 - 【手法】エキスパートの教師データを用いる模倣学習. 特にモダリティ毎にTokenize及びEmbedding(ViTなど)を行い, Decision Transformerベースの時系列の予測.
 - 【実験結果】ゲーム, 画像のキャプション付け, テキストチャット, 本物のロボットアームによるブロック積上げ, など**604種類のタスク**を学習
 - 【その他】約12億程度のパラメータ数で興味深い結果.
完全/完璧ではないが, 万能AIに近づく取り組みの一つとして注目すべき.



【ゲーム, キャプション付け, ロボットなど全く異なるタスクに一つのモデルで対応】

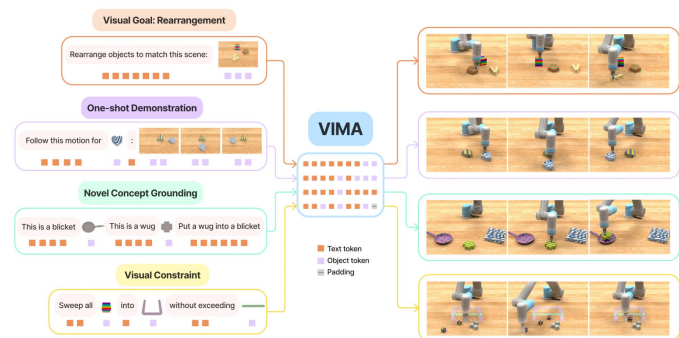


【Gato: 各モダリティに専用の処理を行った上, 統一的空间に知覚を落とし込む.】

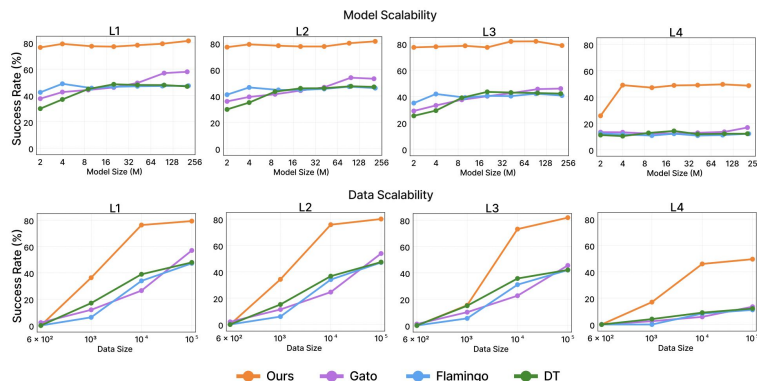


【GatoにおいてもScaling lawが関係】

- 【概要】**視覚に関するいくつかの種類のプロンプトを入力としてさまざまなマニピュレーションタスクを実行するシミュレーション環境とモデルの提案
 - 【新規性】**1つの文としての言語命令だけでなく視覚的な意味を持ったプロンプトに分割して扱い、さまざまなマニピュレーションのタスクを実行可能にした。
 - 【手法】**言語指示と物体情報の画像のトークンの組み合わせから動作を生成する。このとき、過去の動作と物体情報のトークンを共に入力することで連続的に次の動作を生成
 - 【実験結果】**データとモデルサイズにおいてスケーリングの効果が示され、Gatoなどの従来手法より成功率が高い



【さまざまなプロンプトを入力として元の命令を達成するように次の行動を生成する】



【いろんなタスクの難しさ(物か組み合わせが未知かどうか、シーケンスの長さ)で既存の動作生成モデルとの比較】



- 【概要】**事前学習済の大規模言語モデル(LLM)と行動を統合, LLMにより文脈を把握し, ロボットが現実世界にて適切に行動.
 - 【新規性】**現実世界の煩雑な状況においてもLLMにより文脈を把握できた
 - 【実験結果】**主に台所環境にて自然言語による101タスクに対して評価, 溢した飲み物を拭くためのタオルを持って来るように指示した際にもロボットが適切に対応
 - 【その他】**ロボットが現実世界の状況に応じた行動を適切に選択できるようになる. また, SayCanの決定プロセスの可視化による解釈可能性が主張されている.

コメント: SayCanはRT-1につながる流れとして重要!

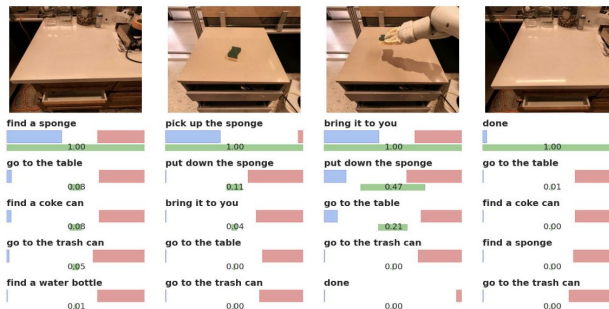


LLMを現実世界で動作させるためにロボットに実装. 事前学習済のロボットスキルを適切に用いるためにLLM/行動の双方を対応付.

Human: I spilled my coke, can you bring me something to clean it up?

Robot: I would
1. Find a sponge
2. Pick up the sponge
3. Bring it to you
4. Done

Language × Affordance
Combined Score

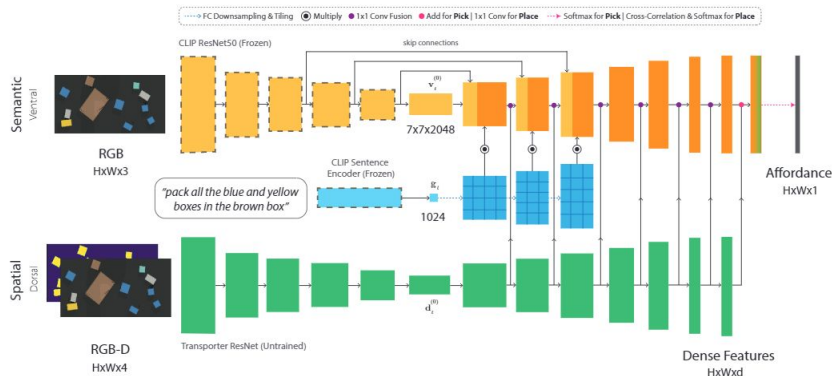


PaLM-SayCanの意思決定を可視化.

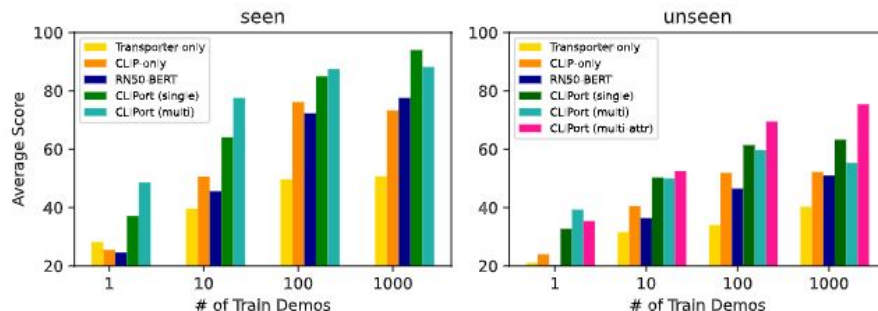


LLM/LVMを用いたロボティクス基盤モデル

- **【概要】**指示やシーンの意味情報と操作精度を両立する2つ経路を持つ模倣学習モデル
 - **【新規性】**Vision-language モデルをマニピュレーションの条件付けに応用
 - **【手法】**指示やシーンの意味情報を CLIP, 空間的特徴を Transporter で抽出し, マニピュレーション用の学習モデルを条件付けする affordance マップを出力
 - **【実験結果】**2つの経路がタスク精度を向上, マルチタスクで汎化性を示した
 - **【その他】**LLM + manipulation で一つのベンチマークとして利用可能

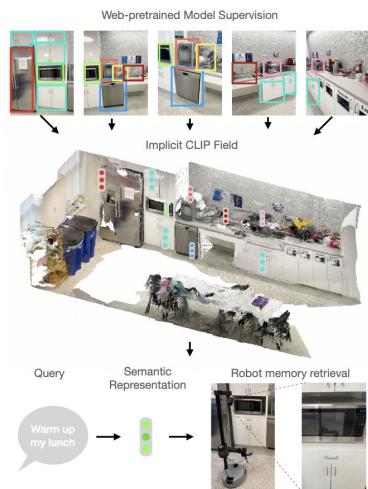


- ① Semantic stream は学習済みの CLIP がシーンの意味情報を,
- ② spatial stream は未学習の Transporter が空間的特徴を抽出する。

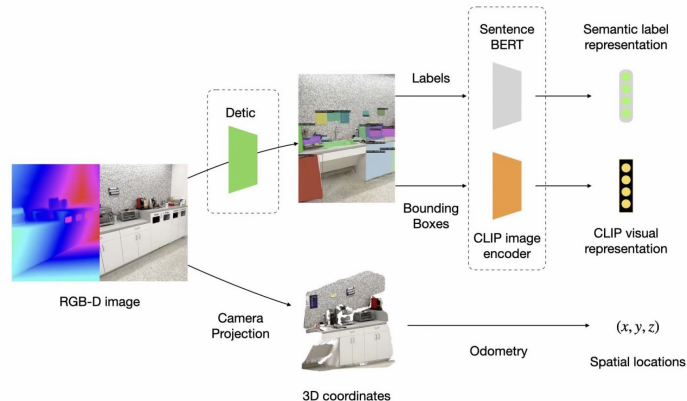


経路が1つの学習モデル (Transporter only, CLIP only) と比較して高い成功率を示し, マルチタスクで他手法より汎化性が見られた

- 【概要】**LLMとLVMを使って3D空間と言語の紐付けを行い、言語クエリに対してナビゲーションを可能にする
 - 【新規性】**セグメンテーション, 空間と言語検索や自己位置推定などを追加の教師付けなしで様々なタスクに使用できる
 - 【手法】**DeticとSentence BERT, CLIPを使った2D空間と言語の対応付け, カメラ軌跡からの2Dと3D対応の2つを統合して学習
 - 【実験結果】**少数サンプルでも3Dセグメンテーションの精度が高い

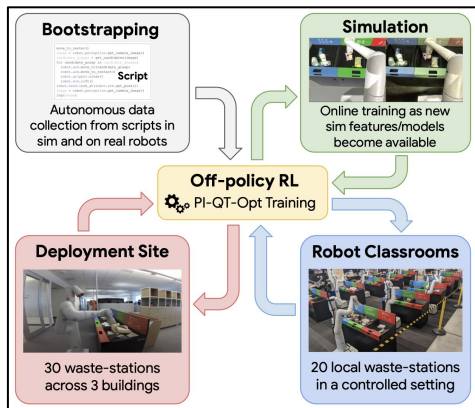


【事前学習済みのセグメンテーションモデルを使って、3D空間に言語情報をマッピングするように学習する。その後、言語クエリから3Dの空間を検索して目的の画像を見つけるようにナビゲーションを実行する】



【RGB,Depthの画像とカメラ軌跡を利用して2Dと言語, 2Dと3Dの対応付けをCLIPで学習する】

- 【概要】リサイクル品とゴミの分別を実行する深層強化学習のシステム。実世界のデータとシミュレーションデータで構成される大量のデータを同時に学習する
 - 【新規性】End-to-Endの枠組みを維持しつつ、シミュレーションと実世界のデータを1つのポリシーとして学習するシステムとした
 - 【手法】スクリプトを実環境とシミュレーションで使えるように用意することで自動でデータを収集し、PI-QT-optでE2Eモデルを学習
 - 【実験結果】84%の物体を正確に仕分けことができ、未知では2年の実証で40-50%の成功率



トレーニングデータやテストシーン

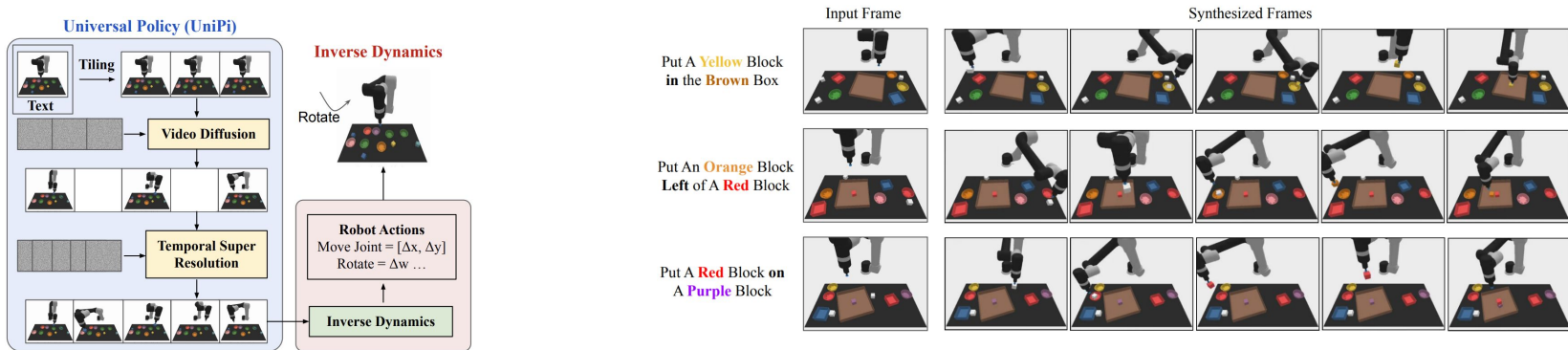


未知物体

論文紹介：メタサーベイ

- **【概要】**テキスト付きの動画生成モデルからロボットの動作を生成
 - **【新規性】**動画ベースなので汎用的. 環境やタスクに応じた報酬設計が不要
 - **【手法】**動画生成は3D Unet, 動作生成は小規模なNN
 - **【実験結果】**シミュレーション上でのピックアンドプレースタスクを実現
 - **【その他】**実世界タスクでロボット行動の動画生成も成功

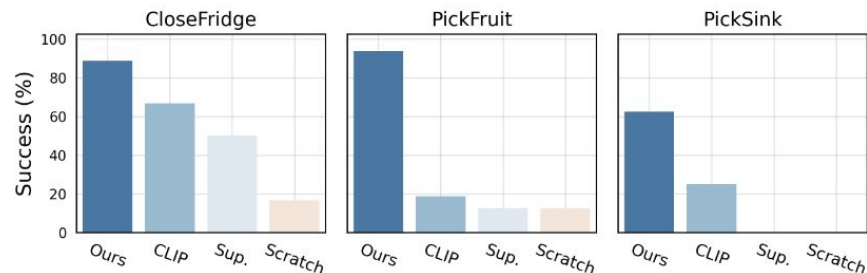
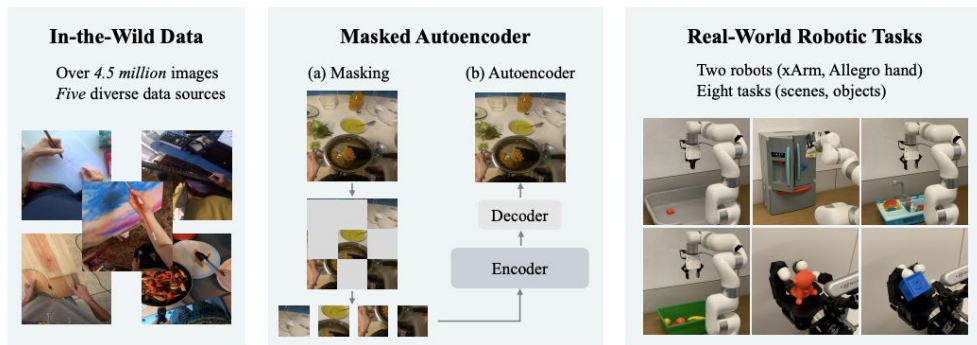
Fine-tuning用ロボットタスク動画・テキストペアは7200個



全体像. Video Diffusionはインプット(画像, テキスト)→アウトプット(動画)の3D U-Net. 超解像した動画から行動(6軸ロボットアーム各関節変位量, グリッパ操作合計7自由度)を小規模なニューラルネットワークで推定(Inverse Dynamicsの部分).

シミュレーション上でのテキスト条件付き動画生成の結果. タスクのAccuracyは46-60%程度. タスクの物理状態はあまり考慮されていない(吸着ON→無条件で近くにある物体をピック). そのためか, 現実ではロボット動作動画を生成するところまでになっている(23.4.18時点).

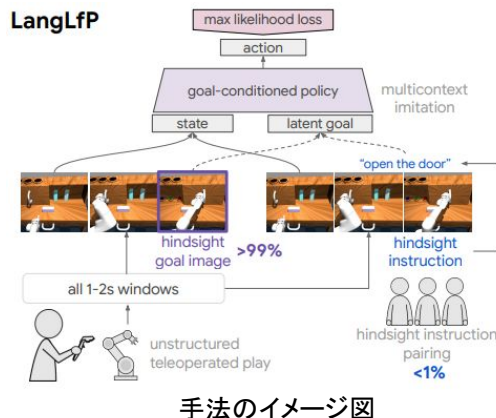
- **【概要】**実世界のロボットタスクのための多様な実画像を用いた自己教師事前学習を検討
 - **【新規性】**NLP/CV分野で成功しているMAEをロボットタスクの事前学習に適用
 - **【手法】**Ego4D含めた複数の一人称視点動画データセット(4.5M)にてMAE
 - **【実験結果】**ImageNet/CLIP事前学習モデルより8種類のタスクにおいて高精度
 - ImageNet: 1.2M (image only), CLIP: 400M (text-image pair)



5つのデータセット(主に一人称視点動画)から4.5億枚の学習画像を確保
Transformer × MAEにて事前学習し、その事前学習モデルをRoboticsに適用

ImageNet事前学習モデル(画像) / CLIP事前学習モデル(画像と言語)の両者と比較して提案手法が高精度

- 【概要】マルチタスクの模倣学習において、僅かな言語ラベルでタスク指示を可能に
 - 【新規性】ラベルなしのdemonstrationデータを活用することで、僅かな言語ラベルで、free-formの言語によるrobotに対するタスク指示を可能にした点
 - 【手法】タスク指示なしのdemonstration(=Play)を使った学習(=Learning from PIY; LfP)は著者らの先行研究. LfPではタスクゴールを画像として与えていたが、これは実環境では難しい. 本研究では後付けでゴール画像と言語ゴールを共通の潜在空間に写像するmulticontext imitationを導入した.
 - 【実験結果】1%以下の言語ラベルでゴール画像を与えていた場合と同等以上の成功率.



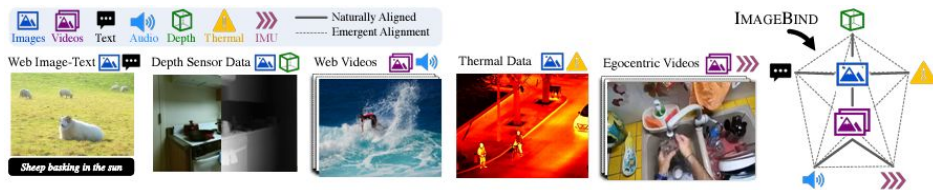
| Method | Input | Training source | Task conditioning | Multi-18 Success (18 tasks) | Chain-4 Success (925 long-horizon tasks) |
|------------------------|--------|---------------------------|-------------------|-----------------------------------|--|
| LangBC | pixels | predefined demos | text | 20.0% \pm 3.0 | 7.1% \pm 1.5 |
| Restricted LangLfP | pixels | unstructured demos | text | 47.1% \pm 2.0 | 25.0% \pm 2.0 |
| LfP | pixels | unstructured demos | image | 66.4% \pm 2.2 | 53.0% \pm 5.0 |
| LangLfP (ours) | pixels | unstructured demos | text | 68.6% \pm 1.7 | 52.1% \pm 2.0 |
| TransferLangLfP (ours) | pixels | unstructured demos | text | 74.1% \pm 1.5 | 61.8% \pm 1.1 |
| LangBC | states | predefined demos | text | 38.5% \pm 6.3 | 13.9% \pm 1.4 |
| Restricted LangLfP | states | unstructured demos | text | 88.0% \pm 1.4 | 64.2% \pm 1.5 |
| LangLfP (ours) | states | unstructured demos | text | 88.5% \pm 2.9 | 63.2% \pm 0.9 |
| TransferLangLfP (ours) | states | unstructured demos | text | 90.5% \pm 0.8 | 71.8% \pm 1.6 |

TABLE I: Human language conditioned visual manipulation experiments

Playを学習に使うとタスク成功率が向上する(LangBC < LfP), Playを使った場合でGoal指示を言語にすると同等(LfP \approx LangLfP, 僅かによくなっている). 言語の部分に事前学習を使うと更に改善(LangLfP < TransferLangLfP)



- **【概要】**画像をキーにして多様なモダリティを関連付け (Image to bind).
 - **【新規性】**6つのモダリティのjoint embedding spaceの構築. 全てのモダリティが揃ったデータは必要としない.
 - images, text, audio, depth, thermal, IMUデータ
 - **【手法】**pairが近く, pairになっていないものが遠くなるように埋め込む
 - **【実験結果】**異なるモダリティの検索, 生成, 演算
 - **【その他】**IMUが入っている (身体情報の一種という点でロボットやドローン応用の参考になるかもしれない)



【異なるデータソースから得られるimage+別モダリティを共通のembedding spaceに埋め込む】

1) Cross-Modal Retrieval

| Audio | Images & Videos | Depth | Text |
|-------------------|-----------------|-------|--|
| Crackle of a Fire | | | "A fire crackles while a pan of food is frying on the fire." "Fire is crackling then wind starts blowing." "Firewood crackles then music..." |
| Baby Cooing | | | "A baby is crying while a toddler is laughing." "A baby is laughing while an adult is laughing." "A baby laughs and something..." |

2) Embedding-Space Arithmetic

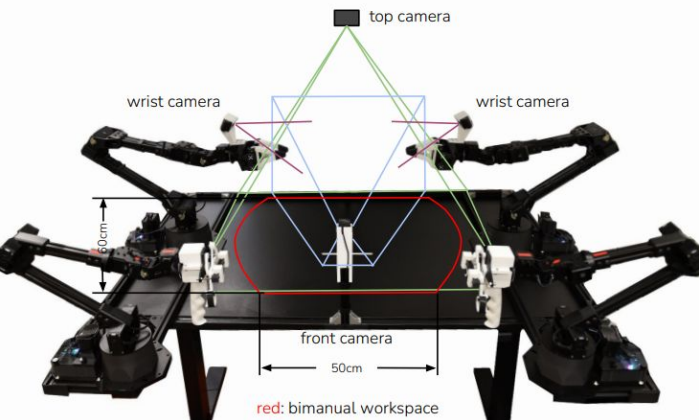
+ =

3) Audio to Image Generation

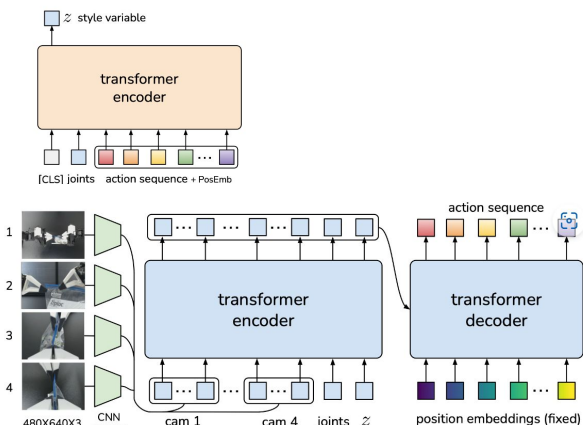
→ → → →

【アプリケーション】

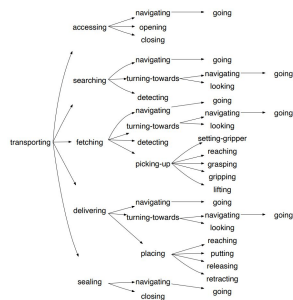
- **【概要】**ALOHAを用いた手先の繊細な作業に対するEnd-to-End学習
 - **【新規性】**低コストシステム「ALOHA」の提案/Transformerによるアクションの学習
 - **【手法】**ALOHAによるデータ収集から条件付きVAEによるアクションの学習
 - **【実験結果】**他手法(RT-1など)が対応しない繊細なタスク(両手での挿入作業やジップロックを開ける, など)への適応を示した.
 - **【その他】**20k USD (約300万円)で実現可能. 多くの研究者が利用できるよう公開.



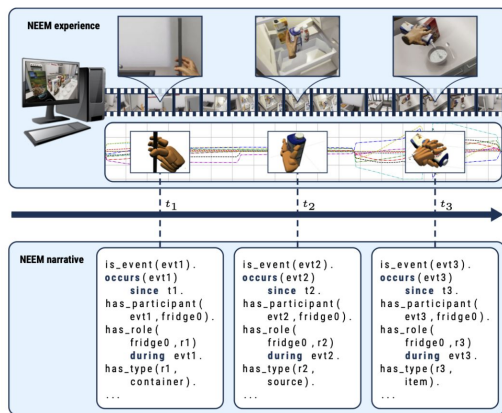
【双腕テレオペレーション・ALOHA】



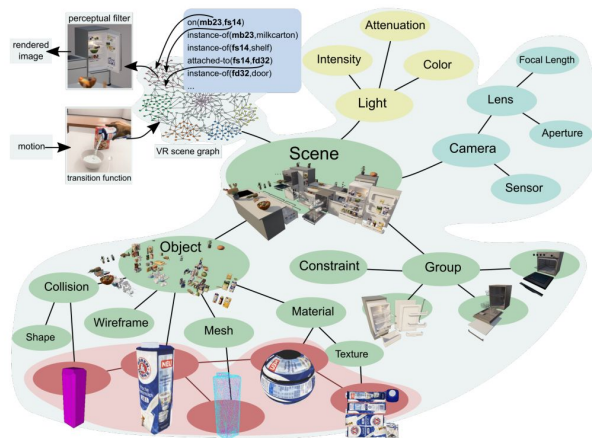
- **【概要】**ロボットの日常タスクの実現に有用な認知モデルCRAMの紹介。
 - **【新規性】**CRAMがロボットの日常タスクの実現の課題点を克服するのにどのように効果的であるかを解説。
 - **【手法】**タスクはロボットの具体的で最小の行動として細分化され理解。構文を構築するスキームにより行動を計画。生成モデルを用いてこれらを構築。
 - **【その他】**ロボットと環境およびタスクに関するクエリに答える具体的な方法を提示。ビデオの動画からモデルを構築し、現実のロボットへの適用の効果を示した。



【最小単位の実現可能な行動に分類】

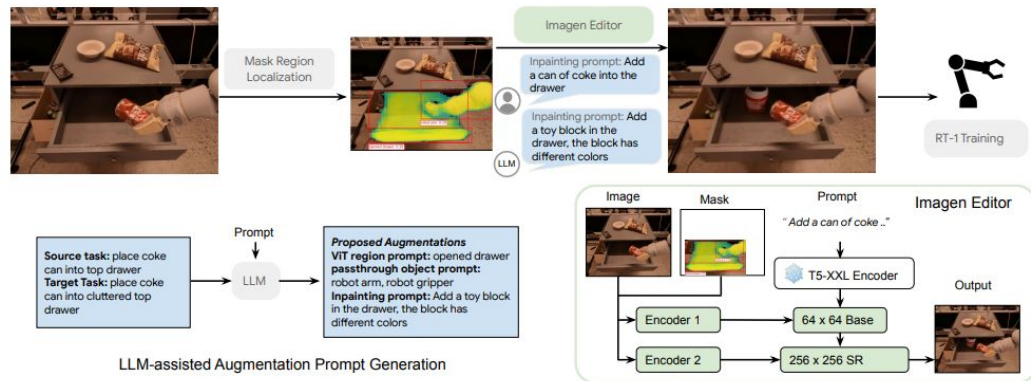
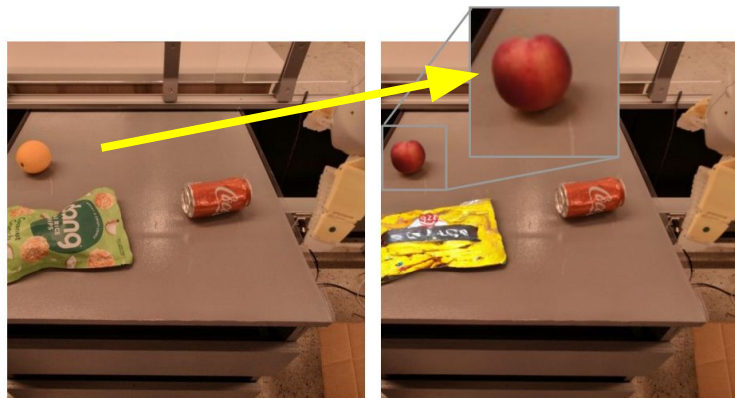


【NEEMs (Narrative-Enabled Episodic Memories)】



【home - KnowRob】

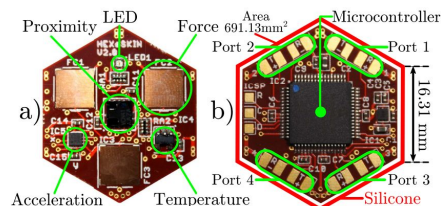
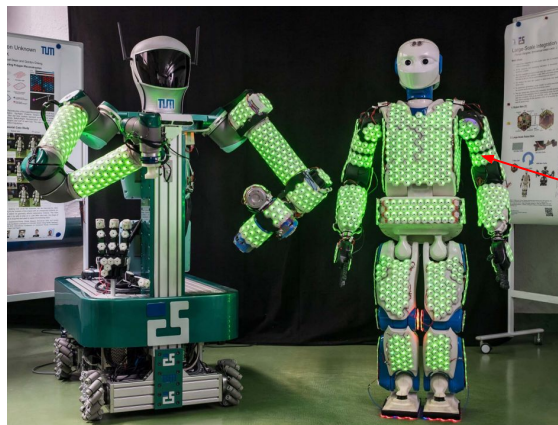
- **【概要】**画像データセットを言語モデルを用いて拡張(教師なしで大規模化)する手法.
 - **【新規性】**最新の画像生成モデルを活用した基盤モデルの拡張に着目.
 - **【手法】**既存の画像生成モデルによりRT-1を拡張. 指定範囲(手動)の対象に言語ベース(GPT-3)に拡張内容を考え, 新しい画像を生成.
 - **【実験結果】**RT-1の拡張→データに含まれない対象物がある場合に成功率向上.
 - **【その他】**既存モデル(複数のFM)を組み合わせることの有用性を示す好例.



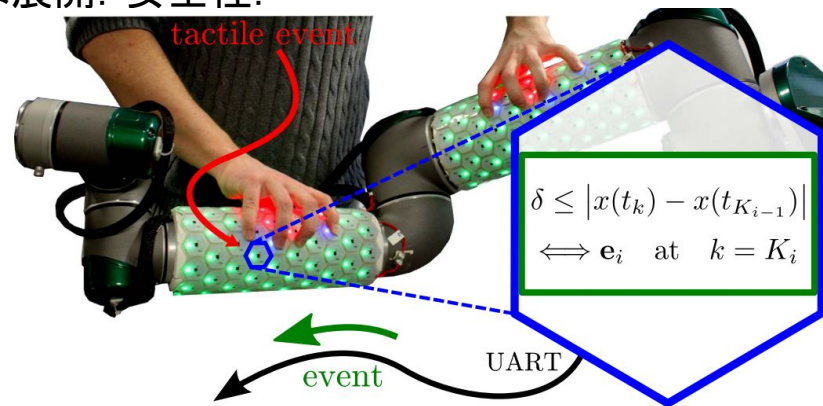
【テキストの補助の下, 拡散モデルによりデータ拡張. タスクの学習に適する”妥当”な画像を生成している.】

【まず画像のマスキング(認識)により領域を特定し, 言語ベースに画像を加工する. 拡張されたデータはRT-1の学習に用いられる】

- 【概要】スキンパッチ/ 皮膚全体の知覚を処理するネットワーク構成を提案。全身人工皮膚を備えた敏感なヒューマノイドロボット。
 - 【新規性】ロボットスキンの開発/ 多量のセンサを処理するアイデアを提案。
 - 【手法】上体, 腕, 脚, 足の裏に1260個のセル(13000個以上のセンサー)を搭載。大量のセンサ情報を扱うために「イベント駆動型」処理を提案(変化のみに着目)
 - 【実験結果】不均一な床面に対応。片足でバランス...人と安全にハグ(抱擁)。
 - 【その他】介護ロボットなど, 人協調への新たな展開。安全性。



【各センサーセル】

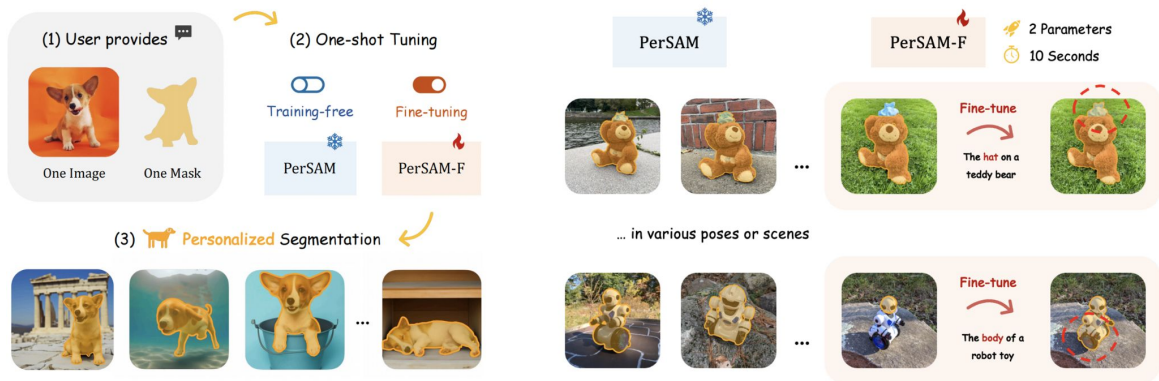


【全体を処理に用いるのではなく、変化があった部分に着目することで通信や処理の問題を解消している】

参考: <https://shiropen.com/seamless/a-comprehensive-realization-of-robot-skin>



- **【概要】** One-shotでSegment Anything modelを任意に調整。
 - **【新規性】** 簡単な入力と学習でSegment Anythingの活用の幅を拡大。
 - **【手法】** 明示的にプロンプト(マスク-画像の対)を与えることで画像中のPositive(対象物)/Negative(背景)の関係を抽出→Segment Anythingへ入力。
 - **【実験結果】** Fine-Tuneの容易さに加え, 高精度に実施可能であることを証明。
 - **【その他】** SAMの重みは固定. 10秒以内/ 2つのパラメータによってマスクを学習可能である。

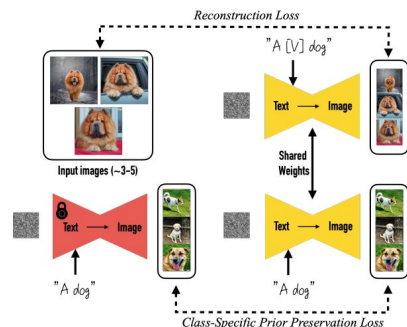


【PerSAMの概要】

| Method | mIoU | Param. |
|-------------------------|--------------|----------|
| <i>Existing Methods</i> | | |
| Painter [51] | 56.35 | 354M |
| Visual Prompting [2] | 65.88 | 383M |
| SEEM [63] | 80.50 | 157M |
| SegGPT [53] | 94.26 | 354M |
| <i>Our Approach</i> | | |
| PerSAM | 89.32 | 0 |
| PerSAM-F | 95.33 | 2 |
| <i>Improvement</i> | +6.01 | |

【ファインチューニングが必要なパラメータ数に着目. 少量で高い精度を実現】

- 【概要】生成モデルをユーザーの目的に特化 (e.g. 特定の「犬」の写真をベースに生成)
 - 【新規性】text-to-imageの課題であった任意の出力方向の調整を実現.
 - 【手法】特定の被写体の画像 (3-5枚) から再生成誤差 (→①特定画像の特徴) と Prior Preventing Loss (→②汎用な特徴) という2つの損失関数を用いてモデルの Fine-tuningを実施.
 - 【実験結果】生成画像と元画像の類似性に関する指標を定義し検証.
 - 【その他】[Stable Diffusion](#)の追加学習方法として界限で着目.
an award candidate at CVPR 2023 [[link](#)].



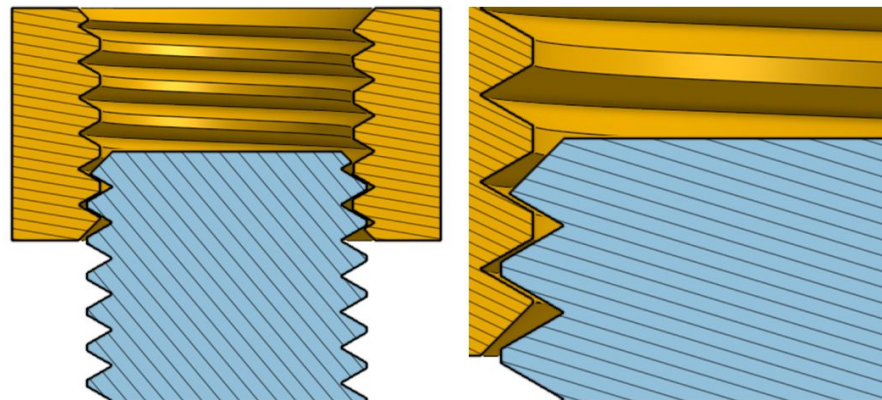
【一般的な生成モデルの場合、普通のテキストで「Dog」とすると様々な種類の「Dogs」が出力。ここでは任意の画像中の「The Dog」が保持可能。】

【二つのLoss関数によってユーザの固有の入力情報と大規模モデルが持つ一般的な特徴とのバランスをとっている】

- **【概要】**Omniverse Isaac Simでのアセンブリタスクのシミュレーションツールの開発
 - **【新規性】**複雑な接触の簡易表現やRL環境の構築
 - **【手法】**SDFに基づいた接触のGPUを用いた統合, 接触削減, Gauss-Seidel solver
 - **【実験結果】**1,000同時のボルト・ナットの接触シミュレーションを高速に実行
 - **【その他】**60のパーツモデル, 3つのアセンブリ環境, 7つのロボットコントローラを含むツール群として公開予定.

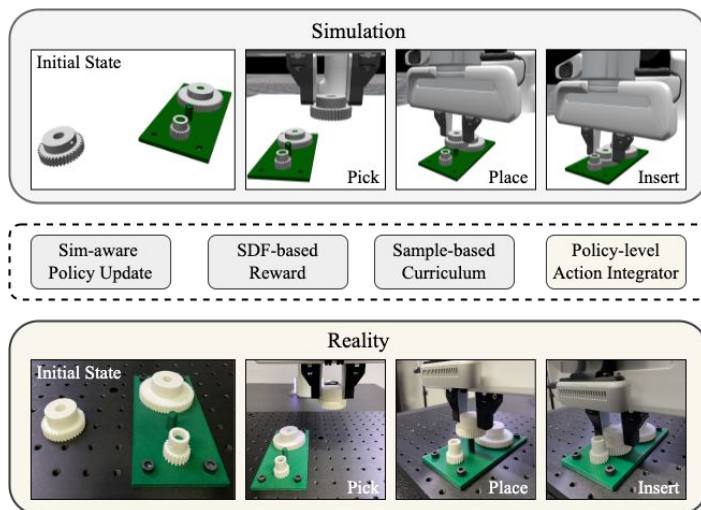


【1,000ペアのボルト・ナットをシミュレーション】



【ボルト・ナット間の複雑な多点接触】

- 【概要】**アセンブリタスクにおけるSim-to-RealでのRLを実現するアルゴリズム・システム・ツールを一体として提案
 - 【新規性】**従来のFactoryに加えて学習ツールや実験環境を一体化して提供
 - 【手法】**学習手法としては1) simulation-aware policy updates, 2) signed-distance-field rewards, and 3) sampling-based curriculaを提案
 - 【実験結果】**右下表

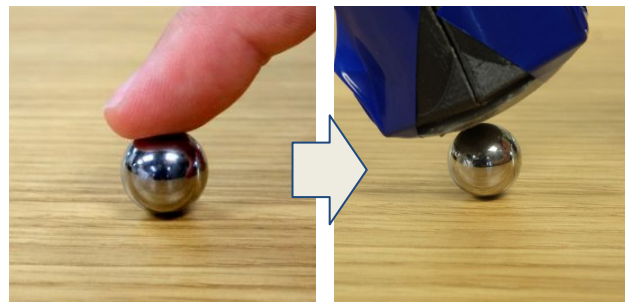
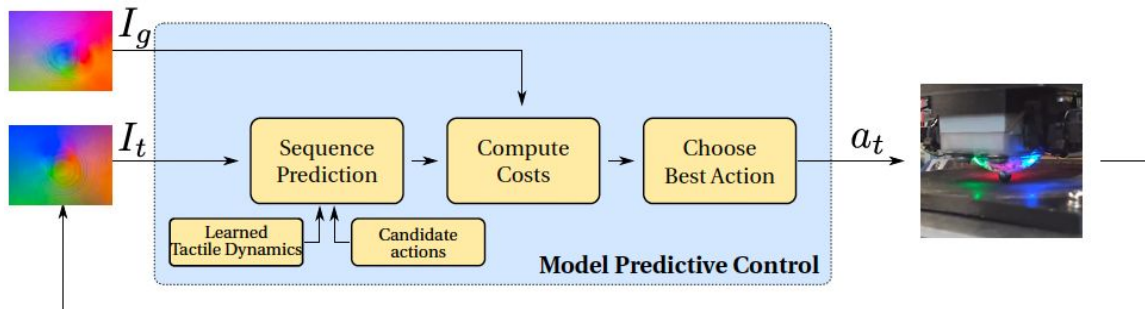


| Asset | Pick | Insert | | Pick-Place-Insert | |
|-----------------------|---------|---------|---------|-------------------|---------|
| | Success | Success | Engage | Success | Engage |
| Round peg 8 mm | 19/20 | 7/10 | 7/10 | 7/10 | 7/10 |
| Round peg 12 mm | 19/20 | 7/10 | 9/10 | 7/10 | 7/10 |
| Round peg 16 mm | 20/20 | 8/10 | 10/10 | 8/10 | 10/10 |
| Rectangular peg 8 mm | 20/20 | 8/10 | 9/10 | 10/10 | 10/10 |
| Rectangular peg 12 mm | 20/20 | 8/10 | 8/10 | 8/10 | 9/10 |
| Rectangular peg 16 mm | 20/20 | 8/10 | 9/10 | 8/10 | 10/10 |
| NEMA 2-prong | - | 10/10 | 10/10 | 10/10 | 10/10 |
| NEMA 3-prong | - | 7/10 | 10/10 | 10/10 | 10/10 |
| Small gear | - | 8/10 | 9/10 | 10/10 | 10/10 |
| Medium gear | - | 9/10 | 9/10 | 9/10 | 10/10 |
| Large gear | - | 10/10 | 10/10 | 10/10 | 10/10 |
| Multi-gear assembly | - | 10/10 | 10/10 | 10/10 | 10/10 |
| Total # | 158/160 | 100/120 | 110/120 | 107/120 | 113/120 |
| Total (%) | 98.75% | 83.33% | 91.67% | 89.16% | 94.17% |

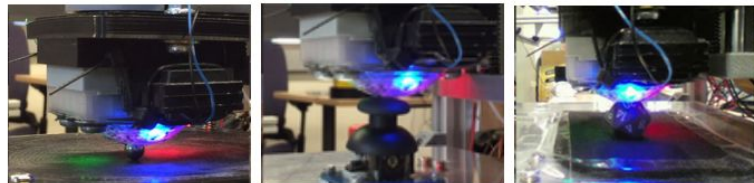
TABLE III: Real-world experimental results for Pick, Insert, and PPI.



- 【概要】GelSightの画像から接触状態を学習により推定. モデル予測制御によりボール操作
 - 【新規性】視触覚として接触モデルを獲得して制御に活用
 - 【手法】下図
 - 【実験結果】右下表



| | Median L2 dist [mm] | | Success Rate |
|-------------------|---------------------|--------------|----------------------|
| | Ball Rolling | Analog Stick | Die |
| Tactile MPC | 2.10 | 5.31 | 86.6% (26/30) |
| Centroid Baseline | 2.97 | 8.86 | 46.6% (14/30) |



←Deep tactile MPCを
3つの異なるタスクで評価
(右に結果を掲載)

TABLE I: Benchmark results for the ball-rolling, analog-stick and die-rolling experiments. The median L2 distances are between the hand-annotated pressure centroid of the final and goal-image. For the die experiment we measure the fraction of examples where the desired face lands on top. Benchmarks are performed with 30 examples.

- **【概要】**視触覚に基づいたSim-to-Real強化学習
 - **【新規性】**視触覚として接触モデルを獲得して制御
 - **【手法】**従来のTactile Gym の拡張として、新たなセンサ形式に対応
 - **【実験結果】**下図参照

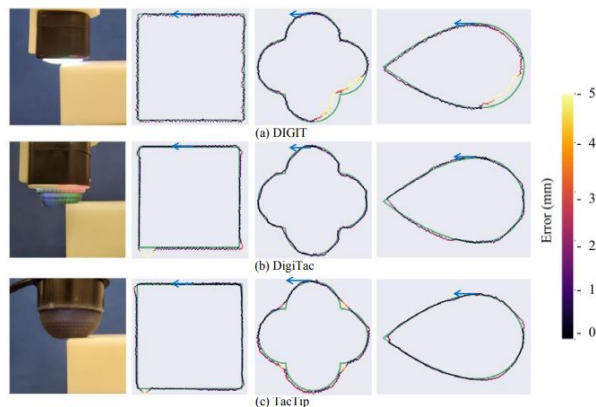


Figure 6. The tactile robot executing edge-following policies on 3 distinct shapes for the (a) DIGIT, (b) DigiTac and (c) TacTip. The ground truth for each object is shown in green and errors of the traced contour from the ground truth are colour-coded (side colour bar). The blue arrow denotes the starting point and direction.

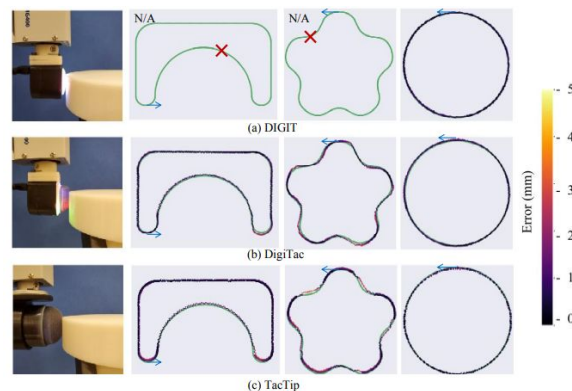
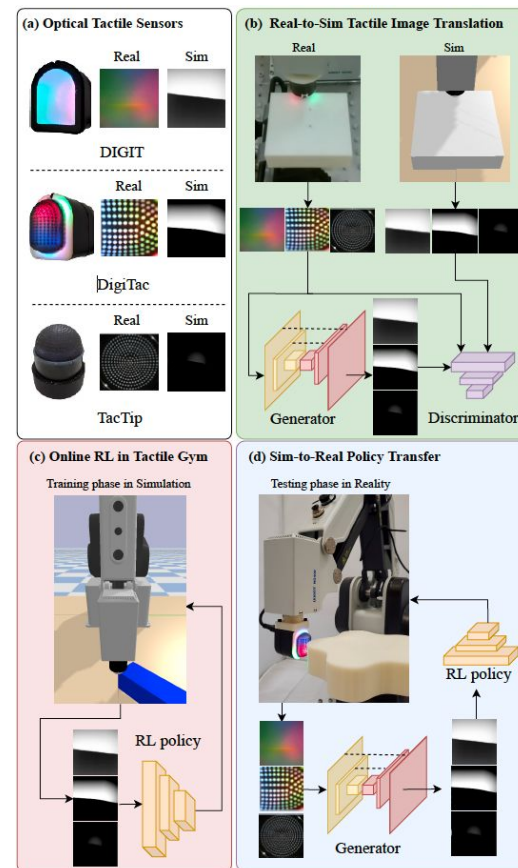
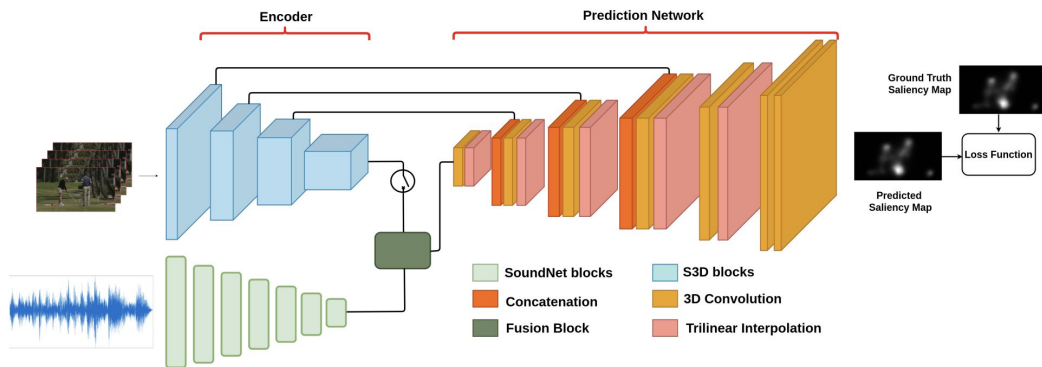


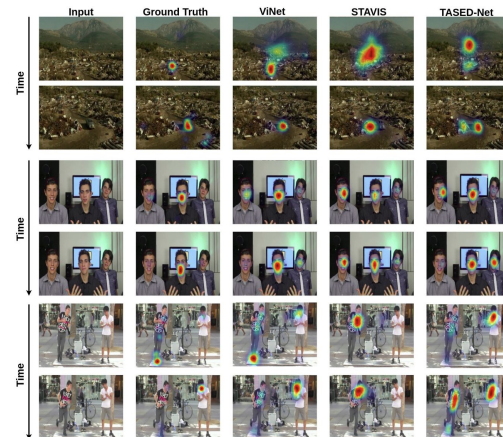
Figure 7. The tactile robot executing 3 surface-following policies on 3 distinct shapes, corresponding to the (a) DIGIT, (b) DigiTac and (c) TacTip. The ground truth for each object is shown in green and errors of the traced contour from the ground truth are colour-coded (side colour bar). The blue arrow denotes the starting point and direction. The DIGIT failed to follow the arch and flower objects at points denoted by the red crosses.



- 【概要】動画と音声から注目する場所であるSaliency Mapを予測する手法
 - 【新規性】SOTAを超える性能. さらに提案手法も現在のSOTAも実は音声を利用しないモデルを学習していたことを示した(音声はもっと上手く使える可能性がある).
 - 【手法】シンプルな畳み込みベースのencoder-decoderアーキテクチャ. backboneは動作認識タスクで事前学習
 - 【実験結果】9つのデータセットでSOTAを更新. AVEデータセットでは3つの指標で人間を上回る性能.

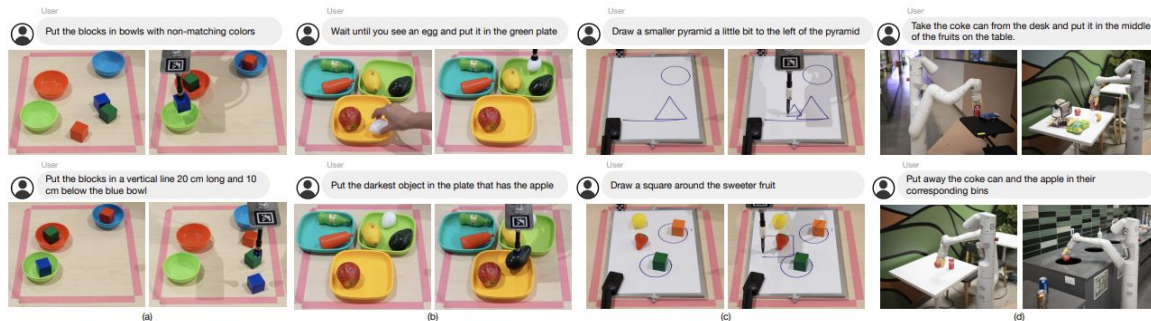
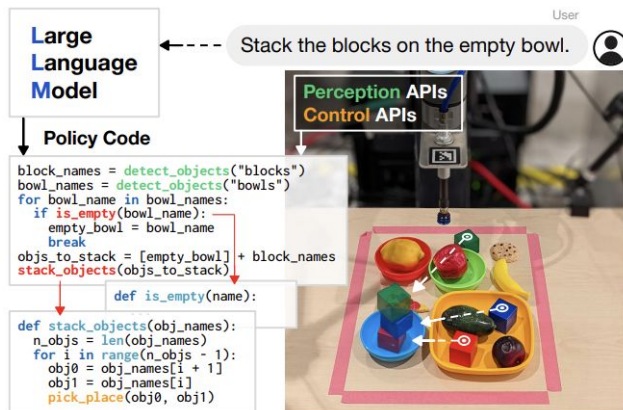


シンプルな畳み込みベースのencoder-decoderアーキテクチャ. skip-connectionで階層的な特徴をdecoderに入力. video encoderはKinectsデータセットで事前学習.



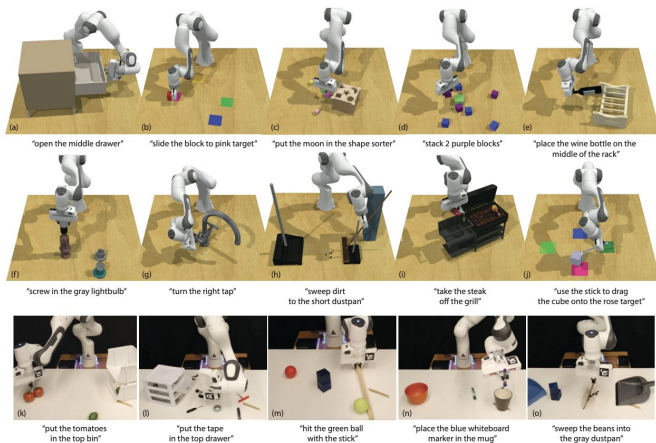
予測したvisual saliencyはSOTAであるSTAVISとTASED-NetよりGround Truthに近い.

- 【概要】LLMを使ってロボットのポリシーを生成することで、空間的な理解を行い新しい命令に一般化し、コンテキスト(行動常識)に応じて曖昧な記述に正確な制御を実行する
 - 【新規性】行動シーケンスに加えて、ピックアッププレースの軌道や速度などの低レベルの動作命令のコードを生成して実行することができる
 - 【手法】いくつかのモジュールにコードとして行動の種類を分割しておき、適切なパラメータまでLLMを使って生成して当てはめて行動生成をする
 - 【実験結果】特にLong-horizonなタスクに対して学習ベースのポリシーより正確

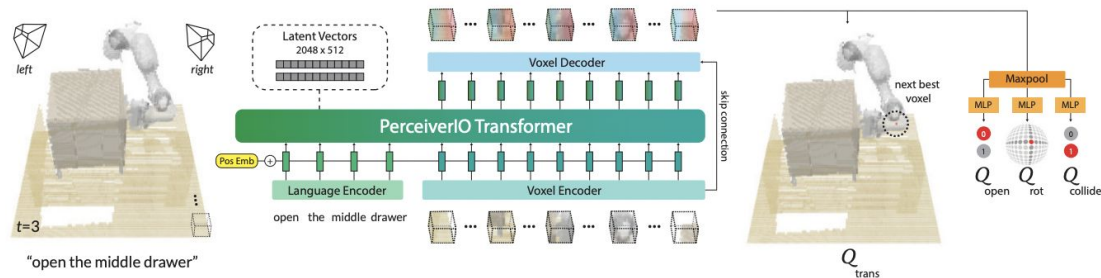


【画像認識モデルと組み合わせることで実環境でも動作可能で、さまざまな空間的理解が必要なタスクをさまざまなロボットで実行できる】

- **【概要】**Transformerベースで言語の条件付けに基づいて6次元の物体操作を行う
 - **【新規性】**Transformerを使って、3D情報と言語から行動生成を行う
 - **【手法】**Perceiver Transformerを改良して、言語による条件付けとボクセル情報を入力して、次の行動のボクセルを予測することから行動を生成する。少数のデモから学習を行う。
 - **【実験結果】**多くの操作タスクにおいて性能が高く、デモ数を増やすと既存手法との性能差が大きくなる

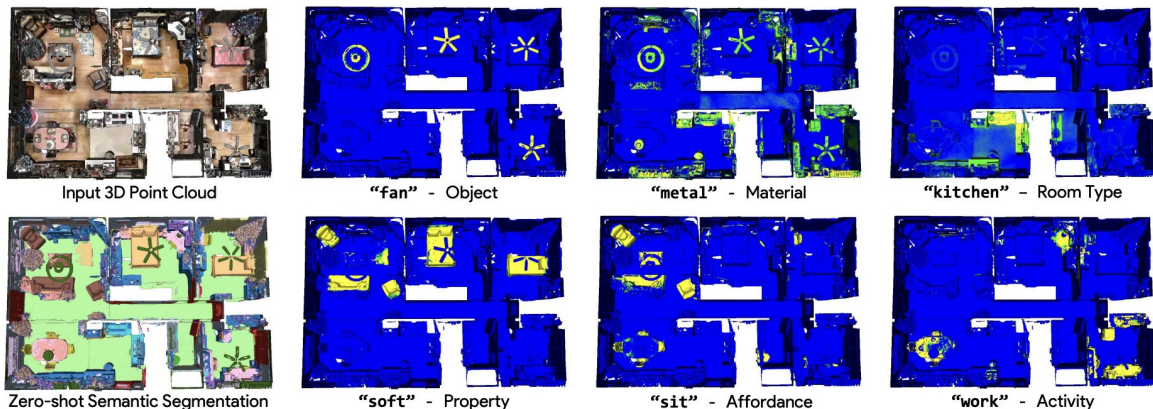


【RLBenchや実環境で言語に対応した少数のデモンストレーションを収集する】

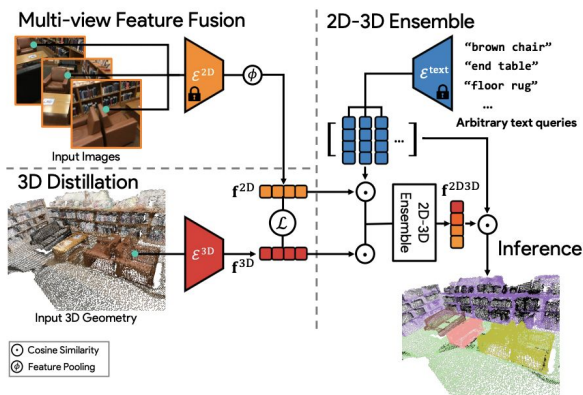


【RGB-Dの3次元情報をボクセルとして、言語条件と共に入力し、将来の3D状態を予測する】

- **【概要】**CLIP特徴空間においてテキストや画像のピクセルと3Dシーンのポイントを対応付けを学習する
 - **【新規性】**タスクに依存することなく, 特定のカテゴリに囚われない3D認識
 - **【手法】**クラスラベルのEmbeddingと2Dとカメラパラメータから紐づく3D情報を対応付けたCLIP特徴で, 2Dのみまたは3Dのみ入力された場合において言語特徴を抽出できるようにする
 - **【実験結果】**複雑な3Dシーンにおける意味的な情報を高い精度で推定できる

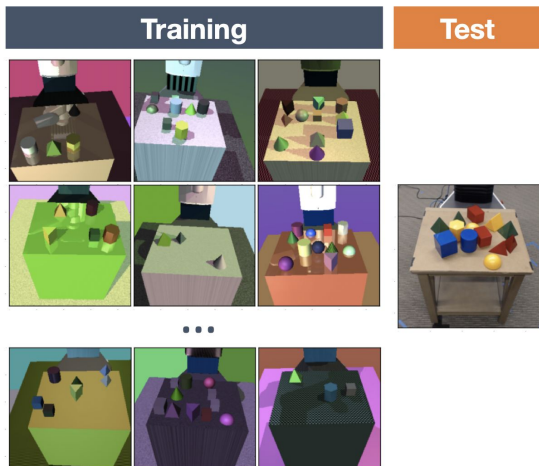


【3Dラベルなし, Zero-shotで3D空間と言語を対応づけられた】



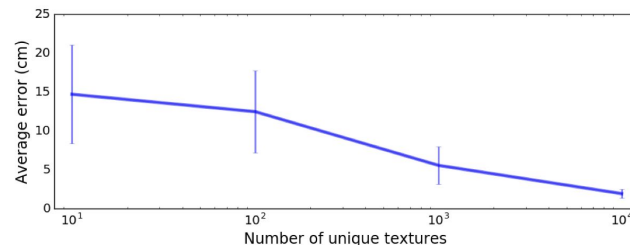
【2Dと3D対応を学習して, CLIPで言語と対応付ける. 推定時は2Dと3D特徴をアンサンブルする】

- 【概要】simulationでの生成データで学習したモデルを実環境で適用できるように画像のランダム化を行う。
 - 【新規性】シミュレータに十分な変動性を持たせて学習させ実環境のドメインに転移
 - 【手法】物体の位置やテクスチャ, カメラ, 光源に関するパラメータ, 画像へのノイズを網羅的に与える
 - 【実験結果】Object Localizationタスクに関して実環境の様々なシーンにおいて高い性能で推定ができた



- Number and shape of distractor objects on the table
- Position and texture of all objects on the table
- Textures of the table, floor, skybox, and robot
- Position, orientation, and field of view of the camera
- Number of lights in the scene
- Position, orientation, and specular characteristics of the lights
- Type and amount of random noise added to images

【上記をランダムに変更してシミュレーション画像をレンダリング】

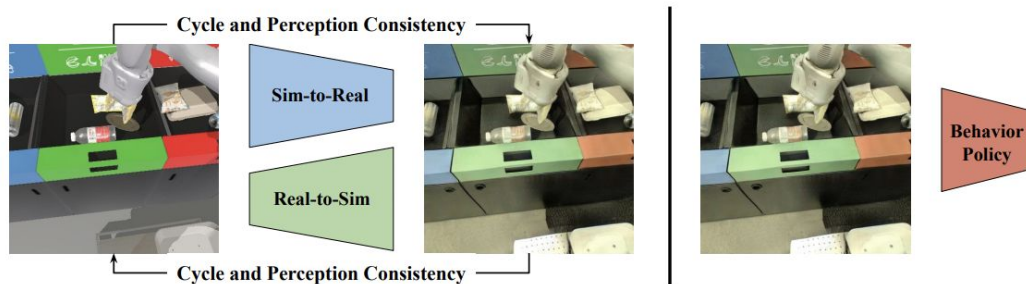


| Average detection error on geometric shapes by method, cm ⁴ | | | |
|--|-------------|-------------|------------|
| Evaluation type | Real images | | |
| | Object only | Distractors | Occlusions |
| Full method | 1.3 ± 0.6 | 1.8 ± 1.7 | 2.4 ± 3.0 |
| No noise added | 1.4 ± 0.7 | 1.9 ± 2.0 | 2.4 ± 2.8 |
| No camera randomization | 2.0 ± 2.1 | 2.4 ± 2.3 | 2.9 ± 3.5 |
| No distractors in training | 1.5 ± 0.6 | 7.2 ± 4.5 | 7.4 ± 5.3 |

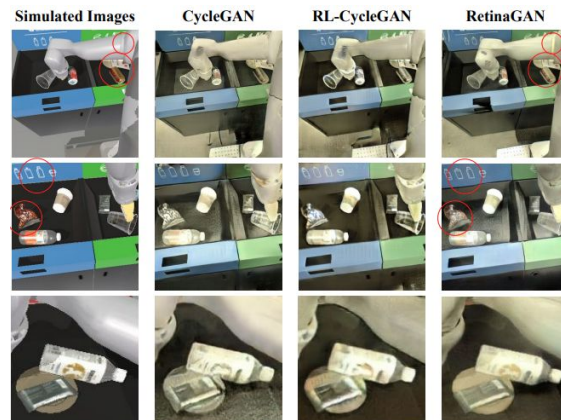
【できるだけ広い範囲でランダム化すると性能がいい】



- 【概要】タスクに適した情報を「物品に関する知識を保存すること」とみなし、タスク非依存な形でドメイン適応に組み込み、強化学習に利用
 - 【新規性】物体検出をsim-to-realのドメイン変換に組み合わせた点
 - 【手法】CycleGANを用いた画像のドメイン変換において、シミュレーション画像と実画像に対して物体検出をした結果が一致するという制約(loss)を追加
 - 【実験結果】把持タスクにおいて、ドメイン乱択化やCycleGAN, RL-CycleGANによるドメイン変換手法より高い成功率を達成。



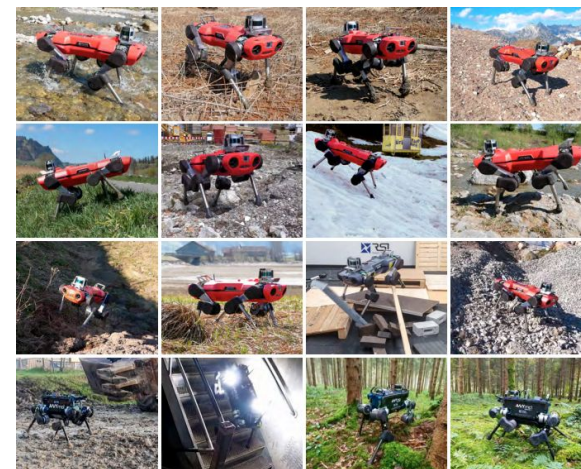
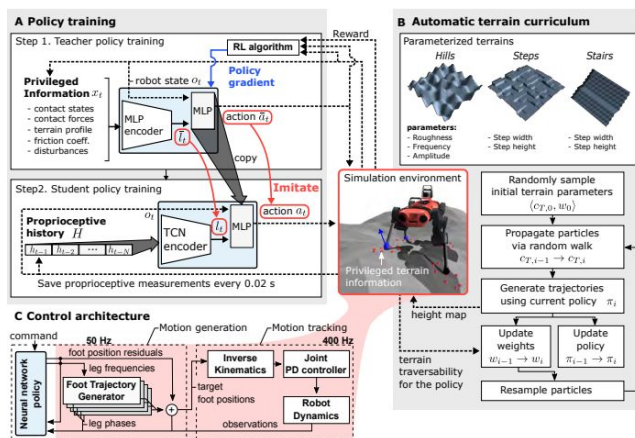
事前学習をしたperception modelを用いてsim-to-realのモデルを学習(左図). 行動のpolicyをこのsim-to-realモデルで生成した画像を用いて学習する(右図)



RetinaGANは他の手法と比べて物体のテクスチャを保存し、実シートの特徴をより上手く再構築している

● 【概要】学習による4足歩行ロボットの移動制御

- 【新規性】複雑な自然地形における4足歩行ロボットの移動制御をシミュレータを使った学習によって実現
- 【手法】teacher-student学習, privileged informationの利用, カリキュラム学習
- 【実験結果】剛体の地面, 少数の手続き的に生成した地面で学習しただけで, 変形も含む複雑な実環境の地面でも歩行が可能



シミュレーションでのみ利用できる情報 (privileged information) を用いて teacher policy を学習。
teacher を模倣するように student policy を学習。カリキュラム学習で地面の難しさを調整。

提案手法により多様で挑戦的な環境における移動を実現

- **【概要】**学習済みのLLMがオフライン強化学習(Offline RL)において、収束速度と報酬の性能改善をもたらすかどうかを調査。
 - **【新規性】**Offline RLを文脈のあるタスクとして捉え言語・視覚などのドメインで事前学習されたシーケンスモデルの転移可能性を調査. 事前学習を強化学習に適用する
 - **【実験結果】**Wikipedia / GPT-2を適用し, 種々の環境における学習を3-6倍加速, 最先端の性能を達成. OpenAI Gym / AtariにおいてDecision Transformerと比較して優位性を確認.
 - **【その他】**Offline RLにおける事前学習の有用性を示唆. 言語事前学習→RLでも効果的と結論.



事前学習済み言語モデルをOffline RLに適用. Wikipediaの文章の事前学習, GPT-2を用いた言語モデルを適用してRLのタスクに対して適用した.

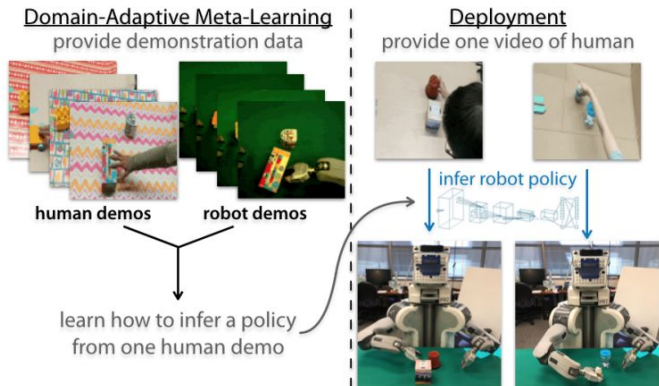
- 【概要】**データ収集コストが高いロボット分野で、大規模モデルのように多数のタスクを1つのモデルで学習させる目的に適した強化学習手法
 - 【新規性】**過去と未来の相関である”predictive information”を強化学習に利用. これは多くのタスクに共通するのでモデル化すると汎用エージェントの性能向上につながる.
 - 【手法】**QT-Optにpredictive informationのauxiliary lossを追加. ちなみにQT-Optは
 - 【実験結果】**Bin pickingタスクにおいて、有名なLevineらの研究(Googleの14台のロボットを使ったもの)より高い成功率



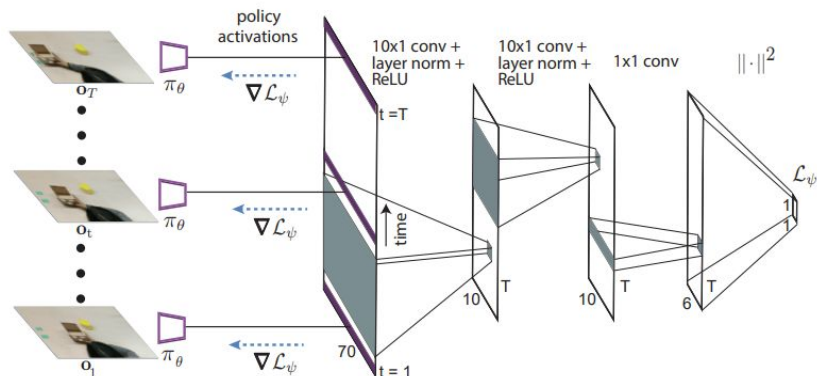
視覚によるbin pickingタスク

| Method | Dataset | Test | Bin emptying | | |
|--------------------|-------------------------------------|------------|--------------|------------|------------|
| | | | first 10 | first 20 | first 30 |
| QT-Opt (ours) | 580k off-policy + 28k on-policy | 96% | 88% | 88% | 76% |
| Levine et al. [27] | 900k grasps from Levine et al. [27] | 78% | 76% | 72% | 72% |
| QT-Opt (ours) | 580k off-policy grasps only | 87% | | | |
| Levine et al. [27] | 400k grasps from our dataset | 67% | | | |

- **【概要】**人のデモンストレーションを模倣する教示システムの提案
 - **【新規性】**人のデモからロボットが方策を推定するメタ学習システム
 - **【手法】**人・ロボットのデモの視覚的なドメインギャップを埋めるようメタ学習
 - **【実験結果】**マニピュレーションで新規デモに対しても方策を推定可能. 複数の画像フレームを考慮するよう学習すると精度向上
 - **【その他】**ロボットの方策を behavioral cloning で学習, BC-Z につながる. ロボットの動作データは3~8秒のデータがおよそ600~1300エピソード

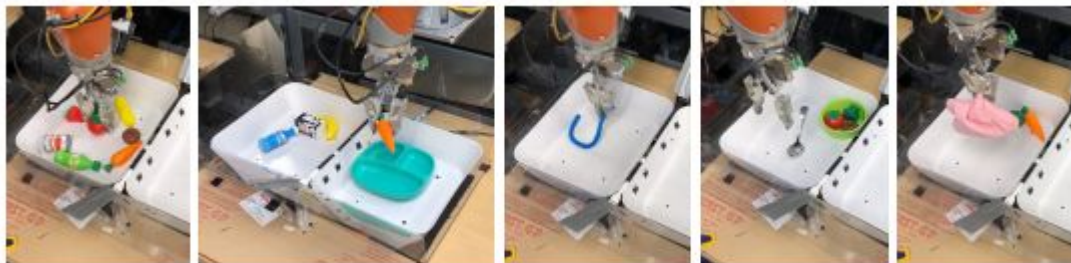
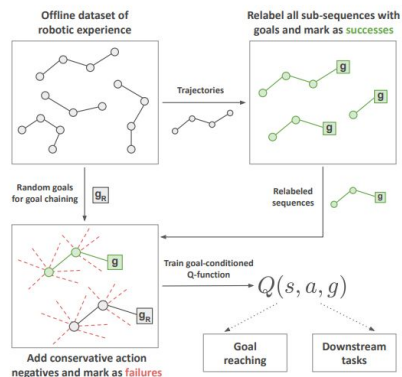


人・ロボットの視覚的ドメインギャップを埋める学習から学習後は人間のデモからロボットの行動方策を決定する.



人のデモで訓練するとき複数フレームを積みこんだ特徴を求める. この特徴でネットワークを正則化して人・ロボットのドメインギャップを埋める.

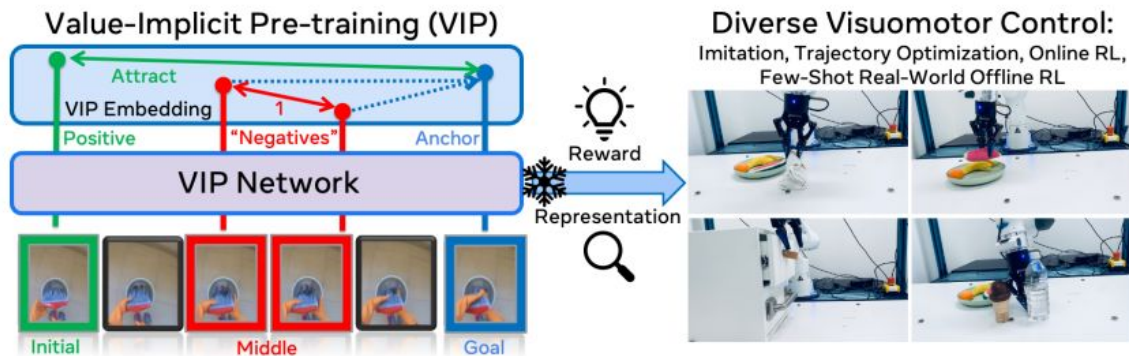
- **【概要】**教師なしオフラインデータによるスキルの獲得, 中間ゴールによる学習効率の向上
 - **【新規性】**手動アノテーションを用いずに過去のデータを再利用し, 新たにゴールを設定するモデルを構築.
 - **【手法】**オフラインのゴール志向方策および大域的な目的に合わせてサブタスクを組み合わせることで様々なタスクに適用.
 - **【実験結果】**特に下流タスクの獲得に有用. 事前学習の効果を確認.
 - **【その他】**Hindsight Experience Replay (HER)の類似. 元のデータセットから教師なし学習に発展させたことに着目.



【サブタスクおよび一連の動作をつなぎ合わせることで多くの作業に展開】

【既存のデータからゴール地点を再ラベル化】

- 【概要】人の作業映像を動作学習に応用できる特徴／報酬を推定する事前学習モデル
 - 【新規性】人の作業の表現学習をオフライン強化学習と捉えて、特徴空間内で状態価値を導出する
 - 【手法】作業画像の特徴空間内で目標画像との距離を評価、時間方向に対照学習する
 - 【実験結果】事前学習モデルから動作学習するとR3Mより高いタスク成功率を示した

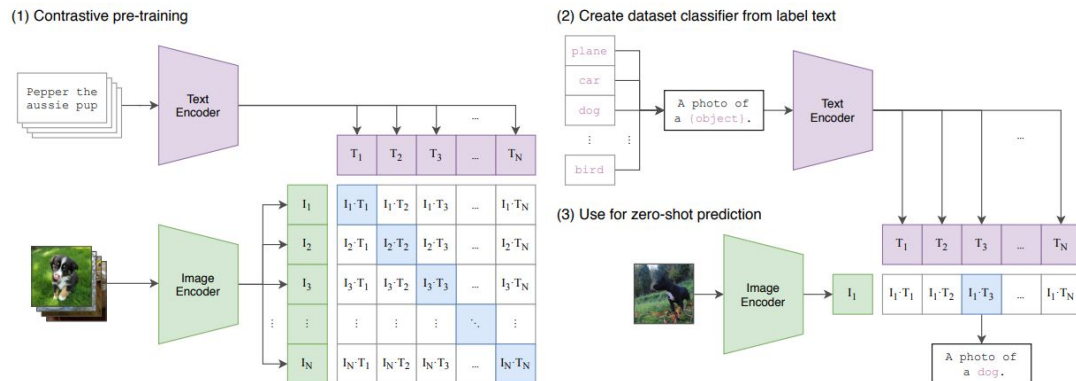


人が作業時の画像を特徴空間に埋め込み、目標画像の特徴との距離に基づいて対照学習する。大規模な作業映像データセット(論文では Ego4D)で事前学習することで、fine-tuning なしに報酬を推定できる。

| Environment | <i>Pre-Trained</i> | | | | <i>In-Domain</i> | | |
|-------------|--------------------|---------|---------|---------|------------------|---------|--------|
| | VIP-RWR | VIP-BC | R3M-RWR | R3M-BC | Scratch-BC | VIP-RWR | VIP-BC |
| CloseDrawer | 100 ± 0 | 50 ± 50 | 80 ± 40 | 10 ± 30 | 30 ± 46 | 0 ± 0 | 0* ± 0 |
| PushBottle | 90 ± 30 | 50 ± 50 | 70 ± 46 | 50 ± 50 | 40 ± 48 | 0* ± 0 | 0* ± 0 |
| PlaceMelon | 60 ± 48 | 10 ± 30 | 0 ± 0 | 0 ± 0 | 0 ± 0 | 0* ± 0 | 0* ± 0 |
| FoldTowel | 90 ± 30 | 20 ± 40 | 0 ± 0 | 0 ± 0 | 0 ± 0 | 0* ± 0 | 0* ± 0 |

事前学習したVIPでオフライン強化学習するとR3Mと比較して高い成功率を示した。

- 【概要】CLIP: 大規模な画像・テキストのペアを対照学習した V&L 基盤モデル
 - 【新規性】画像・テキストを別々にエンコードし、類似度を大きくするよう対照学習
 - 【手法】ウェブから4億組の画像・テキストペアを収集し、クリーニングしたデータセットで学習
 - 【実験結果】学習後はクラスをテキストに埋め込み、類似度スコアを評価することで zero-shot 画像認識を実現
 - 【その他】CLIP を事前学習モデルとして言語と視覚を結びつける手法が次々に登場



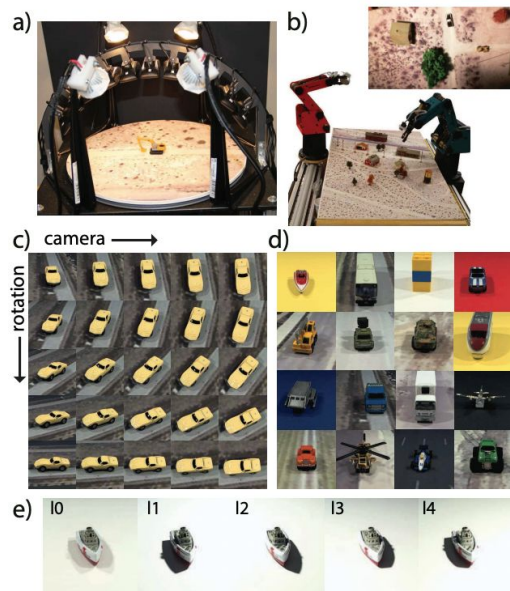
CLIP のシステム図. (1) 画像・テキストの組を各エンコーダでベクトル化, 内積が大きくなるよう学習する. (2) 画像認識にはクラスを埋め込んだテキストデータを作成, (3) 画像とテキストの類似度スコアでクラスを推定する.

- 【概要】多様な条件下での玩具の車画像によるRNN深層学習用大規模データセット
 - 【新規性】RNN深層学習に用いるための大規模データセットを構築
 - 【手法】15カテゴリ，88カメラ設置条件，5照明条件，3カメラ焦点設定，14-40背景画像。合計22百万画像のデータセットを構築
 - 【実験結果】下図参照

| train \ test | Without fine tuning | | With fine tuning | |
|--------------|---------------------|----------|------------------|----------|
| | Natural | iLab-20M | Natural | iLab-20M |
| Natural | 95 | 75 | 93 ↓ | 65 ↓ |
| iLab-20M | 78 | 97 | 70 ↓ | 100 ↑ |

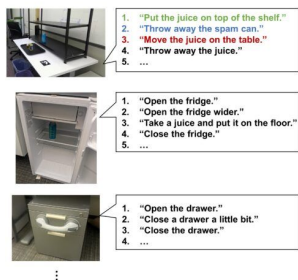
Table 3. Domain adaptation on boat vs. tank classification (in percentage).

| train \ test | Without fine tuning | | With fine tuning | |
|-------------------------------------|---------------------|-------------|------------------|-------------|
| | Natural | iLab-20M | Natural | iLab-20M |
| Natural [2000] | 96.48 (0.5) | 55.6 (2.7) | 95.56 (0.6) | 68.06 (2.0) |
| iLab-20M [2000] | 66.92 (3.2) | 96.90 (0.2) | 65.22 (1.4) | 99.72 (0.1) |
| iLab-20M [1000] + Natural [1000] | 94.42 (0.8) | 93.94 (0.4) | 92.52 (0.2) | 98.70 (0.2) |

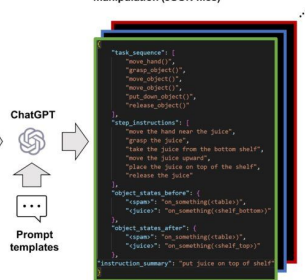


- **【概要】**ChatGPTで自然言語指示からロボットアクションシーケンスのJSONファイルを生成
 - **【新規性】**ChatGPTをロボット行動生成にいち早く応用
 - **【手法】**自然言語をアクションシーケンスに変更. Teach by showingで実動作推定
 - **【実験結果】**少数タスクで適切なJSONファイルを獲得できた
 - **【その他】**Teach by Showing ([2212.09242.pdf \(arxiv.org\)](#)) ありき

Multi-step instructions in various environments



Readable sequence of object manipulation (JSON files)



自然言語で指示:例)青いカップをとって, 赤い皿の上に置いて
 →JSON形式のアクションシーケンスを出力
 →視覚的なデモンストレーションで物理量を推定

Online robot teaching interface

ChatGPTの利用インターフェース例.

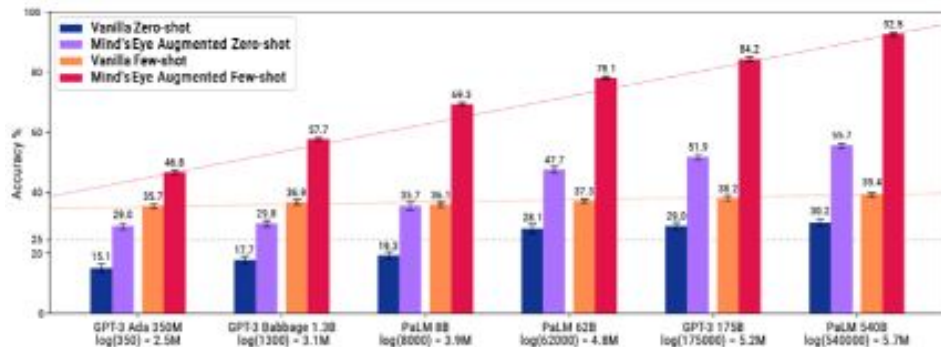
自然言語で指示し, 人間に視覚的にデモンストレーションしてもらうことで, ロボット実行に必要なパラメータ(=具体的アクション)を取得する.



- 【概要】**大規模言語モデル(LLM)が得意でない問題に対して, 外部知識(ロボティクスで重要となる物理エンジン)を利用することでLLMの推論能力を強化
 - 【新規性】**物理現象に関する質問に対し, 物理エンジンを使った試行の結果を解釈して回答
 - 【手法】**質問文から「物理エンジンMuJoCoのコード」を生成するモデルを学習
 - 【実験結果】**GPTやPaLMで3~4割の正答率にとどまる質問に対して, few-shot, PaLMベースで92.5%の正答率, ChatGPTで使われているInstructGPTベースでは99.1%の正答率を達成

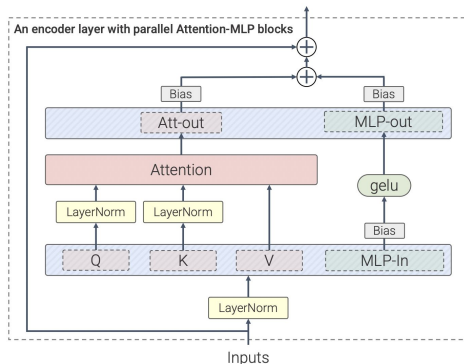


生成したMuJoCoコードには, シミュレーション結果において関心がある物理量(加速度や運動エネルギーなど)をテキストとして埋め込む。



物理量に関する質問からなるUTOPIAデータセットを提案。絶対量を問うのではなく, 大小など相対的な関係を問う。39タスク。通常のLLMはスケールしないが, 提案手法であるMind's Eyeはモデルパラメータ数に関してスケールする。

- 【概要】ViTのスケーリングに関して、パラメータ数を大きくするとさまざまタスクに関してLLMと同様なスケーリングの効果が見られた。
 - 【新規性】ViTのスケーリング
 - 【手法】学習を安定化させるためにQueryとKeyの正規化の組み込みを行った。また、大規模モデルを効果的に学習するための並列演算のテクニックを用いている。
 - 【実験結果】認識精度以外にも、公平性と性能のトレードオフの改善、形状/テクスチャの偏りに関する人間の視覚認識との最新の整合性、ロバスト性の改善が見られた



【Q,KのNormalizeを入れたViT】

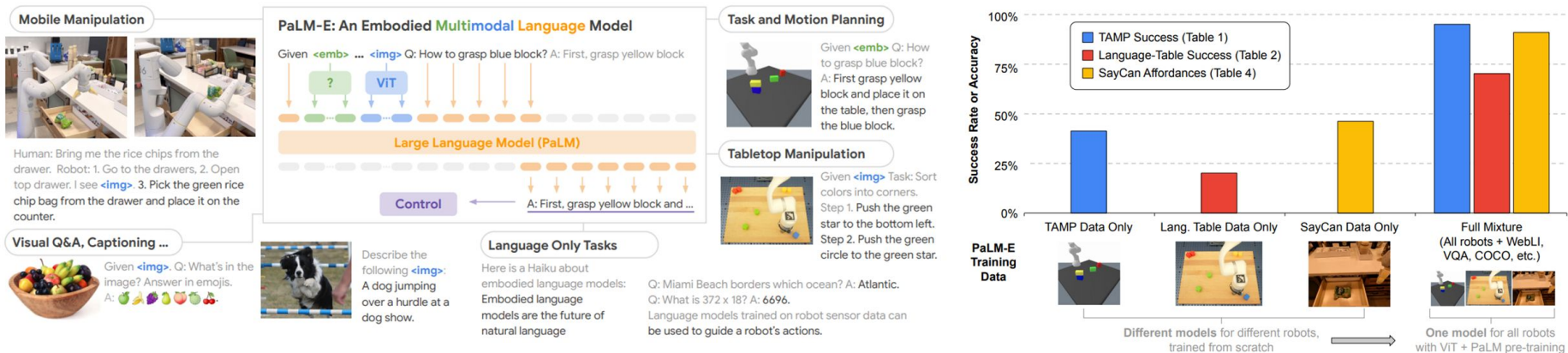
Table 1: ViT-22B model architecture details.

| Name | Width | Depth | MLP | Heads | Params [M] |
|----------------|-------------|-----------|--------------|-----------|--------------|
| ViT-G | 1664 | 48 | 8192 | 16 | 1843 |
| ViT-e | 1792 | 56 | 15360 | 16 | 3926 |
| ViT-22B | 6144 | 48 | 24576 | 48 | 21743 |

| Model | IN | ReaL | INv2 | ObjectNet | IN-R | IN-A |
|------------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| <i>224px linear probe (frozen)</i> | | | | | | |
| B/32 | 80.18 | 86.00 | 69.56 | 46.03 | 75.03 | 31.2 |
| B/16 | 84.20 | 88.79 | 75.07 | 56.01 | 82.50 | 52.67 |
| ALIGN (360px) | 85.5 | - | - | - | - | - |
| L/16 | 86.66 | 90.05 | 78.57 | 63.84 | 89.92 | 67.96 |
| g/14 | 88.51 | 90.50 | 81.10 | 68.84 | 92.33 | 77.51 |
| G/14 | 88.98 | 90.60 | 81.32 | 69.55 | 91.74 | 78.79 |
| e/14 | 89.26 | 90.74 | 82.51 | 71.54 | 94.33 | 81.56 |
| 22B | 89.51 | 90.94 | 83.15 | 74.30 | 94.27 | 83.80 |
| <i>High-res fine-tuning</i> | | | | | | |
| L/16 | 88.5 | 90.4 | 80.4 | - | - | - |
| FixNoisy-L2 | 88.5 | 90.9 | 80.8 | - | - | - |
| ALIGN-L2 | 88.64 | - | - | - | - | - |
| MaxViT-XL | 89.53 | - | - | - | - | - |
| G/14 | 90.45 | 90.81 | 83.33 | 70.53 | - | - |
| e/14 | 90.9 | 91.1 | 84.3 | 72.0 | - | - |

【Classification】

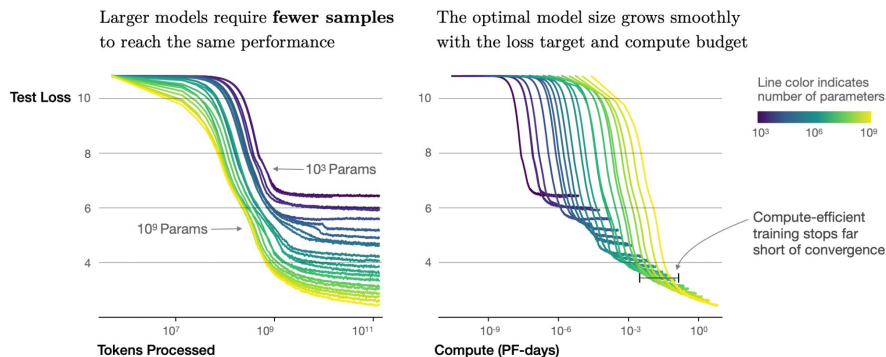
- **【概要】大規模言語モデル(LLM)を用いたロボットタスクモデル**
 - **【新規性】大規模言語モデルの言語空間にViTを用いて画像情報を埋め込むことに成功**
 - **【手法】ViTとPaLMの融合**
 - **【実験結果】OK-VQA でSOTA**



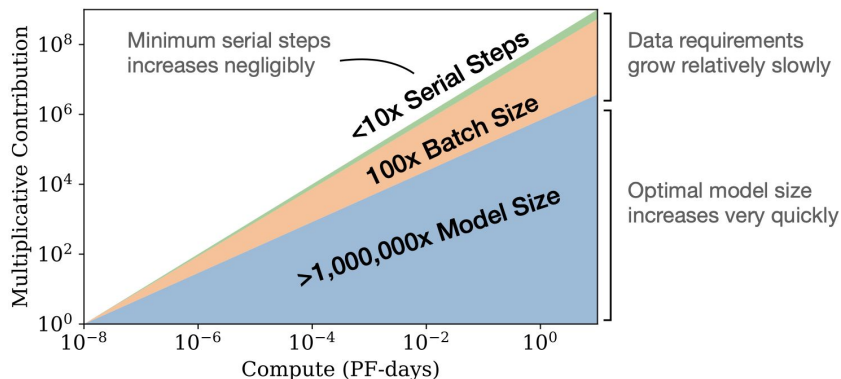
様々なモダリティーの画像をViTを介して大規模言語モデル空間(PaLM)に入れることで行動計画を行うことに成功している

3つの異なるロボットタスク。独自に学習するよりも言語モデルを介して統合した方が良い成績

- 【概要】LLMに関して、ある一定以上の巨大なモデル、データサイズ、計算資源で表されるスケーリングのべき乗則によって性能が向上する。
 - 【手法】テスト時のLossを性能と定義し、PF-days, token数, パラメータ数をBillionスケールまで増やしてWeb2Testデータセット検証する
 - 【実験結果】各パラメータ(モデル, データサイズ, 計算資源)と性能に関してそれぞれべき乗則を確認し, それぞれの組み合わせについても成り立つことがわかった
 - 【その他】



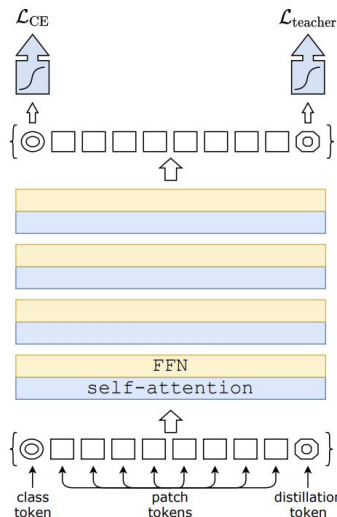
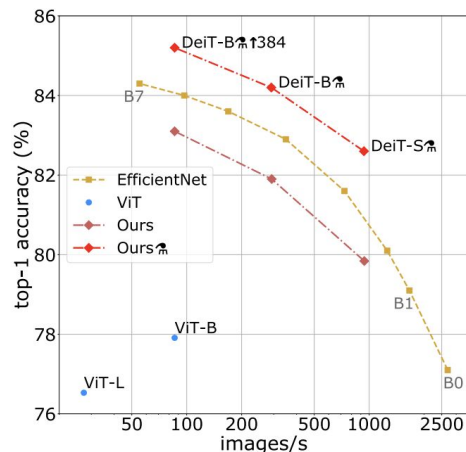
【トークン数と計算量を増やして行った時に、モデルのサイズを増やすとTest lossがどう変化していくか】



【モデルサイズと計算時間を増やすと性能が上がりやすいが、データに関しては並列計算の量が増えるだけなので比較的效果は少ない】

- **【概要】**DeepMindが提案した基盤モデル・Gopherの提案論文. DeepMindが収集したデータセットに対して, 4400万~2800億パラメータのTransformerモデルを適用して学習. 最大パラメータのモデルをGopherと呼称.
 - **【新規性】**
 - **【手法】**Massive Multitask Language Understanding (MMLU) ベンチマークにより学習され, 発表当時の既存言語モデルの性能よりも高いことを示した.
 - **【実験結果】**モデルパラメータを増加させることでゼロショットによる読解・ファクトチェック, 倫理解析などのタスクにおいて性能が向上することが明らかになったが, 一方で論理推論などについてはそれほど性能向上に寄与しないことがわかった. 対話により受け答えが一貫性を持つことも明らかにしている. また, 生成が繰り返されること, バイアスによる影響, 誤った情報も含まれることも改善点として指摘された.
 - **【その他】**下流タスクにおけるリスク分析も行われており, 実際に使用される際のリスクに対しても議論が行われている.

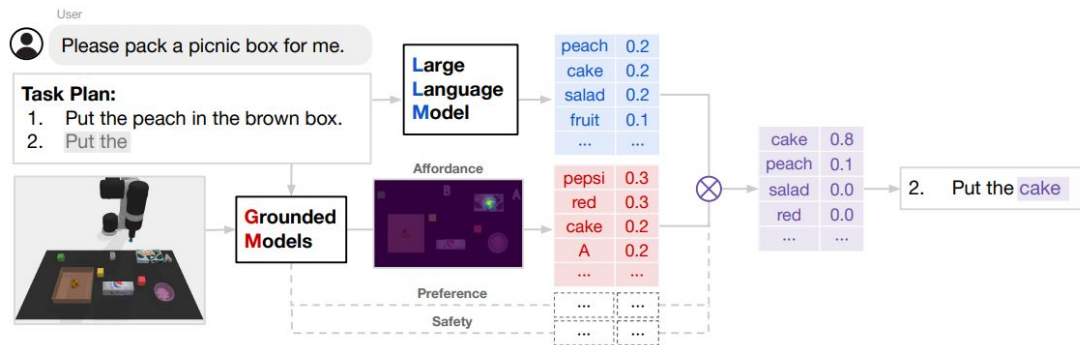
- 【概要】**従来, ViTは何百万枚もの画像での事前学習が必要であったが, 蒸留手法を用いることで, ImageNetのみで高精度な結果を達成するDeiTを提案した.
 - 【新規性】**蒸留トークンを採用することにより, 蒸留の損失を計算する.
 - 【手法】**ViTと同じ構造を用い, 蒸留手法を組み込み, データ拡張も行うことで, 効率的な学習を可能にした.
 - 【実験結果】**ImageNetのみの学習で最新のモデルと同等の精度を達成し, さらに学習コストも大幅に下げた.



【蒸留トークンを用いることで, 精度が向上】

【ImageNet top-1 accuracyで高い精度を達成】

- 【概要】LLMと環境に接地したモデル (grounded model, GM) を組み合わせた行動系列の計画
 - 【新規性】LLMの目的指向な推論をロボット環境に接地するアルゴリズムの提案
 - 【手法】LLMや複数のGMの提示から同時確率をもとに対象トークンを決定する
 - 【実験結果】GMと組み合わせると長期かつ複雑なタスクの計画で高い成功率



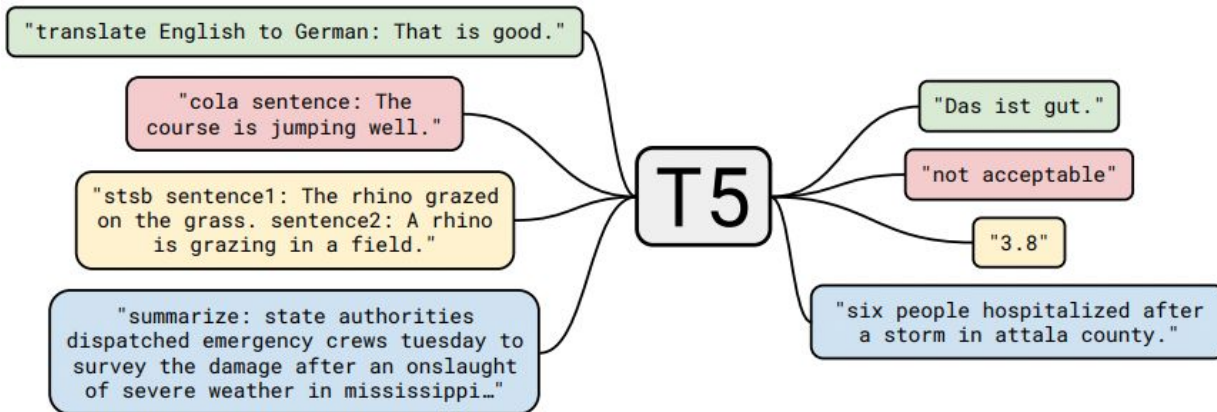
LLMは指示に対する目的指向な提案を行う。加えて、環境に根差した複数のGM (affordance, preference, safety など) の提案から得られる同時確率を基に対象トークンを決定する。

| | CLIPort | | +LLM | +GD | |
|---------------------------|---------|------|------------|------------|------------|
| | Short | Long | Ungrounded | Greedy | Beam |
| Seen Tasks | | | | | |
| Letters | 7% | 40% | 20% | 43% | 57% |
| Blocks & Bowls | 2% | 62% | 35% | 60% | 77% |
| Box Packing* | 15% | 28% | 11% | 79% | 78% |
| Unseen Tasks | | | | | |
| Letters | 6% | 10% | 19% | 37% | 41% |
| Blocks & Bowls | 6% | 10% | 28% | 44% | 50% |

Tabletop におけるタスクを比較すると、ベースライン (CLIPort: LLM なしでプランニング) は未学習のシーンに不安定。接地モデルを含むモデル (+GD) は LLM のみで推論した場合と比較して長期かつ未学習のシーンで高い成功率を示した。

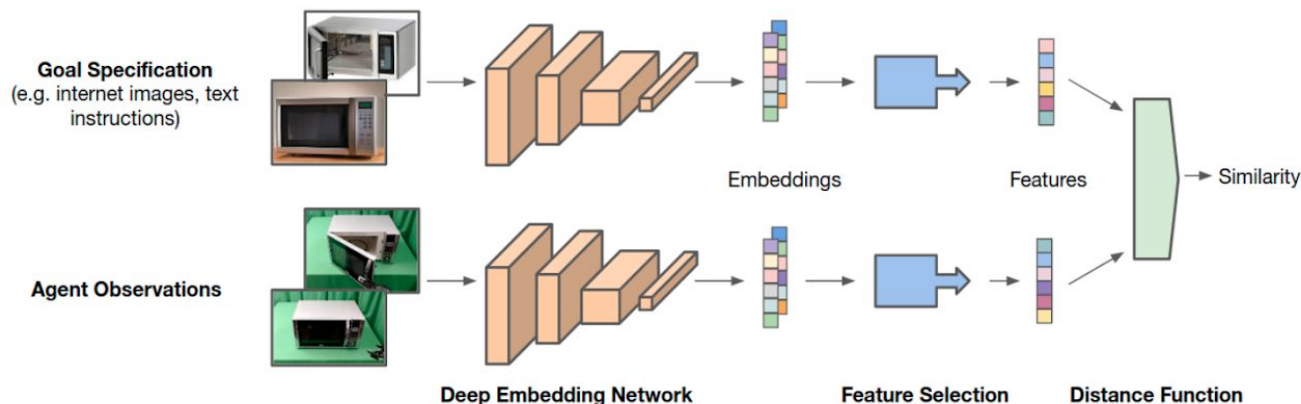
- 【概要】Text-to-Text Transfer Transformer (T5) の提案論文

- 【新規性】テキスト同士の入出力でNLPタスクの転移学習を高度化したこと。
- 【手法】テキスト同士の入出力形式で言語タスクを行い事前学習。要約・翻訳などを統一的な枠組みで扱うことに成功した。
- 【実験結果】要約, 質問回答, テキスト識別, 翻訳などにより有効性を確認した (WMT, GLUE, SQuAD, RACEデータセットで検証)。
- 【その他】単一モデル・複数タスクに対応。BERTの提案後に発表されているため, MLMに関する検討も含まれている。



T5のフレームワーク。翻訳, 質問回答, 識別, 要約などをモデルで処理させることにより転移学習のための言語モデルの事前学習ができることを明らかにした。

- 【概要】ZeSTスコアを報酬関数の代理として使用することで、行動クローンベースラインよりも優れたパフォーマンスを発揮するポリシーの学習が可能。
 - 【新規性】ロボット操作タスクのゼロショットタスク指定を可能にする
 - 【手法】基盤モデルの研究のためのフレームワークZeSTを提案.
 - 【実験結果】ランダム推測のベースラインと比較して14倍の性能向上が得られる.
 - 【その他】ロボットの動作成功試行では、観測画像とタスク内容の類似性が高まっていることが観測できた



● 【概要】人との対話による閉ループフィードバックを導入したLLMベースの行動計画手法

- 【新規性】アクション間のフィードバック(LLMのプロンプト/人間の回答)の追加により, より詳細なシーンの制御が可能に.
- 【手法】成否判断, 受動的記述, 能動的記述の3種のフィードバックを含む対話形式のモノローグを構成し, リアルタイムにロボットの動作を計画・修正.
- 【実験結果】物体及びシーンを含む本手法は「現在の理解」「長期的な記憶」に有用性
- 【その他】対話から自発的に代替案を提案するなど言語による応用の範囲が広い

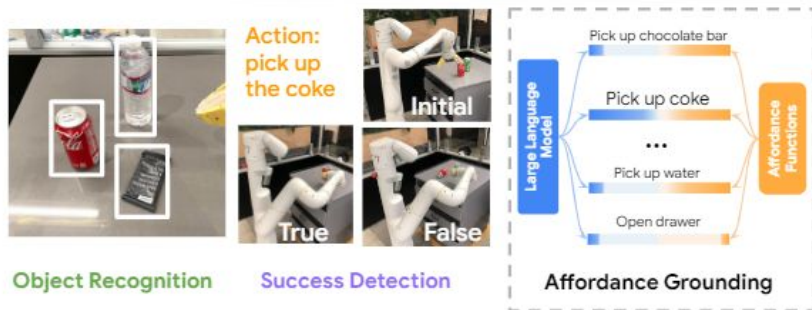


Active Scene Description in context:

Robot Action: Go to drawers
 Robot Ask: Is the drawer open?
 Human: The drawer is closed.
 Robot Action: Open the drawer

Kitchen Mobile Manipulation (Real)

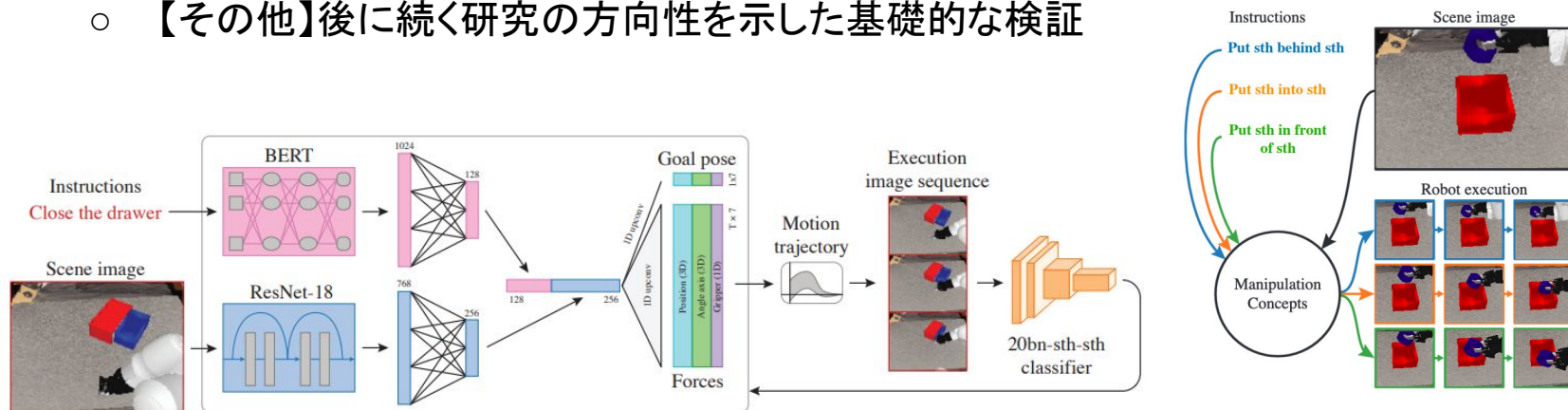
Human: Bring me a drink from the table
 Robot: go to the table
 Scene: I see a coke and lime soda
 Robot: What kind of drink would you like?
 Human: Something with caffeine
 Robot: Pick up coke
 Success Detector: False
 Robot: Pick up coke
 Robot: Bring it to you
 Robot: Put down the coke
 Robot: Done



【直接的なシーンに関する質問によりフィードバックを行う】

【物体識別による環境の理解から, クエリを構成し, 必要なフィードバックを人から直接的に得ることで, 複雑なタスクにも対応できる】

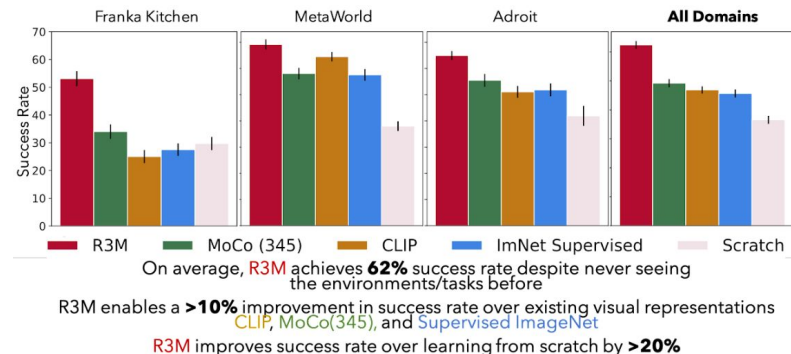
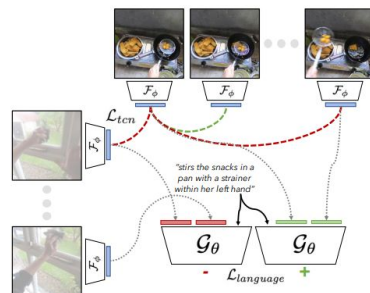
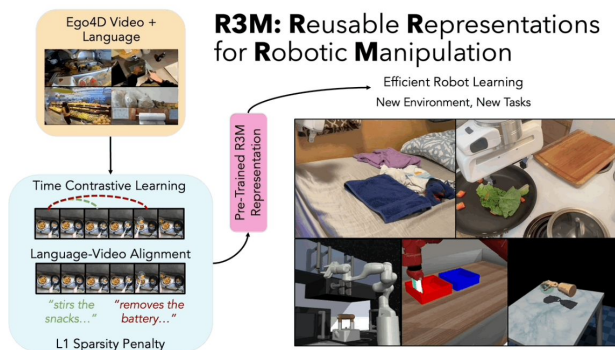
- **【概要】**視覚的教示および言語による指示を組み合わせたマニピュレーションの学習
 - **【新規性】**人のデモ動画+言語から行動を得るための学習モデルを提案
 - **【手法】**デモ動画をタスクごとに分類. 78のタスクをベースに個別には強化学習にて, 複合したタスクは模倣学習により学習. これらは画像と言語による入力からニューラルネットワークで最終的な出力が得られるようにモデルを設計. cf. DMP
 - **【実験結果】**未知の指示に対する動作の生成の可能性を示した.
 - **【その他】**後に続く研究の方向性を示した基礎的な検証



【フレームワークの全体図-画像および指示から7dofの終端姿勢おとび力を表現するベクトルを出力】

【自然言語を含む人のデモ動画によりmanipulation conceptsを獲得. 本モデルはロボットの動作軌道が出力可能】

- 【概要】事前学習済みの視覚表現はロボットのマニピュレーションの学習効率を高める
 - 【新規性】既存の認識モデルを組み合わせることで、サブタスクの学習効率が向上
 - 【手法】時間対比学習, キャプション付き動画の学習, スパースな誤差関数を組み合わせることで, 基盤モデルであるR3Mを構築.
 - 【実験結果】R3Mをベースにした模倣学習が非常に効率的であることを示した
 - 【その他】効果は模倣学習などの小規模なタスクを目的とした場合に限定されるが, 単なる画像への他の情報の埋め込みの可能性を示した.



【既存のビデオによる学習から認識に重要な前処理が行われる】

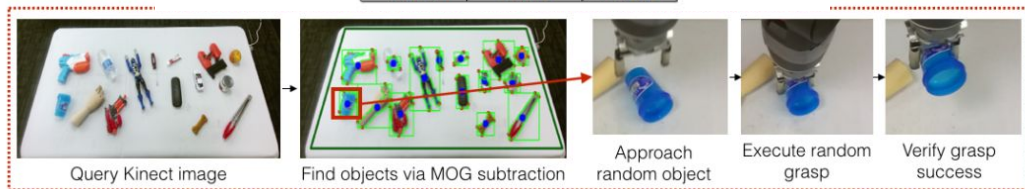
【時間の流れと対応するキャプションから空間にEmbeddingするイメージ】

【他の優れた手法による認識に比べて模倣学習の効果がもっとも良いことを示した】

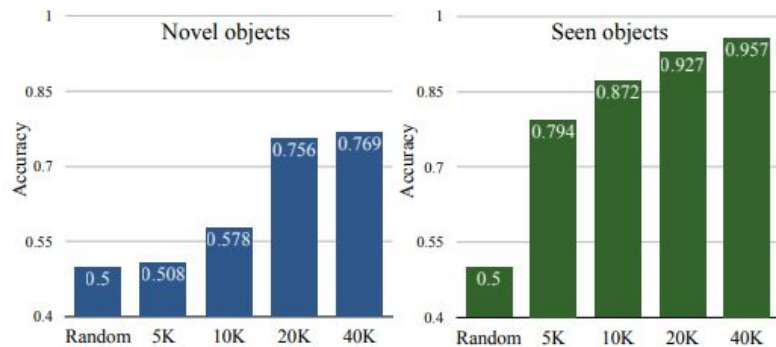
- 【概要】大量のデータセットを用いた学習による把持位置予測
 - 【新規性】700時間・5万回のデータセット。
 - 【手法】大規模場把持試行データを基に画像パッチに対する多段階学習モデル。
 - 【実験結果】訓練用のデータ量を多ければ認識の正確さも向上（未知/既知ともに）
 - 【その他】2016年当時のロボット学習のScaling Law？画像をパッチに分割し把持の向きを離散的に分けて判断する多段階の行動計画を採用したことも効果の一因。

GRASP DATASET STATISTICS

| Data Collection Type | Positive | Negative | Total | Grasp Rate |
|----------------------|--------------|---------------|---------------|------------|
| Random Trials | 3,245 | 37,042 | 40,287 | 8.05% |
| Multi-Staged | 2,807 | 4,500 | 7,307 | 38.41% |
| Test Set | 214 | 2,759 | 2,973 | 7.19% |
| | 6,266 | 44,301 | 50,567 | |



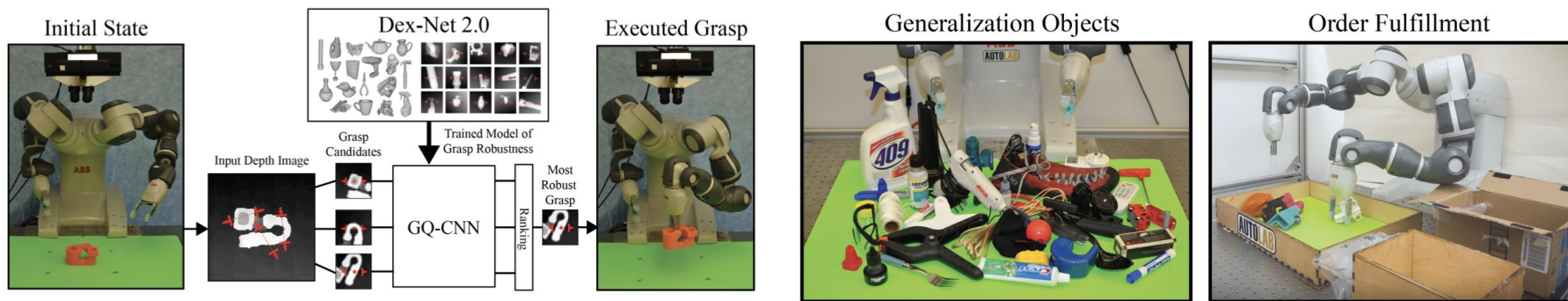
【上記のランダムな把持データを実際のロボットを用いた実験を通じて取得している】



【データセットサイズ別のAccuracyの比較. データ量に対する相関関係は明らかである.】

- 【概要】シミュレーションで生成した大量の画像データを学習して幅広い単一対象物を実環境でピックアップする

- 【新規性】シミュレーションのみを用いて画像からの把持のデータを収集して学習した
- 【手法】サンプリングされた把持位置とそのスコアをペアとしてデータを自動生成し、把持性能を推定するGQ-CNNを学習させる
- 【実験結果】従来の点群ベースの方法より高速で高い成功率でピックアップに成功した
- 【その他】実環境で動くシミュレーションベースの把持の学習で最初の論文？

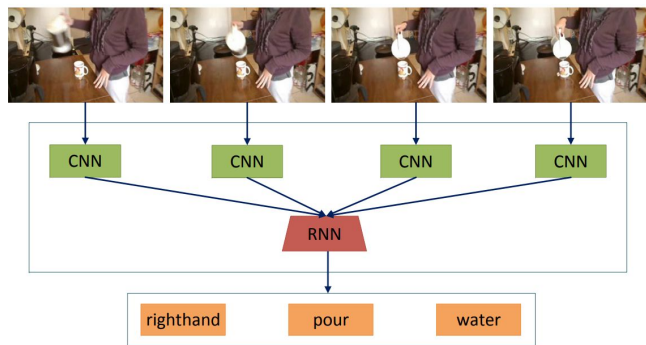


【大量の3Dモデルを利用して、シミュレーションの画像データと把持性能のペアを生成し、モデルを学習したのちに実環境で動作させる】

【学習したデータによって現実の様々な形状の物体に対してピックアップを可能にしている】

● 【概要】人のデモ動画をロボット用のコマンドに変換する手法

- 【新規性】動画から一連の構文を抽出することでロボットに有用な情報を提供.
- 【手法】CNNによる視覚情報のエンコード. フレーム間を二層のRNNにより解析し視覚的な特徴とロボット用のコマンドの関連性を学習.
- 【実験結果】人型ロボットでデモ動画からの操作生成を簡単に適用.
- 【その他】Transformerは流行前の方法であるため, VGGやResNetなどモデルの選定が重要な点に注目. 論文内でモデル別の効果などについて検討が行われている.



GT: righthand carry spatula
Ours: righthand carry spatula

SGC: lefthand reach stove
S2VT: lefthand reach pan



GT: righthand cut fruit
Ours: righthand cut fruit

SGC: righthand cut fruit
S2VT: righthand cut fruit



GT: righthand crack egg
Ours: righthand carry egg

SGC: lefthand reach spatula
S2VT: righthand carry egg



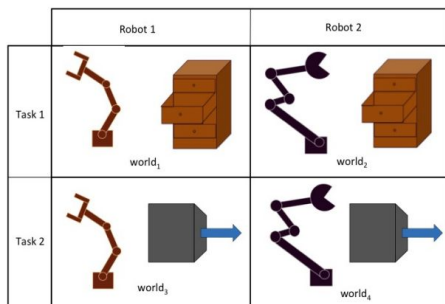
GT: righthand stir milk
Ours: righthand hold teabag

SGC: righthand place kettle
S2VT: righthand take cacao

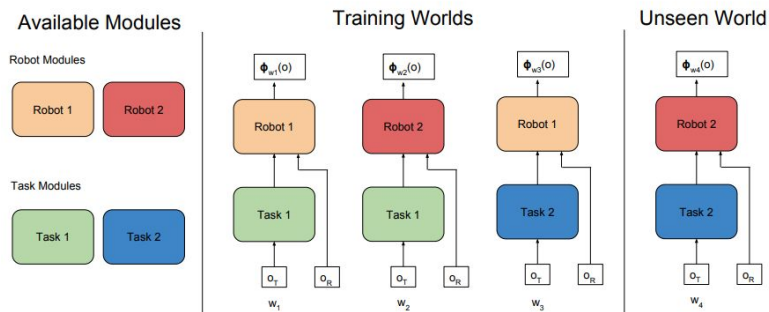
【動画に対するラベルの結果】

【フレーム毎の視覚情報のエンコードとそれらからRNNを用いてラベルとなる構文を解析している】

- 【概要】**タスク別, ロボット別にモデルをモジュール化した方策の学習を展開. マルチタスク強化学習に効果的であることを示した.
 - **【新規性】**タスクorロボットの共通する特徴の抽出を想定した転移学習.
 - **【手法】**モジュール化した構築. あるタスクについてロボット別に学習を実施し, 特徴の共有を目指している.
 - **【実験結果】**4DoF のロボットでの検証では未知のロボットとタスクの組み合わせにも対応する zero-shotの結果を示した.
 - **【その他】**少ないデータも, 転移学習の工夫からzero-shotを目指せる工夫.

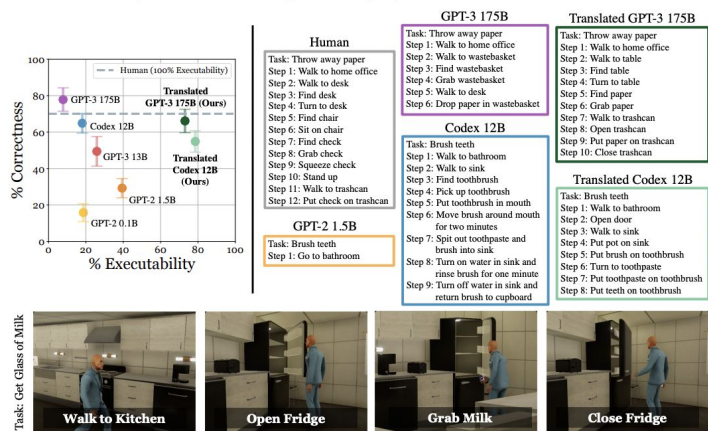


【ロボットやタスクが変わると強化学習でも, データセットの拡充が必要になってしまう】



【モデルの時点でタスク別・ロボット別にモジュール化. 組み合わせで表現し, 未知のタスクへの対応を示す】

- 【概要】LLMを使って事前学習なしでAgentの行動シーケンスを生成する
 - 【新規性】言語で表現できるハイレベルな行動タスクをLLMを使ってZero-shotで行動可能なシーケンスに分割できるかの調査
 - 【手法】分割されたステップを既存のデモ行動との類似度から選択して実行する
 - 【実験結果】LLMのモデルサイズが大きければハイレベルなタスクに対してより実行可能な行動シーケンスを生成できる



【正確な行動シーケンスをLLMで生成し、シミュレーションでエージェントの行動を実行する】

| Methods | Executability | LCS |
|-----------------------------|---------------|---------------|
| Translated Codex 12B | 78.57% | 24.72% |
| - w/o Action Translation | 31.49% | 22.53% |
| - w/o Dynamic Example | 50.86% | 22.84% |
| - w/o Trajectory Correction | 55.19% | 24.43% |
| Translated GPT-3 175B | 73.05% | 24.09% |
| - w/o Action Translation | 36.04% | 24.31% |
| - w/o Dynamic Example | 60.82% | 22.92% |
| - w/o Trajectory Correction | 40.10% | 24.98% |

【アブレーション: Action Translationを省くと大きく性能が低下することから、行動と言語の意味的変換が最も重要であることがわかった】



メタサーベイ

メタサーベイ 目次

- Google DeepMindの軌跡
- sim-to-realの動向
- 学習方法
- データ作成方法
- モデル構造



Google DeepMindの軌跡 (1/4)

機械学習とロボット工学のシナジーによるAGI開発と知能の解明へ

- Google ResearchのBrainチームとDeepMindを統合し「Google DeepMind」を設立

Google DeepMind: Bringing together two world-class AI teams

Apr 20, 2023 · 2 min read

Share

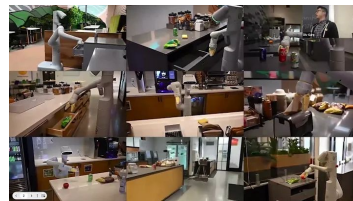
<https://blog.google/technology/ai/april-ai-update/> (2023/05/25 時点)



Demis Hassabis @demishassabis · 4月17日

I believe AGI will be one of the most important inventions humanity will ever make. Thanks to @60Minutes for visiting @DeepMind last month and learning about how we're working with @Google to drive AI breakthroughs in a bold and responsible way.

<https://twitter.com/demishassabis/status/1647893015032365057?s=20> (2023/05/25 時点)



参考: Stanford Seminar
<https://youtu.be/ct4tdyyNDY4>



Co-founder & CEO
Demis Hassabis



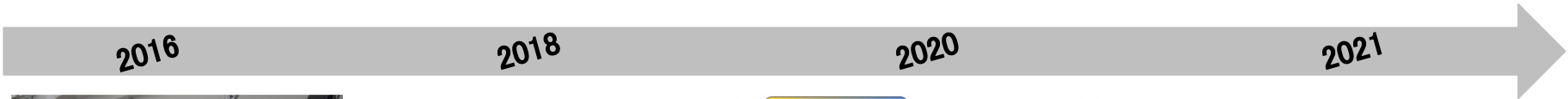
robotpaper.challenge



cvpaper.challenge

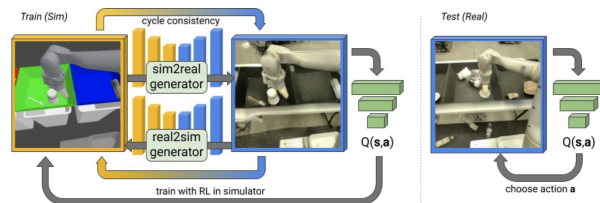
Google DeepMindの軌跡 (2/4)

Google DeepMind (旧Google Research) におけるRobotics研究動向



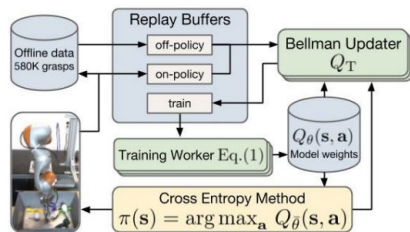
Arm Farm [Sergey Levin+, IJRR2018]

深層学習によるRobot Pickingの火付け役



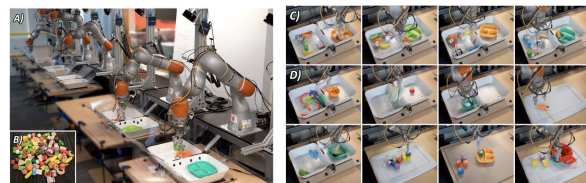
RL-CycleGAN [Kanishka Rao+, CVPR2020]

CycleGANを活用したSim2Real



QT-Opt [Dmitry Kalashnikov+, arXiv]

Experience Replayに過去の学習時における経験データを活用



MT-Opt [Dmitry Kalashnikov+, arXiv]

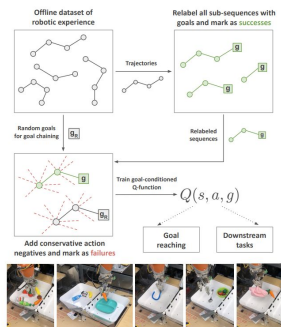
複数タスクの経験データを複数ロボットから同時収集可能なフレームワーク



Google DeepMindの軌跡 (3/4)

Google DeepMind (旧Google Research) におけるRobotics研究動向

2021



AM-Z [Yevgen Chebotar+, arXiv2021]

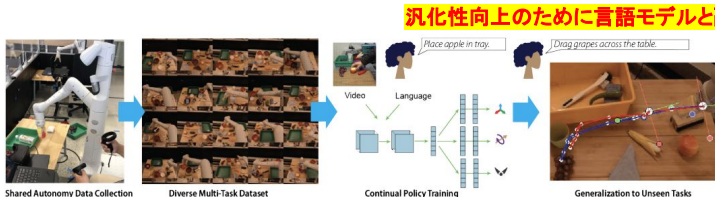
ユーザが報酬を設定なしに、大規模マルチタスクDBを Goal-reaching Policyにて事前学習可能に

2022



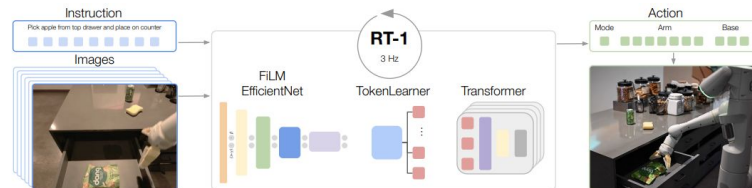
PaLM-SayCan [Michael Ahn+, arXiv2022]

周囲環境を理解した上でユーザからの指示文に対して最適な行動を実現



BC-Z [Eric Jang+, CoRL2021]

言語モデルを利用して学習データに含まれない未知タスクへ汎化



RT-1 [Dmitry Kalashnikov+, arXiv]

ロボティクスにおいてタスク・環境に依存せず、ゼロショットを実現する初の基盤モデル



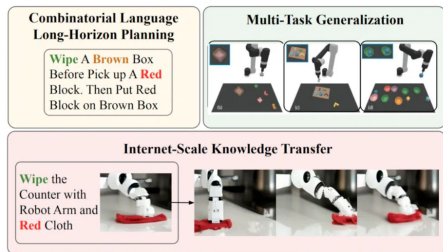
Google DeepMindの軌跡 (4/4)

Google DeepMind (旧Google Research) におけるRobotics研究動向

2023

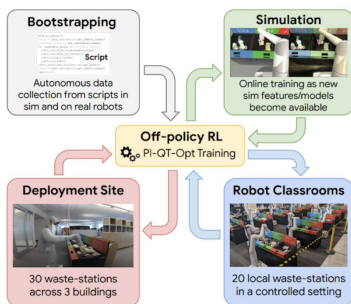
2024

2025

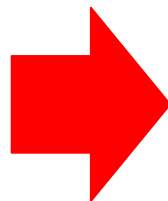


UniPi [Alexander Herzog+, arXiv2023]

テキストガイドを利用して異なる環境&異なるタスクにて動作



RLS [Alexander Herzog+, arXiv2023]



What's next?



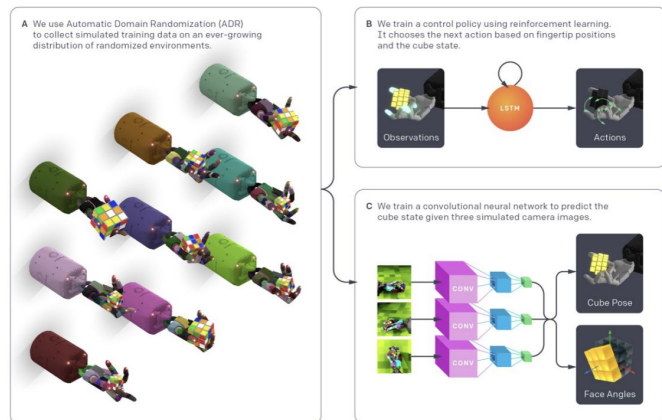
sim-to-real (1/5)

sim-to-realのブームは過ぎたのか？

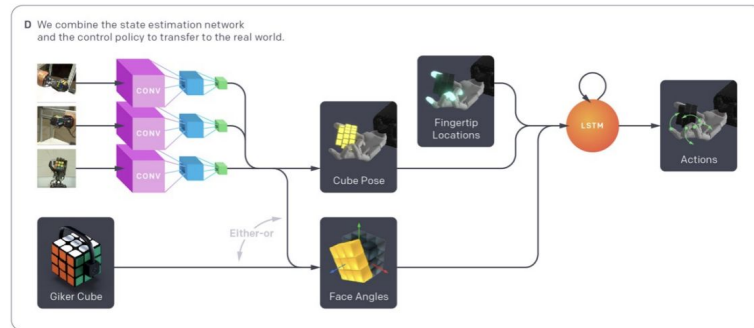
- OpenAI [Akkaya+, arXiv2019]
 - 2019年にシミュレータを使った強化学習でルービックキューブの in-hand manipulationを実現し話題になった
 - しかし、2021年にロボットから撤退

- 膨大なドメイン乱択化
- 視覚・物理・システムパラメータ
- 乱択化範囲の自動調整 (カリキュラム学習)

Train in Simulation



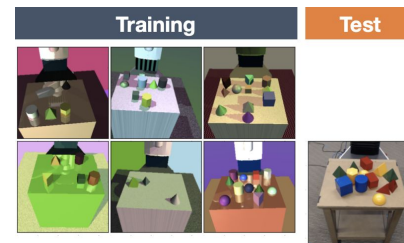
Transfer to the Real World



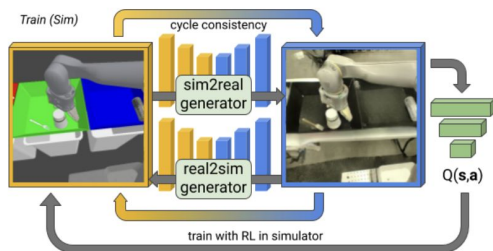
sim-to-real (2/5)

ブームは過ぎたのか？ (GoogleとUC Berkeleyの取組み)

- sim-to-realと強化学習を進めていた
 - ドメイン乱択化のロボットマニピュレーションへの応用
 - ドメイン適応の強化学習における利用

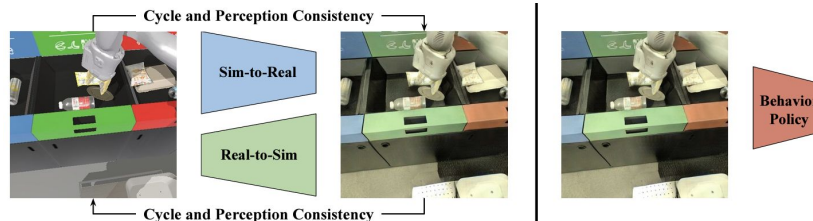


非現実的なくらい大胆な乱択化
[Tobin+, IROS2017]



RL-CycleGAN [K. Rao+, CVPR2020]

ドメイン変換と強化学習タスクを一緒に学習させることでタスクに有用な情報を保存



RetinaGAN [D. Ho+, ICRA2021]

ドメイン変換と物体検出を一緒に学習させることで物理解解に必要な情報を保存

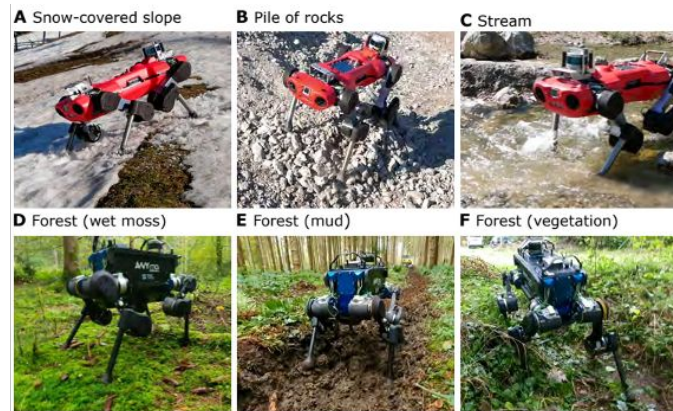
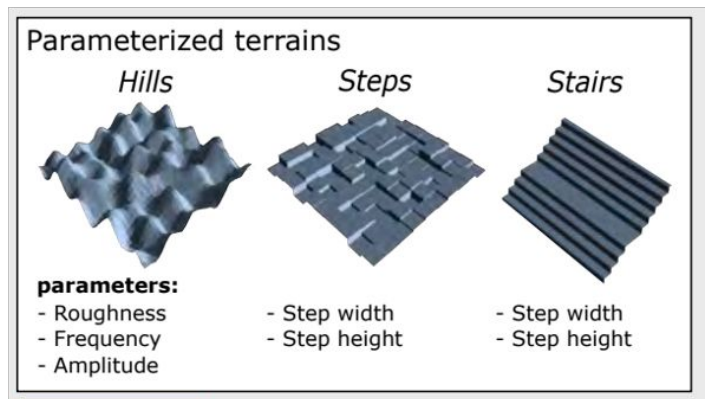
- 近年のRT-1では、実ロボットでデータを集める方向に向かっている



sim-to-real (3/5)

simulator \neq realが必須とは限らない

- 4足ロボットの不整地歩行 (ANYmal), ドメイン乱択化の成功例の1つ
- simulationを实世界に近づけなくとも, 実世界でロバストなコントローラを学習可能な点が興味深い



学習は, 剛体の地面, 少数の手続き的に生成した地面
[J. Lee+, ScienceRobotics 2020]

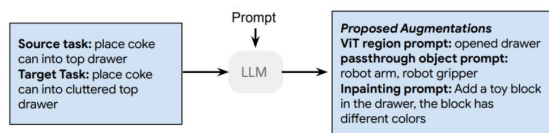
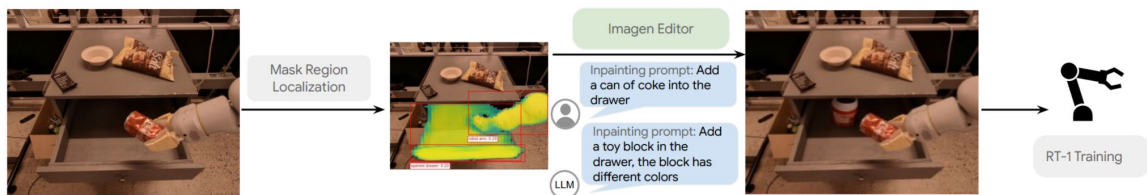
変形する複雑な地面で歩行が可能



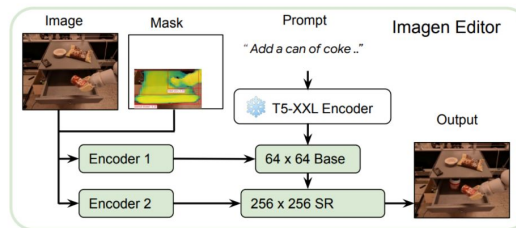
sim-to-real (4/5)

実画像への拡張

- シミュレータにおける乱択化でなく、**実画像を意味的に拡張**する研究も複数出てきている
 - Diffusion modelを利用 (ROSIE, CACTI (Meta), GenAug)
 - ○ 実画像に近いdemonstrationと異なるシーンを生成可能
 - × 動作の多様性を増やすことはできない, Diffusion Modelはデータ生成コストが大



LLM-assisted Augmentation Prompt Generation

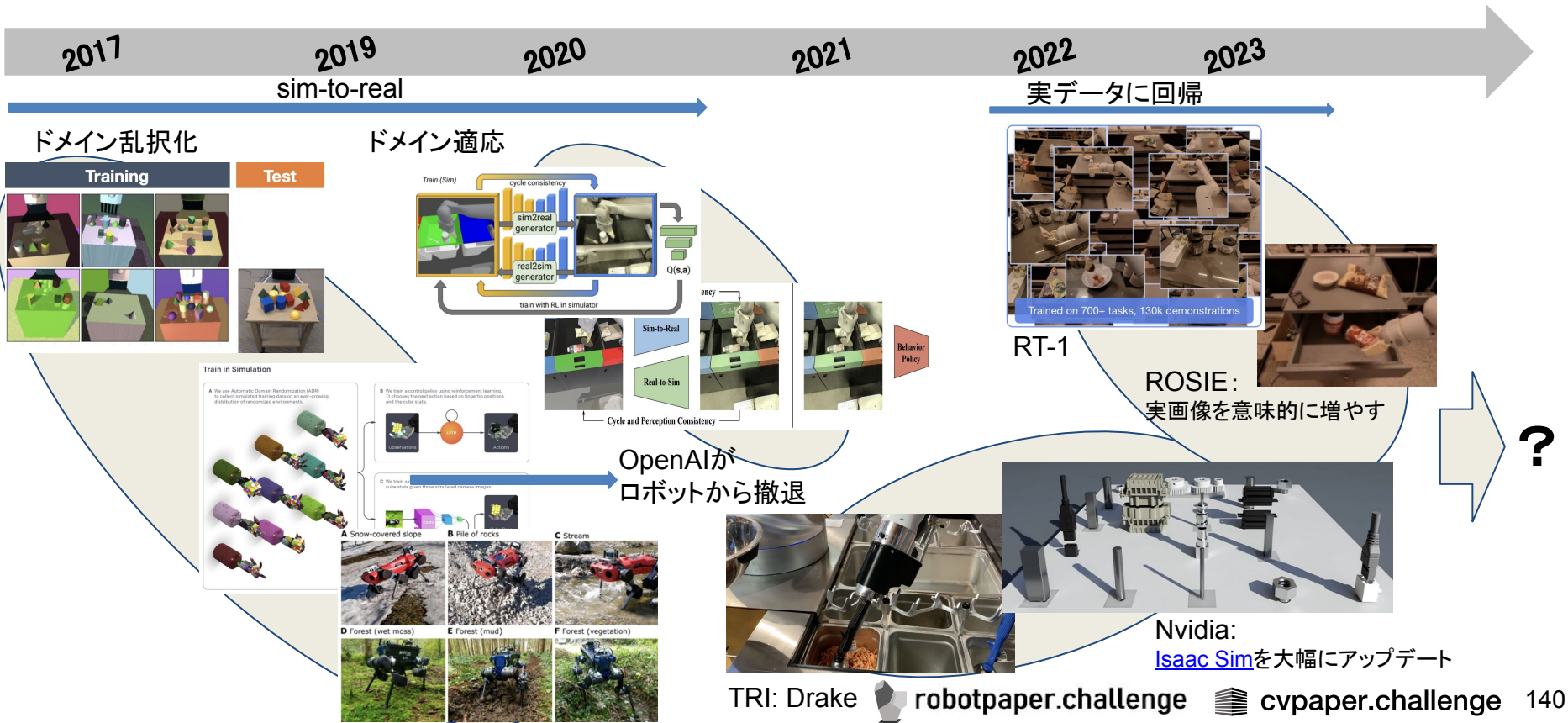


ROSIE [Google, RSS2023]
LLMとdiffusion modelを用いた
意味的データ拡張



sim-to-real (5/5)

シミュレーションをベースにした大規模モデルは来るか？

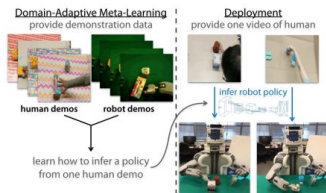


?

学習方法: オンライン学習からオフライン学習へ

タスクの複雑化・データスケールアップ

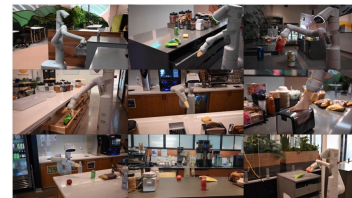
2018年頃から
模倣学習手法が
増える



one-shot imitation (2018)



BC-Z (zero-shot generation, 2022)



RT-1 (2022)

2016



Arm Farm
(on-policy RL, 2016)



QT-opt
(offline data & on-policy
RL, 2018)



MT-opt
(off-policy RL, 2021)

強化学習手法はon-policy から
off-policy に

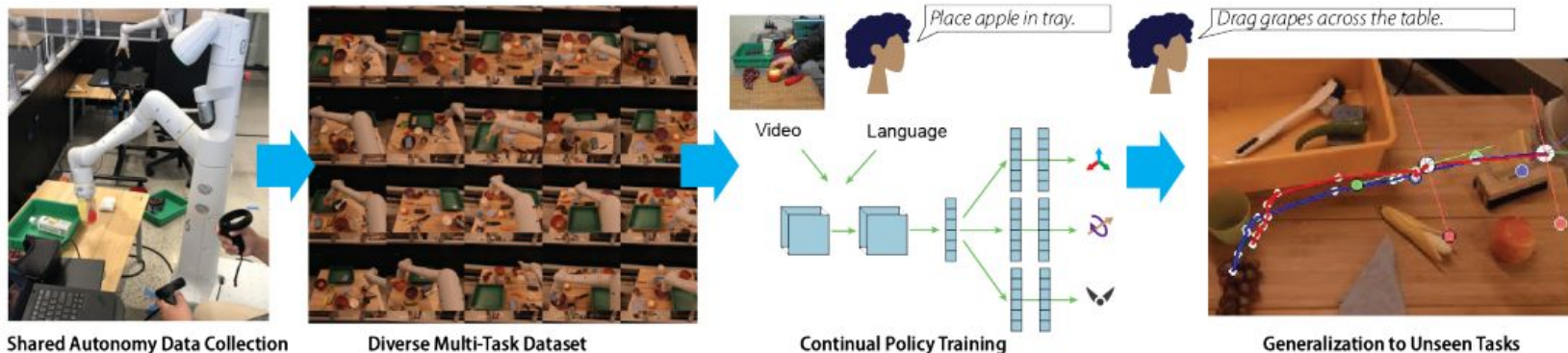
2023



学習方法: 模倣学習によるオフライン学習

Imitation learning / behavioral cloning

- ロボットから画像や関節角・速度・トルクなど運動データを記録, データを再現するように学習する
- ユーザによる直接教示や遠隔操作で画像や運動データを収集する
- どのように複雑な学習データやタスクバリエーションを増やすか?

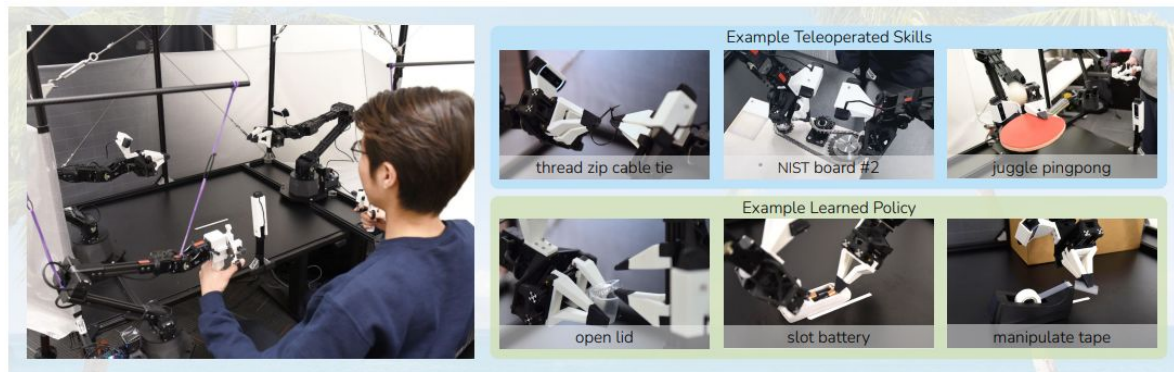


BC-Zではユーザの遠隔操作で学習データを収集する。ビデオと言語指示で条件づけられたネットワークがロボットの運動情報を再現するように模倣学習。

データ作成方法: 効率的な模倣学習データ収集

ALOHA [Zhao+, 2023]

- Leader-followerで双腕を遠隔操作するハードウェアシステム
 - 低コスト (\$20kほど) ながら細かい手作業が可能
 - ユーザがfollowerを一人称視点で操作できる設置が学習データの収集に効果的
- 4台のカメラで画像を撮影
 - 学習に扱うデータとしてビジョン・モータのみのモダリティで十分か？

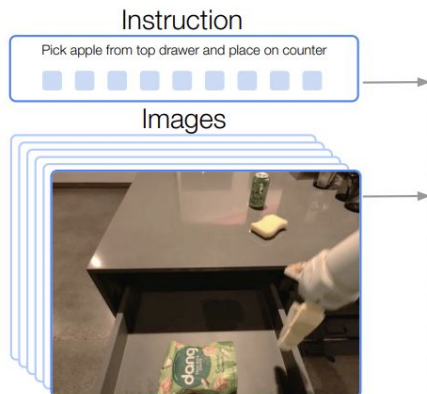


データ作成方法: どんなデータが使われるか

ロボットの一人称 or 俯瞰画像・制御入力+何か

制御入力は

- マニピュレーション: 関節角度, 角速度, 手先位置姿勢など
- 移動ロボット: ベース速度など
- さらにマルチモーダルなセンサ情報, 言語インストラクションやアクションなどが加えられる
- この新たなモーダルをどのように選択し, どのように教示・獲得して学習するかが重要



[RT-1](#) [Brohan+, arXiv2022]



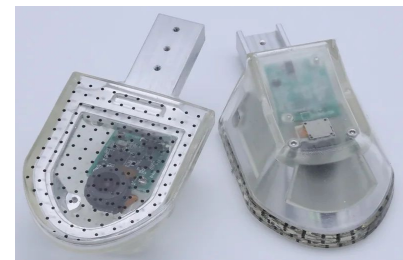
[Tian+, ICRA2019]



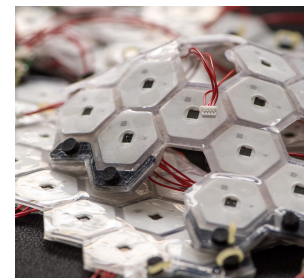
データ作成方法: どのように計測するか

センサによって直接に計測・間接的に推定

- ベースとなる関節角度・手先位置姿勢・ベース速度などモータ動作と構造に起因する情報
 - 回転軸に取り付けるエンコーダ: 高分解能, 高精度, 低ノイズ
 - インクリメンタル or アブソリュート
- 画像情報
 - RGBカメラ(画質と情報量とのトレードオフ)
- 接触情報
 - 力覚センサ(歪ゲージ, 静電容量)
 - 視触覚センサ(ロボットハンドの指先に追加可能な画像を用いた接触推定)
- 非接触(近接, depth)情報
 - ステレオカメラ(2画像の視差から距離を推定する)
 - ToF(赤外線反射がはね返ってくる時間)
- その他
 - 表面接触, 温度, 加速度



[FingerVision](#)

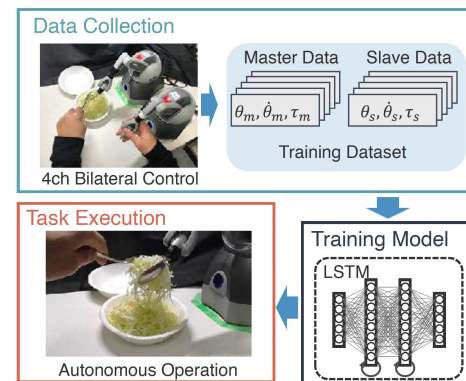


Robot Skin [Cheng+, Proc. IEEE 2019]

データ作成方法: 人デモによる教師データの獲得

手軽に多様なデータを集めるインターフェイス

- 手先運動を教示するためのインターフェイス
 - ジョイスティック操作による手先速度操作
 - 手先位置姿勢操作: 従来の操作に近い教示
 - XRデバイスの振動などによるハプティックフィードバック
 - バイラテラル制御
 - ダイレクトティーチング
 - 操作インターフェイス自体を学習によって獲得する研究もある
- どのような人が教示するか
 - エキスパート: 高い精度, 無駄のない動き
 - 素人: ばらつきは大きくとも最終的にはタスクが遂行できる
- 収集されたデータの質・量との関係
 - 現在の基盤モデルでは質よりも量を優先させた収集が有効



A. Sasagawa, et al., *IEEE RA Letters*

データ作成方法: Simulation or Real data

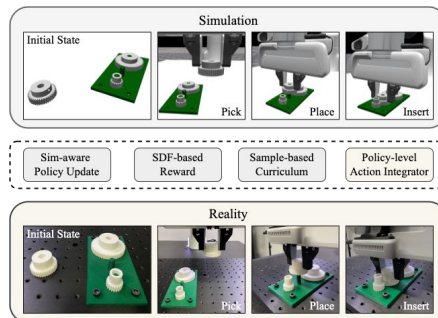
リアルの手間かシミュレーションギャップ

- リアルでのデータ収集

- 多数ロボットを並列して長期間作業
 - RT-1 :使用ロボット数:13台 収集期間:17ヶ月
- シミュレーションでのデータ収集
 - 力接触含めてリアルデータが得られるか/どのようにギャップを補完するか



[R. Elandt+, IROS20219]



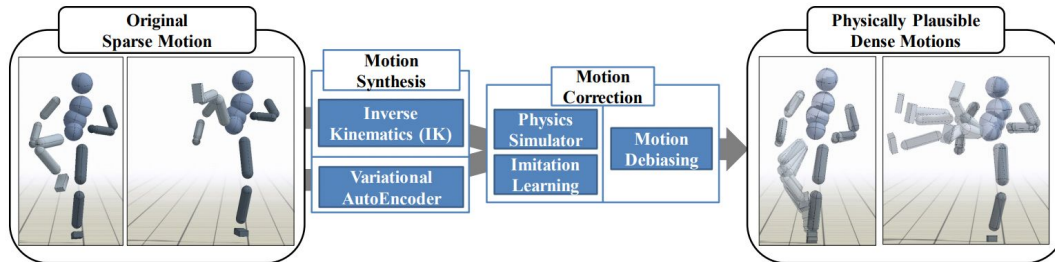
[B. Tang+, RSS2023]



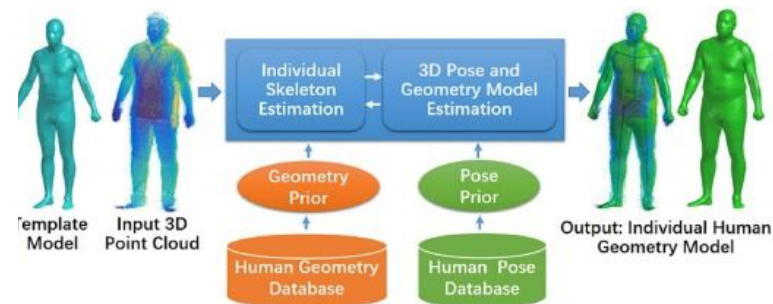
データ作成方法: ロボット作業データの権利問題

人権に配慮したデータセット

- 将来スマート社会として気軽に運動データが収集できるようになったときに、現在の言語モデルのような権利関係の問題は起こりうる。
- 作業データに対する
 - 匿名化技術
 - データオーグメンテーションなどが今後重要になってくる。



[T. Maeda+, CVPR2022]



[Su, Le+, *Journal of Visual Languages & Computing* 2018]

モデル構造 (1/8)

アーキテクチャと入出力

- ニューラルネットワーク (NN)
- 畳み込みニューラルネットワーク (CNN)
例) Dex-NetのGrasp-Quality-CNN [Mahler et al., 2017]
その他ResNet, VGG, LSTMなどから特性合わせた設計
→ 構築の一部が標準でライブラリに提供



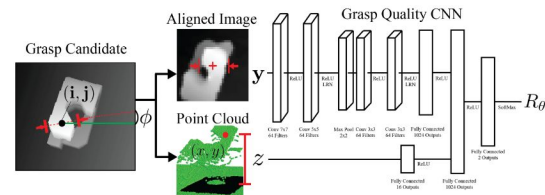
- 大規模言語モデルへの連携・展開・活用

→ **Transformer**を用いた学習と応用がブーム

- 画像やロボットアクションを入力とする既存モデルに Transformerの処理の組み込み、さらにテキスト(あるいは文脈)情報を付加することで高品質な出力を獲得.

特にCV分野の進捗は**対象を認識する上で非常に有用なツール**となりうる

例) Segment Anything, CLIP等は今後の活用が期待される



[Dex-Net by BerkeleyAutomation](#)

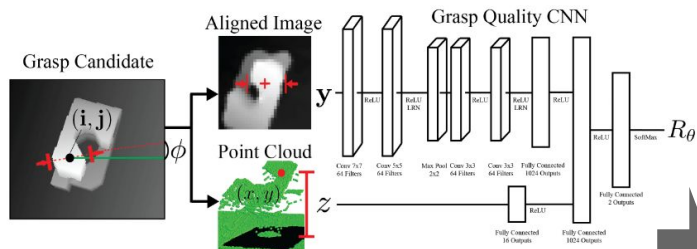
Segmentation: U-net/FCN/AutoEncoder ect. → Transformerに基づく物体識別が主流に? (後述)



モデル構造 (2/8)

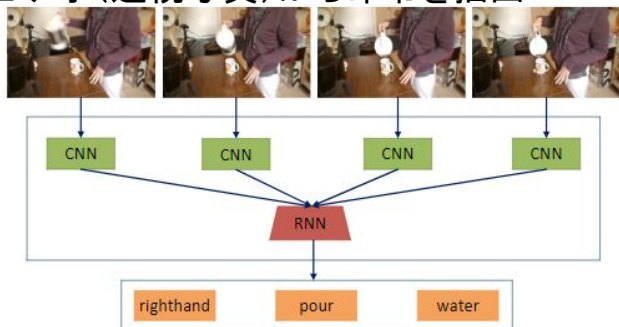
アーキテクチャと入出力

- 深度画像から把持姿勢の検出



Dex-Net by BerkeleyAutomation [Mahler+, 2017]

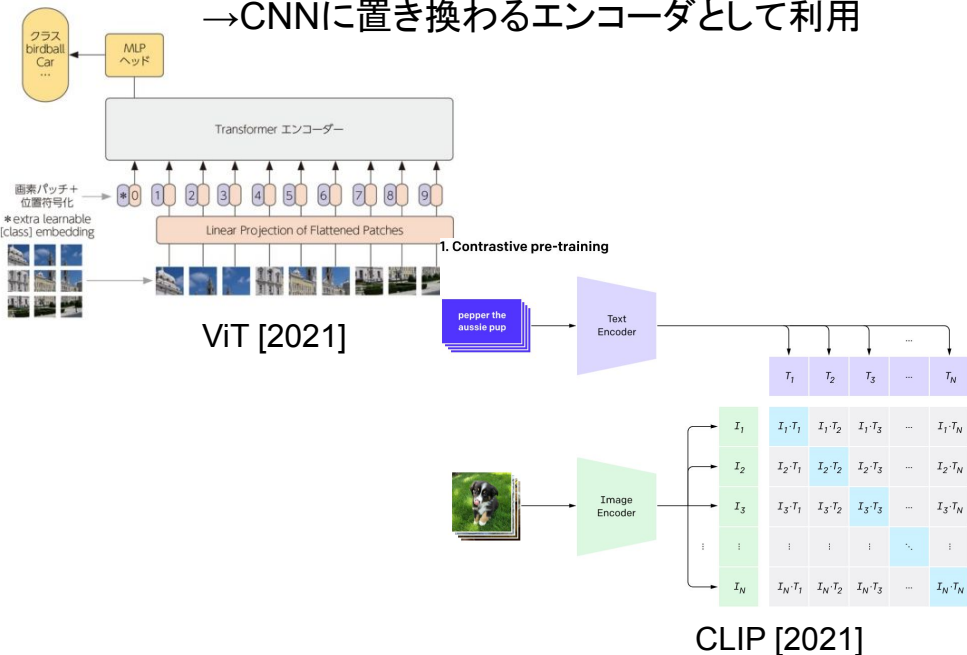
ビデオ(連続写真)から命令を抽出



Translating Videos to Command [Nguyen+, 2017]

Transformer主体のモデルが開発

→CNNに置き換わるエンコーダとして利用



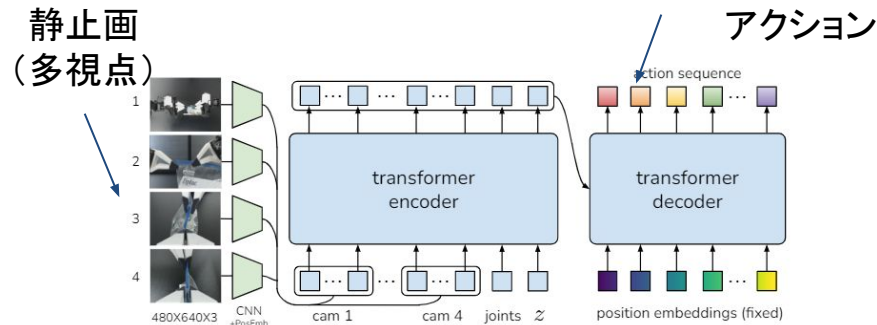
モデル構造 (3/8)

Transformerを軸とした実装モデル

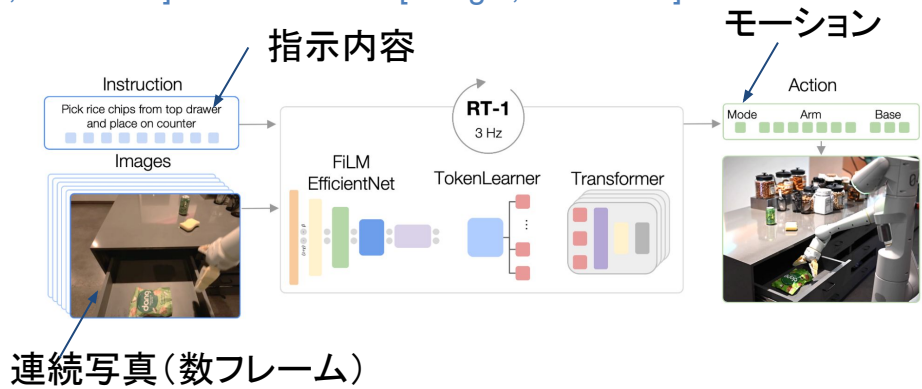
Transformerの適用: 入出力関係の設計 / 学習時の更新則の工夫

- 画像の畳み込みによる適用事例: 画像(複数視点)や動画からスキル/行動を取得

例) Fine-Grained Bimanual Manipulation [Zhao+, RSS2023]



RT-1 [Google, arXiv2023]



- 強化学習への適用事例: 方策や価値関数などによらない意思の決定が可能

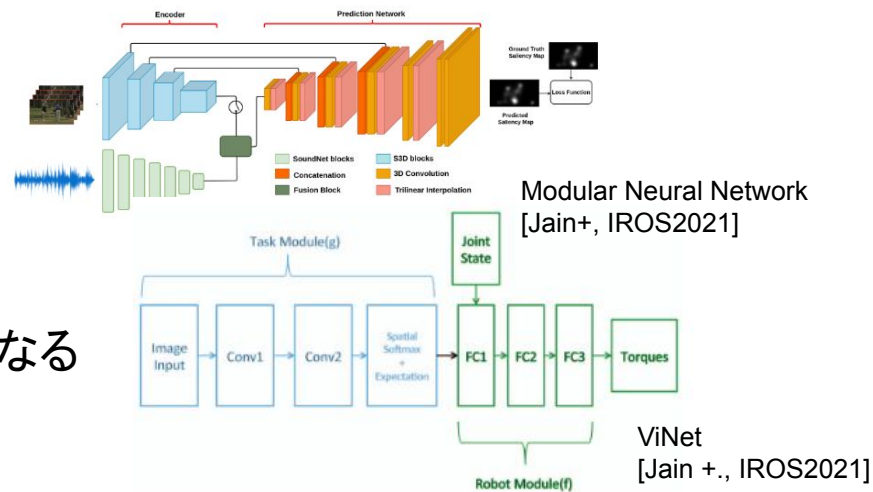
Decision Transformer [Chen+, arXiv2020]



モデル構造 (4/8)

モダリティの課題

- モダリティのマージと選定
 - 入力情報に応じた性能向上
例) ViNet 【sound+image】
 - 今後はLMからの適切な選択が必要になる



- ハードウェア及びタスクの汎化性能
 - 個別の学習フローとその統合
例) Learning Modular NN 【task+robot】
 - 大規模モデル≡模倣学習であるため性能の限界が見える
 - 異なるモダリティを統一的に連携する→参考) [ImageBind](#)



モデル構造 (5/8)

推論速度の課題

応答速度重視のタスクの実行時には...？

- 画像などをトークン単位すると精度向上
- × 指数関数的に計算量増加, 推論時間が長い

精度とリアルタイム性がトレードオフになる



例: コンタクトレンズをつける様子

[Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware \(tonyzhaozh.github.io\)](https://github.com/tonyzhaozh/learning-fine-grained-bimanual-manipulation)

次元圧縮や入力データの選択による処理速度向上

- 畳み込みニューラルネットワークによる前処理
- [TokenLearner](#)によるトークン圧縮(2.4倍の推論速度向上) 参考 [RT-1](#)
 - RT-1の3HzはTransformerを用いた手法としては高速だが...
 - 繊細/器用な作業については学習データの質に関わらずに不十分？



モデル構造 (6/8)

基盤モデルの組み合わせ

→ 対象が異なるモダリティ/スケールのモデルの統合

- 『言語と画像および指示に基づくロボットスキル獲得』

RT-1, Google 2022

- ① 個別のスキル獲得→RT-1
- ② 各スキルをつなぐ戦略→PaLM-SayCan

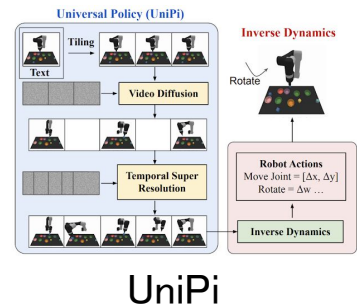


SayCan

- 『言語による動画生成』と『ロボットの行動制御』

UniPi (Universal Policy), Google 2023

- ① 生成されたシーン(時間ごとのロボットの状態)
- ② 逆動力学によりロボットの行動を取得



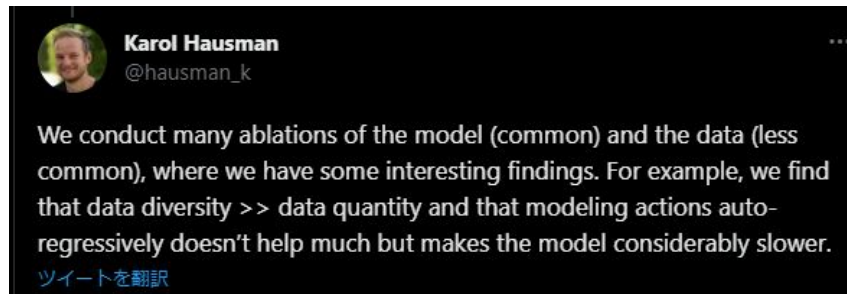
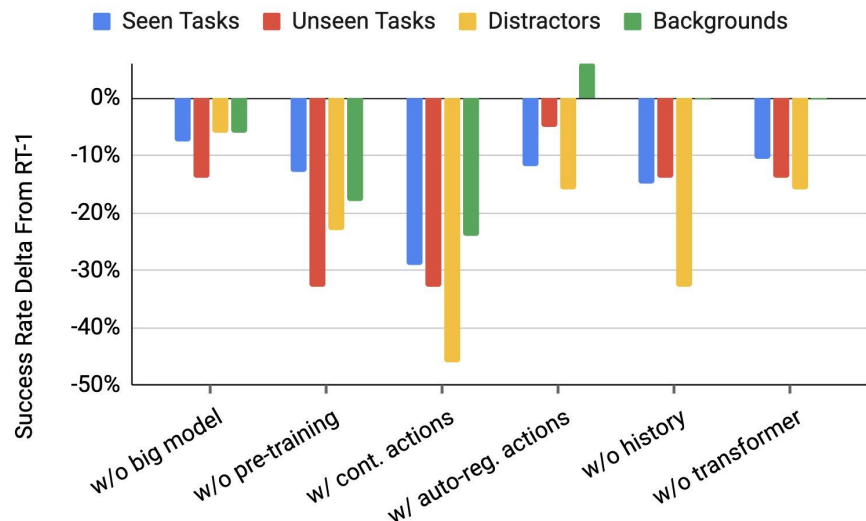
→ ロボットの仕様や安全性に対応. モデルの切り替えが重要でないか?



モデル構造 (7/8)

基盤モデルの注意点

- 大規模な基盤モデルの利用について
- 各タスクへの適用はこれまでの組み合わせを考慮すべき
- データの多様性 >> データの量



RT-1の事例

[引用: @hausman k](#)

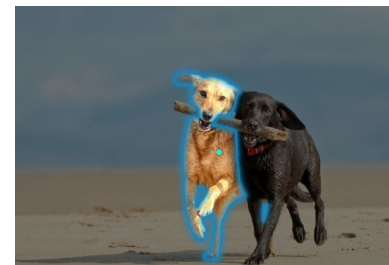
- データ量は模倣の性能を向上(見たことある, に強くなる)
- データの多様性(見たことない, に強くなる)



モデル構造 (8/8)

基盤モデルの利用 (ゼロショット転移)

- 別のタスクに転用することで真価を発揮！
 - [Segment Anything Model \(SAM\)](#) [Meta, arXiv2023]
 - 11M の画像, 1B+のマスクの大規模データ
 - Zero-shotにあらゆる対象のマスクングの可能性
 - [Personalize-SAM](#) [Zhang+, arXiv2023]
 - One-shotのドメイン適用を実現
 - [DreamBooth](#) [Ruiz+, CVPR2023] を強化 (Few-shot特徴空間を構築)
 - ・ 数枚で任意の画像が生成可能



[Segment Anything](#)

➡ **Zero-shot**(ゼロショット転移)のモデルを用いた新しい対象への展開
→ **組み合わせるだけで新しいタスクが解けるようになる!**

CLIPとGPTを接続

CLIP: 言語と2D画像 X GPT: 言語を拡張 → **2D,3Dシーンの拡張**

参考) [OpenScene: 3D Scene Understanding with Open Vocabularies](#)

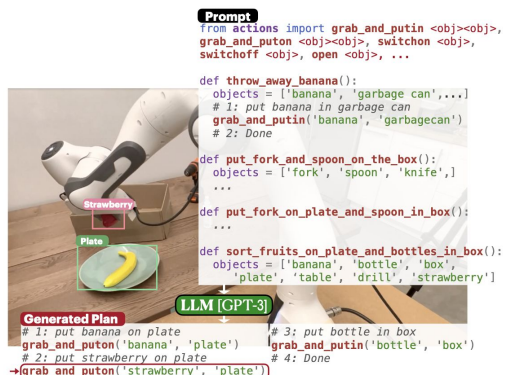


Prompting vs Learning

LLMまたは強化・模倣学習どちらで行動生成するか

- 言語で指定可能な簡単な行動(Pick-and-Placeなど)の組み合わせであれば, Long-termなタスクでもLLMのCodingとLVMを組み合わせることで実環境でもZero-shot実現できる(Progprompt, Code as Policies)
- 一方で言語での記述が困難な行動(Grasping, Assemblyなど)は強化学習または模倣学習を使って行動を生成する必要がある(DeepRL at Scale)
- タスク階層構造を理解し, PromptingとLearningどちらで実現可能かを切り替えるといい?

Progprompt [I. Singh+, ICRA2023]



Prompt

```
from actions import grab_and_putin <obj><obj>,
grab_and_puton <obj><obj>, switchon <obj>,
switchoff <obj>, open <obj>, ...

def throw_away_banana():
    objects = ['banana', 'garbage can', ...]
    # 1: put banana in garbage can
    grab_and_putin('banana', 'garbagecan')
    # 2: Done

def put_fork_and_spoon_on_the_box():
    objects = ['fork', 'spoon', 'knife', ...]
    ...

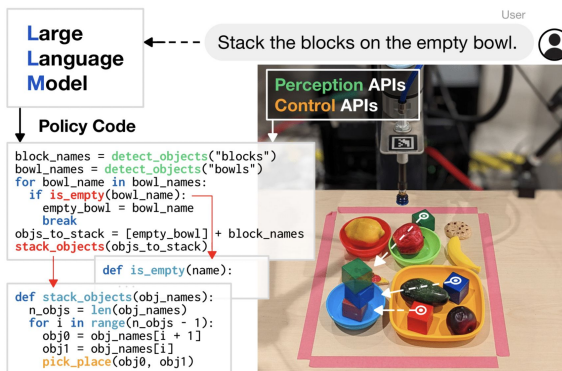
def put_fork_on_plate_and_spoon_in_box():
    ...

def sort_fruits_on_plate_and_bottles_in_box():
    objects = ['banana', 'bottle', 'box',
              'plate', 'table', 'drill', 'strawberry']
    ...
```

Generated Plan LLM (GPT-3)

```
# 1: put banana on plate
grab_and_puton('banana', 'plate')
# 2: put strawberry on plate
grab_and_puton('strawberry', 'plate')
# 3: put bottle in box
grab_and_putin('bottle', 'box')
# 4: Done
```

Code as Policies [J. Liang+, ICRA2023]



Large Language Model

Stack the blocks on the empty bowl.

Perception APIs
Control APIs

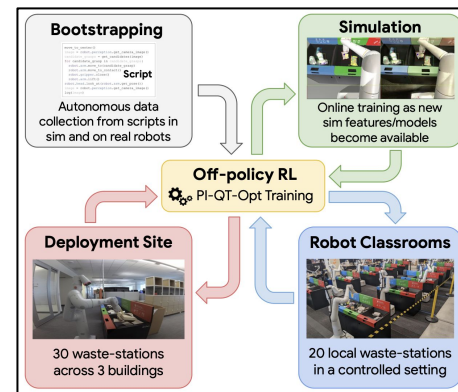
Policy Code

```
block_names = detect_objects("blocks")
bowl_names = detect_objects("bowls")
for bowl_name in bowl_names:
    if is_empty(bowl_name):
        empty_bowl = bowl_name
        break
objs_to_stack = [empty_bowl] + block_names
stack_objects(objs_to_stack)

def is_empty(name):
    ...

def stack_objects(obj_names):
    n_objs = len(obj_names)
    for i in range(n_objs - 1):
        obj0 = obj_names[i + 1]
        obj1 = obj_names[i]
        pick_place(obj0, obj1)
```

DeepRL at Scale [A. Herzog+, arXiv2023]



研究協力者募集・著者紹介

産総研 ICPS研究センター オートメーション研究チーム

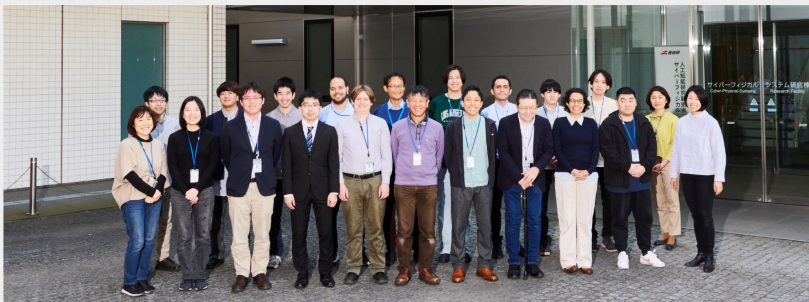
多様な人にとって働きやすく、

かつ、生産性も持続する。

活力のあるインクルーシブな社会の実現へ

Welcome to **ART, AIST!**

国立研究開発法人産業技術総合研究所
インダストリアルCPS研究センター
オートメーション研究チーム



<https://unit.aist.go.jp/icps/icps-am/>

共に研究を推進する方(研究者、企業、学生)を広く募集しています。

博士前期・後期課程の学生:産総研リサーチアシスタント制度
[こちら\(産総研リサーチアシスタント制度紹介ページ\)](#)

博士号取得済みの若手研究者:ポスドク採用, 常勤採用
常勤:[こちら\(産総研研究職員採用情報ページ\)](#)
ポスドク:[こちら\(産総研契約職員採用情報ページ\)](#)

エンジニア志望者:テクニカルスタッフ採用
[こちら\(産総研契約職員採用情報ページ\)](#)

相談連絡先:下記のメールアドレス先にお送りください。

M-icps-am-contact-ml[[@](mailto:M-icps-am-contact-ml@unit.aist.go.jp)]aist.go.jp

* [[@](mailto:M-icps-am-contact-ml@unit.aist.go.jp)]を半角の@にしてお送りください。



robotpaper.challenge



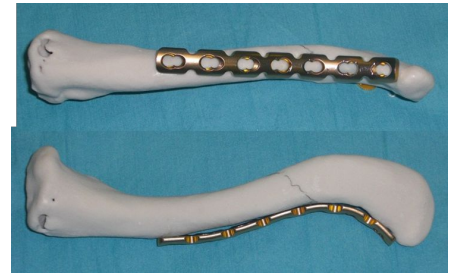
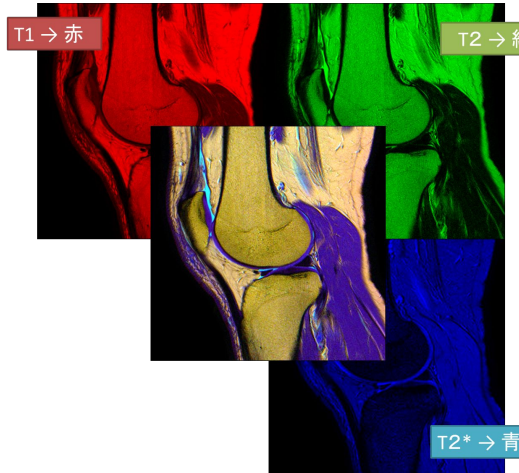
cvpaper.challenge



- 名前: 中原龍一
- 所属: 岡山大学整形外科
- 出身: 愛媛県
- 趣味: 読書, 数学, 散歩

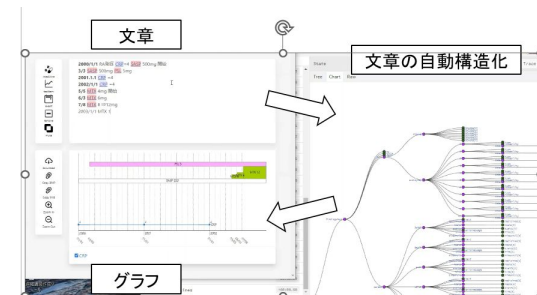
【研究テーマや研究の趣味】

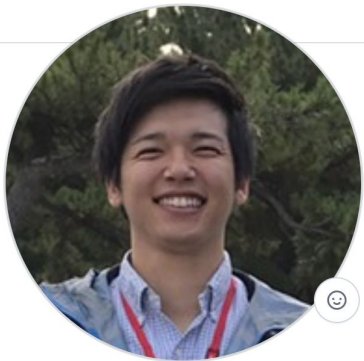
- 医療画像のAI研究
- 電子カルテの言語情報のAI研究
- AIの社会人教育, リスキリング



【何かひとつ】

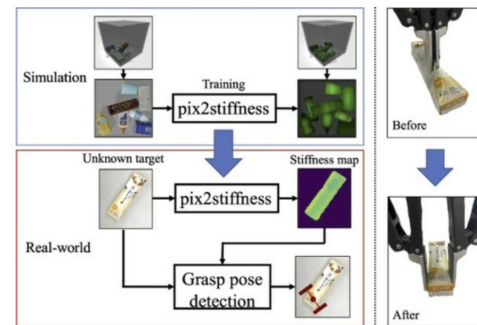
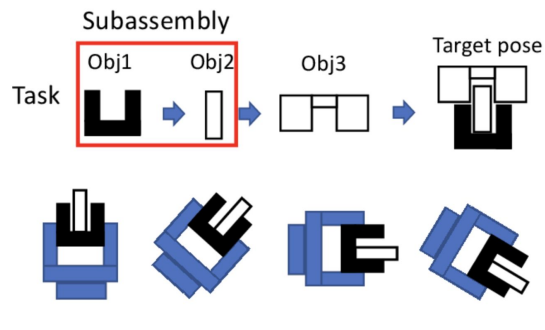
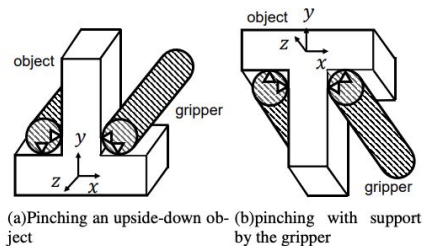
- 医療と工学の懸け橋になれるように頑張ります





- 名前: 牧原昂志(まきはら こうし)
- 所属: 大阪大学 D3 / 産総研 RA
- 出身: 佐賀県
- 趣味: ゴルフ

【研究テーマ: 物体把持】



【何かひとつ】

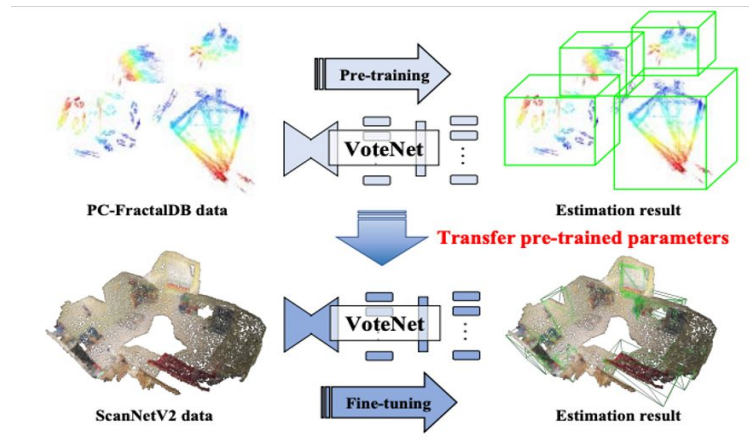
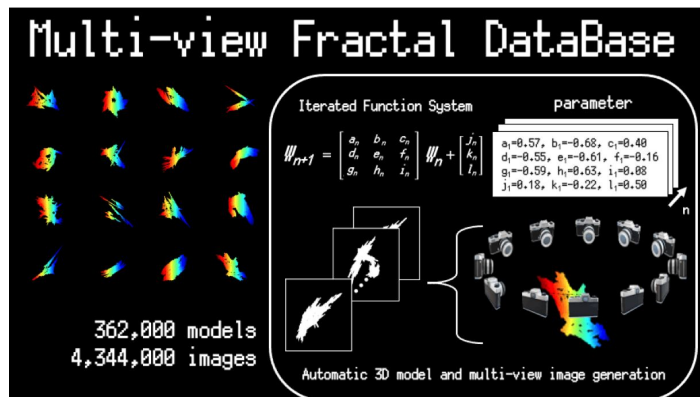
楽しく頑張ります



山田 亮佑 (やまだりょうすけ)

- ❑ 筑波大学 博士課程2年
- ❑ 産業技術総合研究所 リサーチアシスタント
- ❑ 日本学術振興会 特別研究員 (DC1)

【研究テーマ: 3D物体認識】



【最後に一言】

AIスターに俺はなるっ!!!!

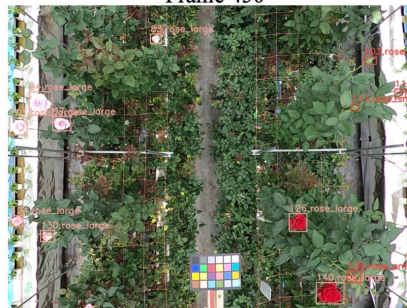


- 名前: 篠田理沙(しのだ りさ)
- 所属: 京都大学 D2 / 産総研・東工大 RA
- <https://sites.google.com/view/risashinoda/home>

【研究テーマ: セグメンテーション、農業✕画像認識】



Frame 450

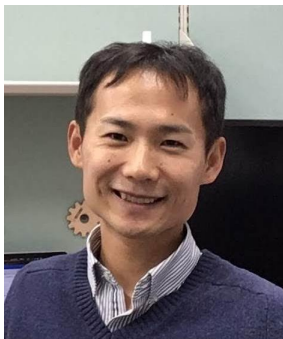


Frame 500



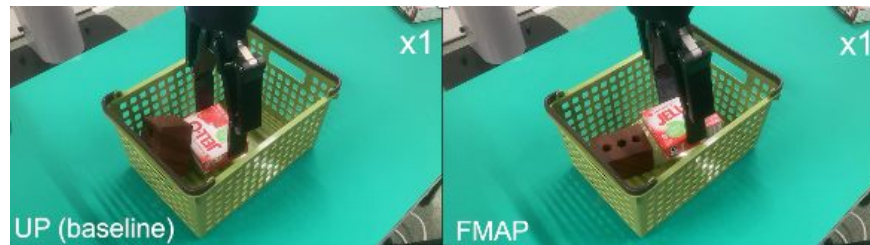
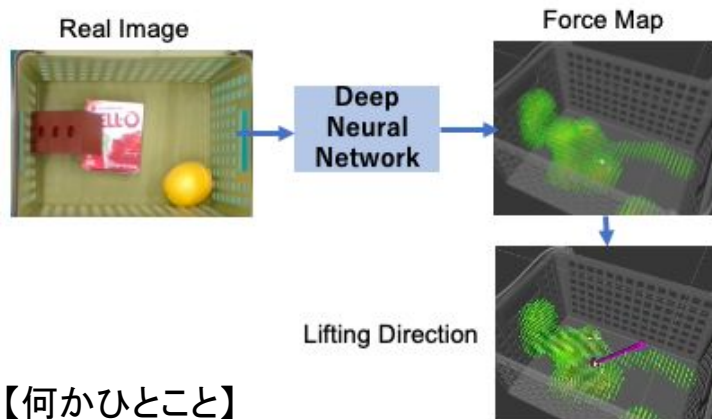
【何かひとつ】

ロボットのこともっと知りたい。



- 名前: 花井 亮(ハナイ リョウ)
- 所属: 産業技術総合研究所 主任研究員
- 出身: 和歌山県
- 趣味: テニス

【研究テーマ: 視覚から力の推定, ロボット動作生成への応用】



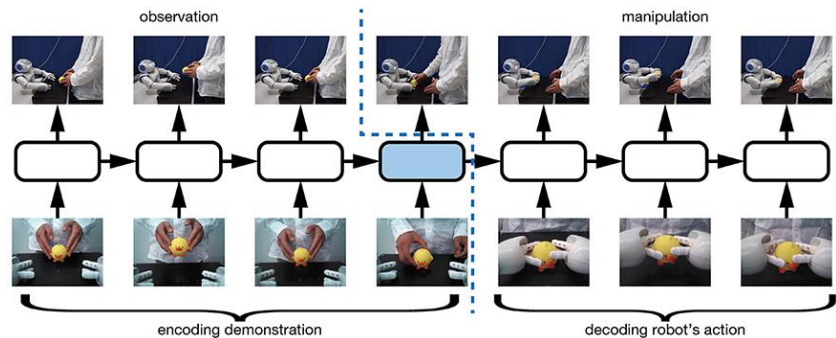
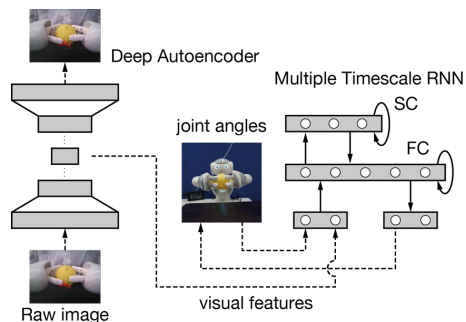
【何かひとつ】

- 生涯現役



- 名前: 中條 亨一 (なかじょう りょういち)
- 所属:
 - 産業技術総合研究所 産総研特別研究員
 - 早稲田大学 次世代ロボット研究機構AIロボット研究所 客員次席研究員
- 出身: 神奈川県

【研究テーマ: 模倣学習によるロボットの動作生成、人-ロボットインタラクション】



【何かひとつ】

とても楽しいメタサーベイでした！

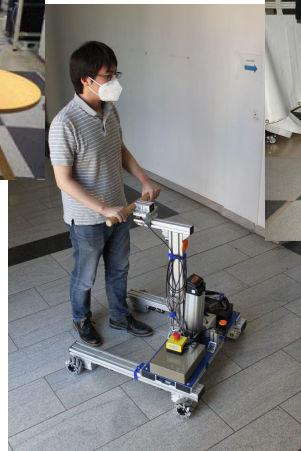
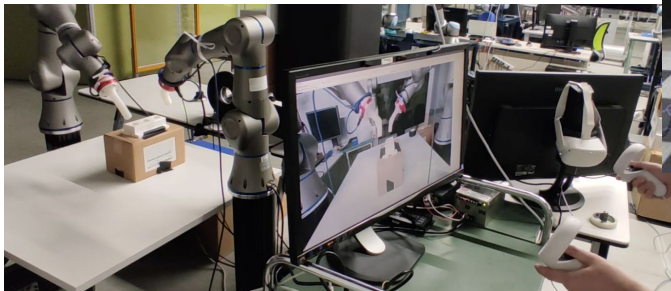


- 名前: 板寺駿輝(イタデラ シュンキ)
- 所属: 産業技術総合研究所 研究員
- 出身: 千葉?
- 趣味: ドライブ

<https://staff.aist.go.jp/s.itadera/>

【研究テーマ: 遠隔操作インターフェイス, 物理的インタラクション, 人支援ロボティクス】

<https://github.com/itadera/OpenHRC>



【何かひとつ】

上記テーマでも共同研究者, 学生さん募集中



- 名前:元田 智大(モトダ トモヒロ)
- 所属:産業技術総合研究所 研究員
- 出身:大阪府
- 趣味:実家の愛猫/ ピアノ演奏

個人ウェブサイト
:<https://tomohiromotoda.github.io/>

Researchmap: <https://researchmap.jp/tomohiro.motoda>



【研究テーマ:双腕ロボットを用いた”安全”なマニピュレーション】

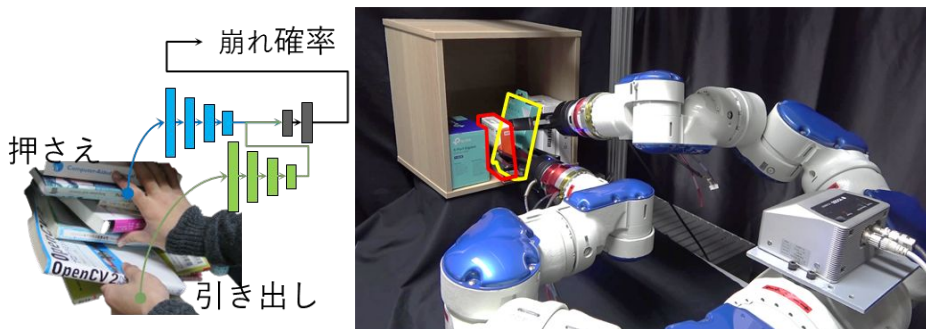


図:両手で補助しながら作業することで
密集したところから取り出せる!

【何かひとつ】

日進月歩→(臥薪嘗胆)→テール・トゥ・ウイン

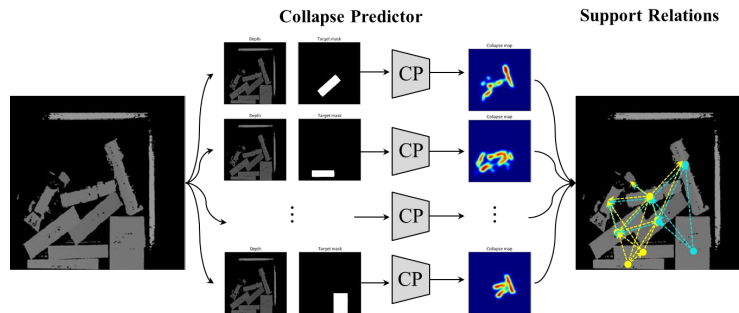


図:複雑なシーンでも物理現象に基づく経験から
「崩れ」の発生が予想可能。
これらの予測から基にサポート関係を推定。

コンピュータビジョン(CV)研究者

研究トレンドを常時創れる研究コミュニティを構築する

■ 経歴等:

茨城県出身, 現在つくば市在住

博士(工学), 2014年 慶應義塾大学(青木義満研究室)

産業技術総合研究所 研究員(2016.4~2020.9) 主任研究員(~2020.10) 上級主任研究員(2023.04~現在)

豊田自動織機-産総研 連携研究ラボ(2016.10~現在)

東京電機大学 訪問研究員(2016.4~現在)

情報科学の達人 メンター(2020.11~現在)

コンピュータビジョン最前線 編集委員(2021.12~現在)

LINE株式会社 クロスアポイントメント出向社員(2022.4~現在)

cypaper.challenge PI | nlpaper.challenge HQ | robotpaper.challenge HQ



■ 研究プロジェクト:

「自然の形成原理に基づく深層学習の真相究明(科研費基盤研究A; 代表)」

「社会インフラ映像処理のための高速・省資源深層学習アルゴリズム基盤(CREST; 分担)」

「容易に構築・導入できるAIの開発(NEDO3.0; 項目代表)」

■ 研究業績(最近2年間の主な業績):

“Visual Atoms: Pre-training Vision Transformers with Sinusoidal Waves (CVPR23)”

”Replacing Labeled Real-Image Datasets with Auto-Generated Contours (CVPR22)”

“Point Cloud Pre-training with Natural 3D Structures (CVPR22)”

“Pre-training without Natural Images (IJCV22, ACCV20 Best Paper H. M. Award)”

“Can Vision Transformers Learn without Natural Images? (AAAI22)”

“Describing and Localizing Multiple Changes with Transformers (ICCV21)”

片 かた

岡 おか

裕 ひろ

雄 かつ

<http://hirokatsukataoka.net/>



@HirokatuKataoka



hirokatsukataoka16



cvpaper.challenge