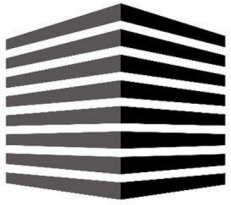


2022年 コンピュータビジョン分野のトレンド
～Computer Vision Trends in 2022～

片岡 裕雄
産業技術総合研究所
人工知能研究センター

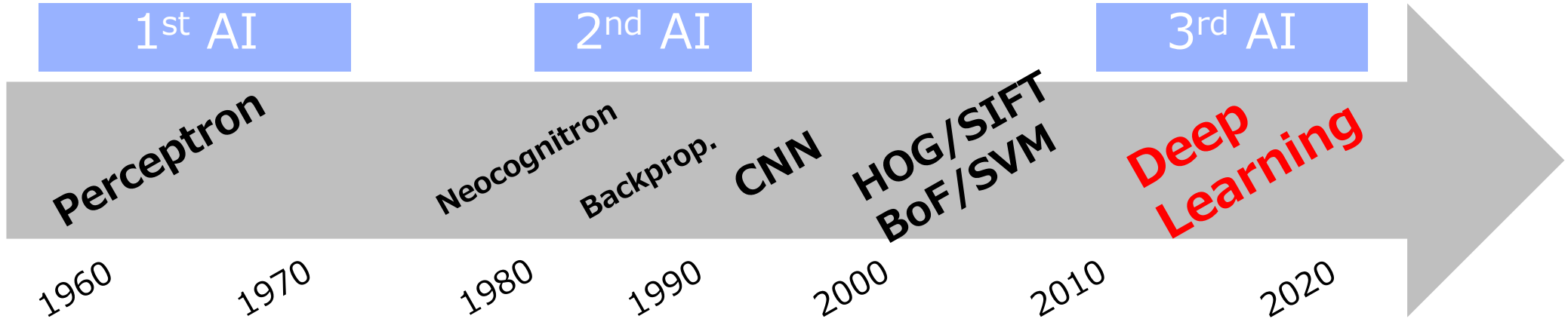
http://xpaperchallenge.org/cv/

DNNの動向・CVのトレンド（1/42）

DNN時代以前の動向

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

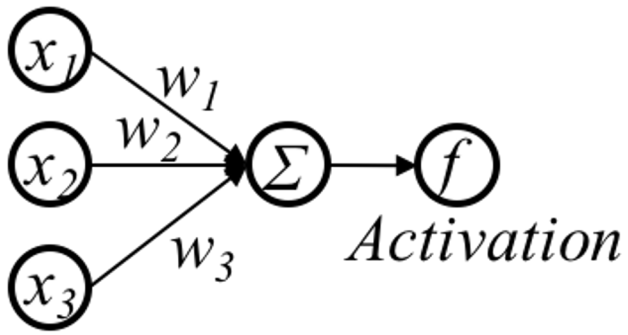
1st - 3rd AIまでの流れ



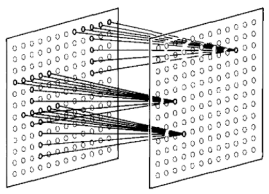
DNNの動向・CVのトレンド（2/42）

Perceptron, MLP, Neocognitron/ConvNet

- 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
- 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

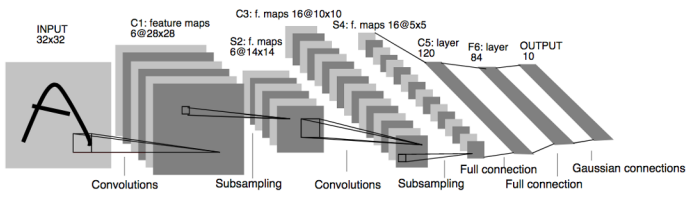


Perceptron（パーセプトロン）



Neocognitron

K. Fukushima, “Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position,” Biol. Cybenetics 36, pp.193-202, 1980.
https://www.rctn.org/bruno/public/papers/Fukushima1980.pdf



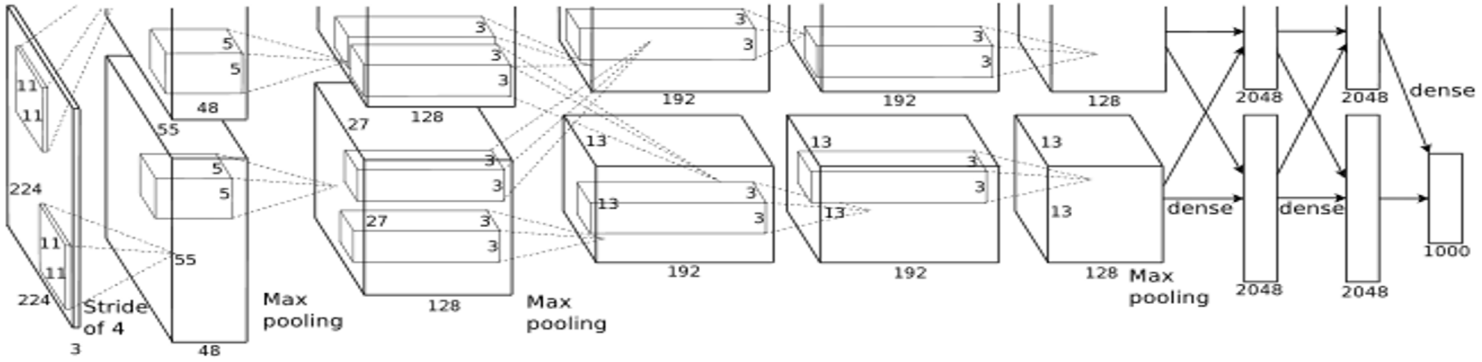
Convolutional Neural Net

Y. LeCun et al. “Gradient-Based Learning Applied to Document Recognition,” IEEE, 1998.
http://yann.lecun.com/exdb/publis/pdf/lecun-01a.pdf

DNNの動向・CVのトレンド（3/42）

ILSVRCを発端とする画像識別タスクへの応用

- AlexNet @画像認識コンペILSVRC2012
  - もはや専門家には説明不要だが，2位に10%以上の大差で圧勝
  - 10年弱で100,000⁺回も引用される論文となる
- 背景には構造をDeepにする技術が揃っていた
  - 最適化，活性化関数，正則化など



AlexNet
A. Krizhevsky et al., “ImageNet Classification with Deep Convolutional Neural Networks,” in NIPS 2014.
https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

DNNの動向・CVのトレンド（4/42）

DNNが勝てた背景

- DNN Architecture!（モデル自体や学習戦略）
- Large-scale Image Datasets!（データは最重要）
- Machine Power!（圧倒的な計算力; NVIDIA!）

Why 3rd AI?

Architecture（Model）
ConvNet，Transformer，GAN…

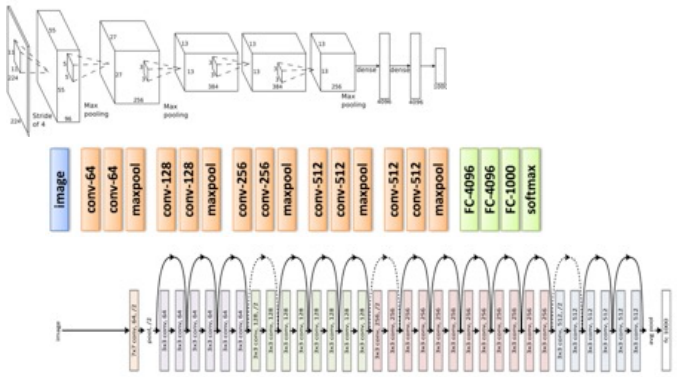


Image Data
ImageNet, Places, COCO…

IMAGENET
http://www.image-net.org/

Machine Power
GPU, TPU, Super Computer…

NVIDIA

https://nvidianews.nvidia.com/multimedia/corporate/corporate-nvidia-logos-legal-filings

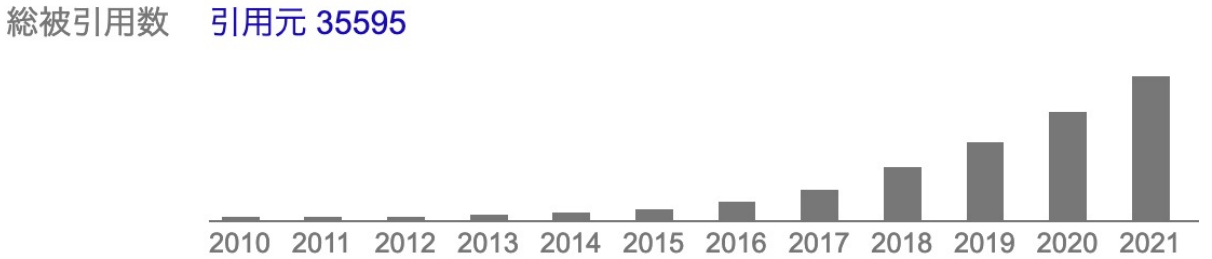
DNNの動向・CVのトレンド（5/42）

大規模画像データセット（ImageNet）の収集について

- 14,197,122 画像 / 21,841 カテゴリ https://www.image-net.org/
- 2007年からデータを収集，2009年CVPR発表
- その後，画像認識コンペを開催し2012年大会でAlexNetが提案される

http://fungai.org/images/blog/imagenet-logo.png

ImageNetのロゴ，右側はStanfordの赤，左は前所属のPrinceton，そして上の緑はWorldPeaceー世界平和ーを示す（らしい）



https://scholar.google.com/citations?view_op=view_citation&hl=ja&user=rDfyQnIAAAAJ&citation_for_view=rDfyQnIAAAAJ:qjMakFHDy7sC

ImageNet論文（CVPR 2009）の被引用数：コンピュータビジョン分野における深層学習の隆盛を象徴する存在

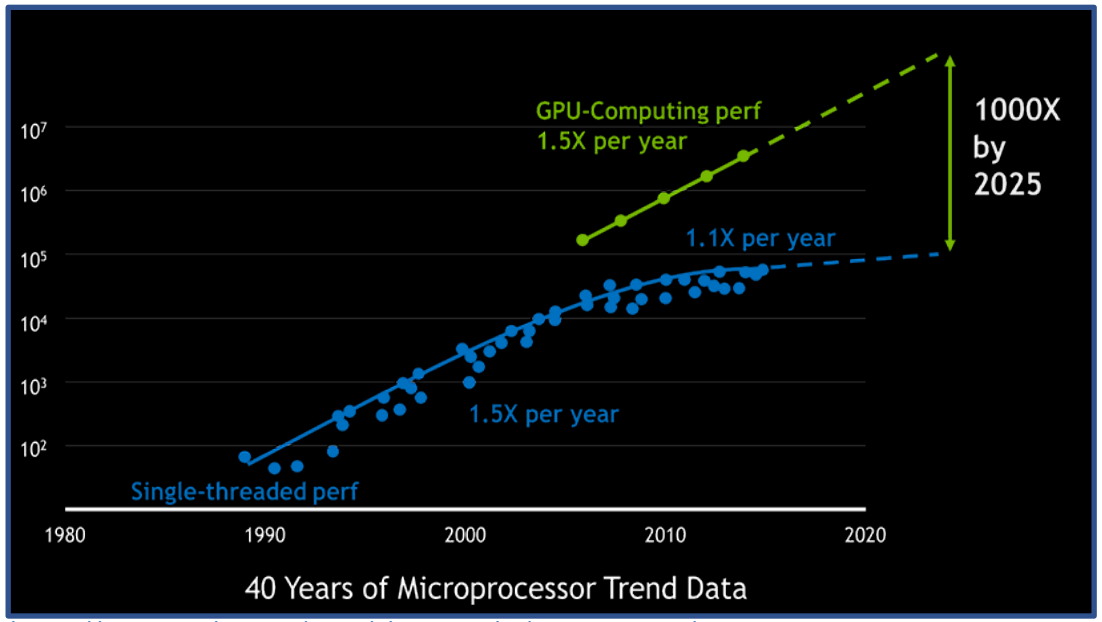
Fei-Fei氏のTED動画では資金繰りの苦労について言及，ワークショップ等では2000年代当時はアルゴリズム至上主義でデータを収集することが理解されなかったとも

https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures

DNNの動向・CVのトレンド（6/42）

計算機環境（主にGPU）の発展

- 特に第３次AIブームからはNVIDIAの隆盛ぶりがすごい
- 当初ゲーム用ボードを売っていたが深層学習始めAIに会社の命運を託すと明言
- 結果，下記の性能向上と世界的な提携を進めるに至る



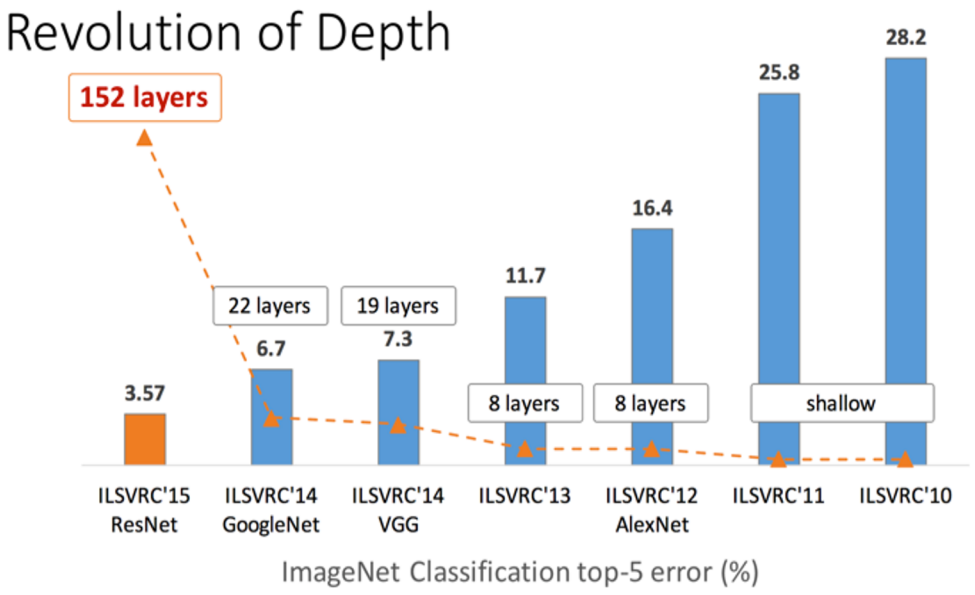
https://www.nvidia.com/ja-jp/about-nvidia/ai-computing/

年々1.5倍のペースでコンピューティングが進化，
2025年までには1,000倍にまで到達すると予想
（NVIDIA公式ページより）

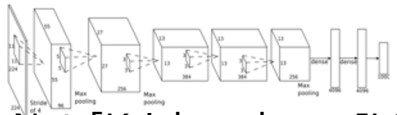
DNNの動向・CVのトレンド（7/42）

構造の深化（2014～2016）

- 2014年頃から「構造をより深くする」ための知見が整う
  - AlexNet, VGGNet, GoogleNetは，ほぼ単純に層を追加
  - ResNetは残差（Residual）の考え方を導入，Skip Connectionにより大幅に深化



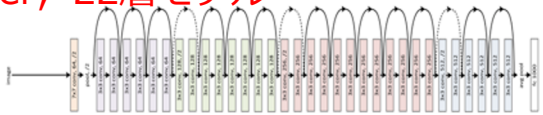
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016.
http://kaiminghe.com/ilsvrc15/ilsvrc2015_deep_residual_learning_kaiminghe.pdf



AlexNet [Krizhevsky+, ILSVRC2012]
ILSVRC2012 winner，DLの火付け役

VGGNet [Simonyan+, ILSVRC2014]
16/19層ネット，deeperモデルの知識

GoogLeNet [Szegedy+, ILSVRC2014/CVPR2015]
ILSVRC2014 winner，22層モデル



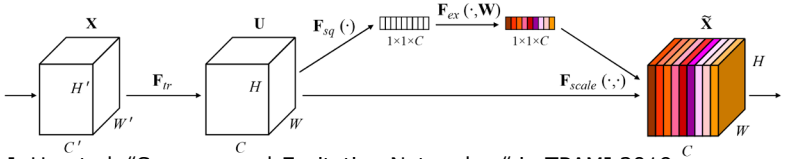
ResNet [He+, ILSVRC2015/CVPR2016]
ILSVRC2015 winner， 152層！(実験では10^3+層も)

DNNの動向・CVのトレンド（8/42）

構造の複雑化・自動化（2016～2019）

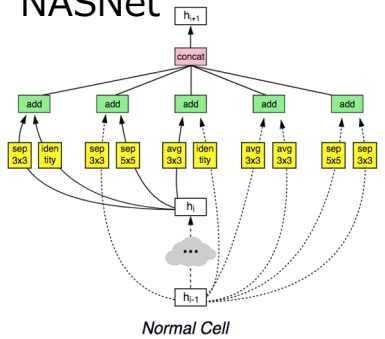
- ResNet以降のアーキテクチャ
  - ResNeXt, DenseNet, SENet, ...
- 自動化・効率化（Neural Architecture Search, EfficientNet）
  - (P)NAS, EfficientNetV1/V2...

SENet（SE-block）



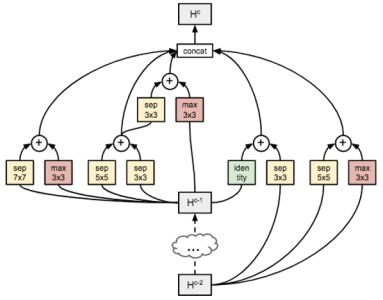
J. Hu et al. “Squeeze-and-Excitation Networks, ” in TPAMI 2019.
https://arxiv.org/abs/1709.01507

NASNet



B. Zoph et al. “Learning Transferable Architectures for Scalable Image Recognition,” in CVPR 2018.
https://arxiv.org/abs/1707.07012

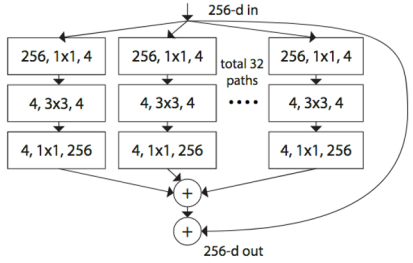
PNASNet



C. Liu et al. “Progressive Neural Architecture Search,” in CVPR 2018.
https://arxiv.org/abs/1712.00559

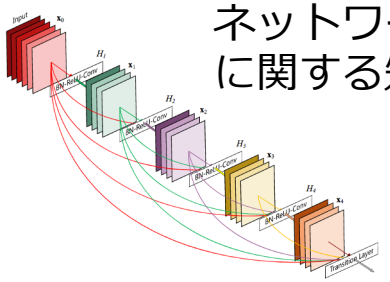
NASはデータセットに合わせてその場で構造を探索するという方針（但し計算リソースを大量に要する）

ResNeXt



S. Xie et al. “Aggregated Residual Transformations for Deep Neural Networks, ” in CVPR 2017.
https://arxiv.org/abs/1709.01507

DenseNet

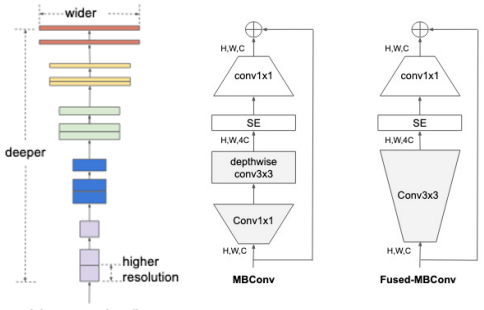


ネットワークの接続に関する知見が強化

G. Huang et al. “Densely Connected Convolutional Networks, ” in CVPR 2017.
https://arxiv.org/abs/1608.06993

EfficientNetは深さ・幅・入力に関するサイズのバランスが重要と解析

EfficientNetV1/V2



M. Tan et al. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in ICML 2019.
https://arxiv.org/pdf/1905.11946.pdf

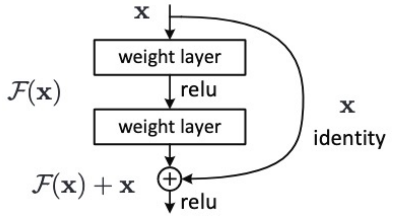
M. Tan et al. “EfficientNetV2: Smaller Models and Faster Training,” in ICML 2021.
https://arxiv.org/pdf/2104.00298.pdf

DNNの動向・CVのトレンド（9/42）

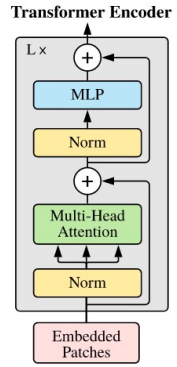
Vision Transformer（ViT）導入（2020～Present）

- 2017にTransformer提案，2020にViT提案
- 2021年，CV分野はTransformerの一年だった
- ViT派生モデル/MLP等で同様のモデル提案に至る

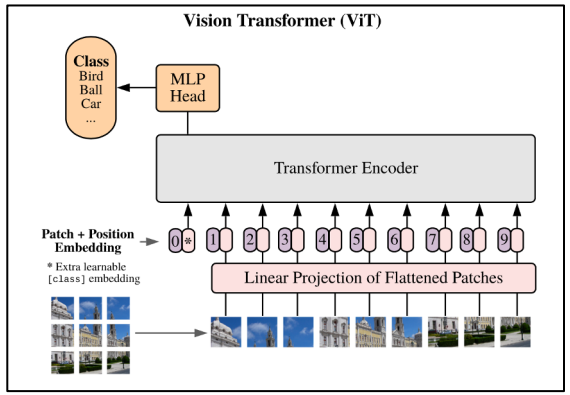
Computer Visionのパラダイムシフト
CNNからViTへ（2020年10月）



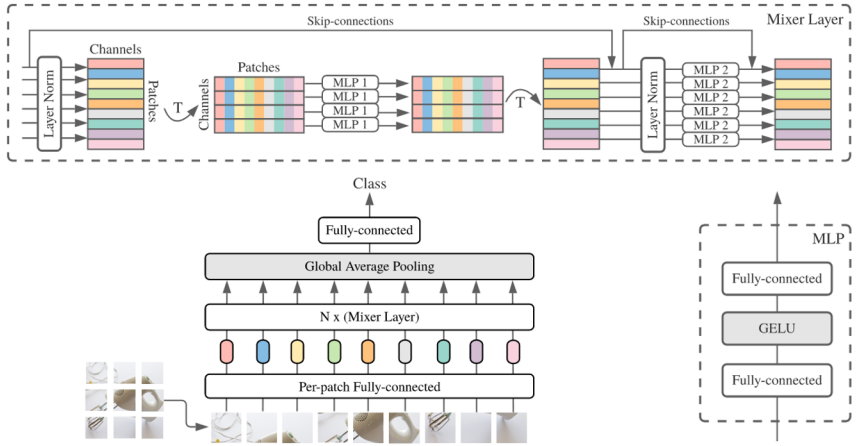
[He al. CVPR16]



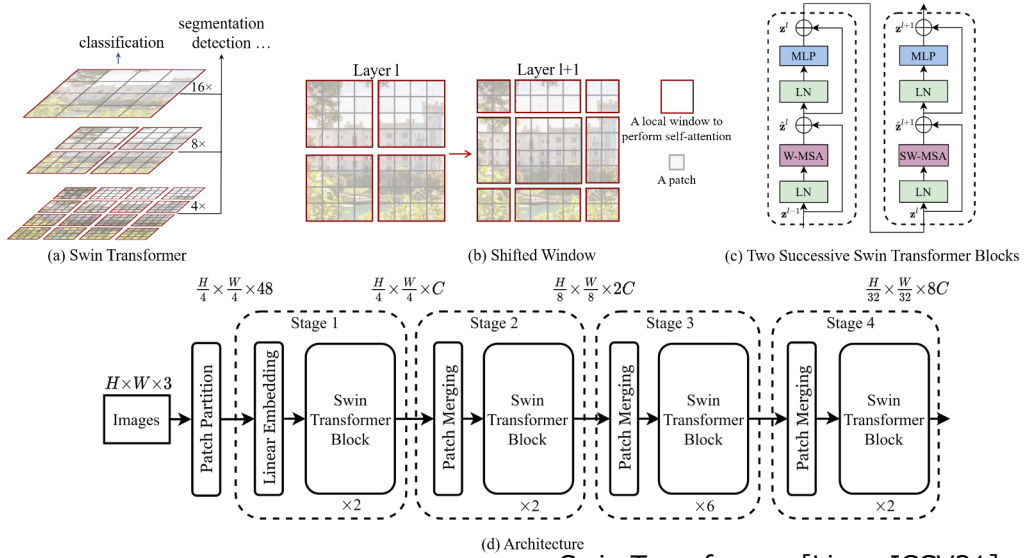
[Vaswani al. NIPS17]
Figure from [Dosovitskiy al. ICLR21]



ViT [Dosovitskiy+, ICLR21]



MLP-Mixer [Tolstinhin+, NeurIPS21]



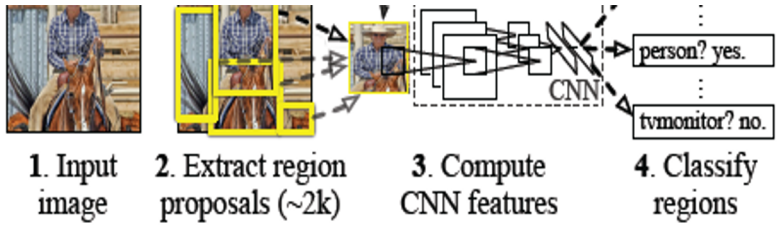
Swin Transformer [Liu+, ICCV21]

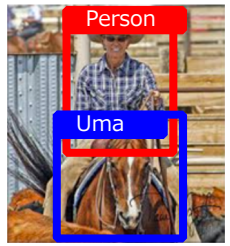
ViTはエンコーダのみのモデル，同様の機能をMLPで実現（MLP-Mixer），近年のベースラインはSwinTransformerか（V2も2021年末に提案）

DNNの動向・CVのトレンド（10/42）

他タスクへの応用（画像認識・動画認識・マルチモーダル）

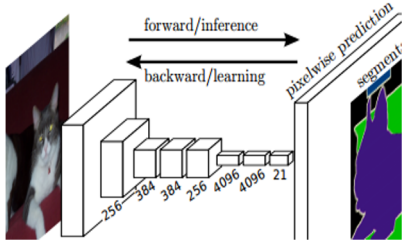
- 物体検出: R-CNN, Fast/Faster R-CNN, YOLO, SSD,,,
- 領域分割: FCN, SegNet, U-Net,,,
- Vision & Language: 画像説明文, VQA, Visual Dialog,,,
- 動画認識: Two-stream ConvNets, 3D Conv., (2+1)D Conv. ,,,





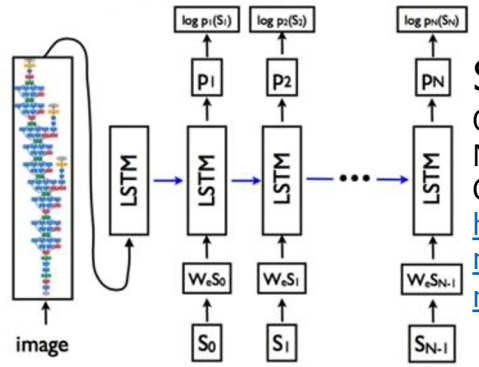
R-CNN

R. Girshick et al., “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” in CVPR 2014.
https://openaccess.thecvf.com/content_cvpr_2014/html/Girshick_Rich_Feature_Hierarchies_2014_CVPR_paper.html



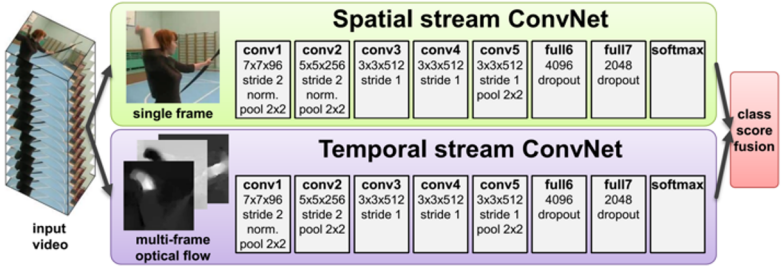
FCN

J. Long et al., “Fully Convolutional Networks for Semantic Segmentation,” in CVPR 2015.
https://openaccess.thecvf.com/content_cvpr_2015/html/Long_Fully_Convolutional_Networks_2015_CVPR_paper.html



Show and Tell

O. Vinyals et al., “Show and Tell: A Neural Image Caption Generator,” in CVPR 2015.
https://openaccess.thecvf.com/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html



Two-stream ConvNets

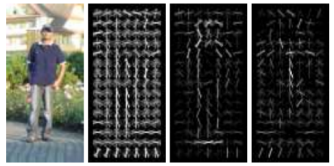
K. Simonyan et al., “Two-Stream Convolutional Networks for Action Recognition in Videos,” in NIPS 2014.
https://www.robots.ox.ac.uk/~vgg/publications/2014/Simonyan14b/simonyan14b.pdf

DNNの動向・CVのトレンド（11/42）

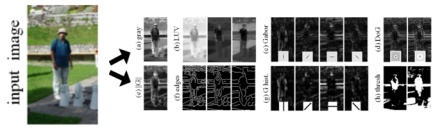
物体検出の流れ（2001～2017）

Hand-crafted feature時代　基礎/枠組みの構築

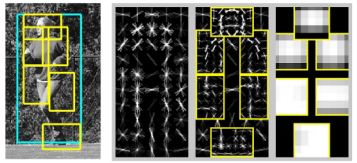
Haar-like [Viola+, CVPR01]
+ AdaBoost



HOG [Dalal+, CVPR05]
+ SVM

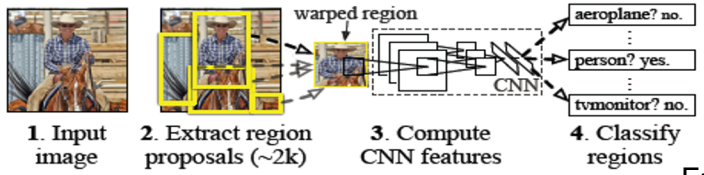


ICF [Dollár+, BMVC09]
+ Soft-cascade

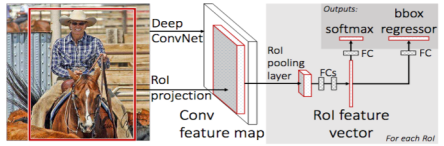


DPM [Felzenszwalb+, TPAMI12]
+ Latent SVM

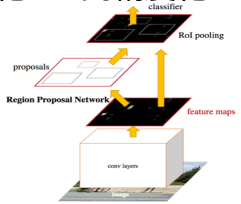
R-CNN時代（それ以前は"Hand-crafted" ObjectNess）高速化 & 高精度化



R-CNN [Girshick, CVPR14]
Selective Search + CNN



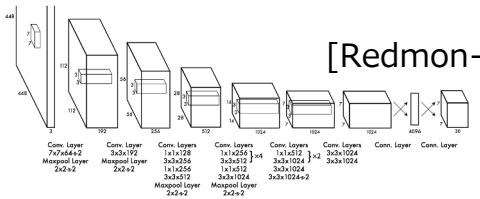
Fast R-CNN [Girshick, ICCV15]
ROI Pooling, Multi-task Loss



Faster R-CNN [Ren+, NIPS15]
RPN

One-shot Detector時代 兎にも角にも（精度を保ちつつ）高速化

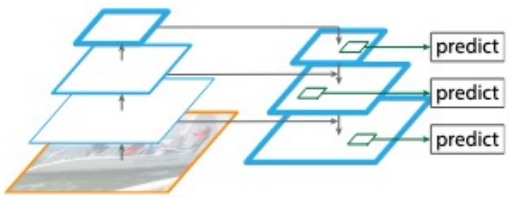
SSD [Liu+, ECCV16]
One-shot detector



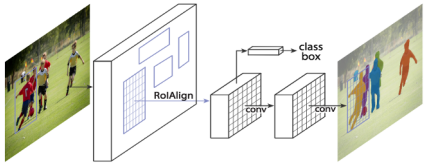
YOLO(v1)/v2/v3
[Redmon+, CVPR16/CVPR17/arXiv18]
One-shot detector

・・・

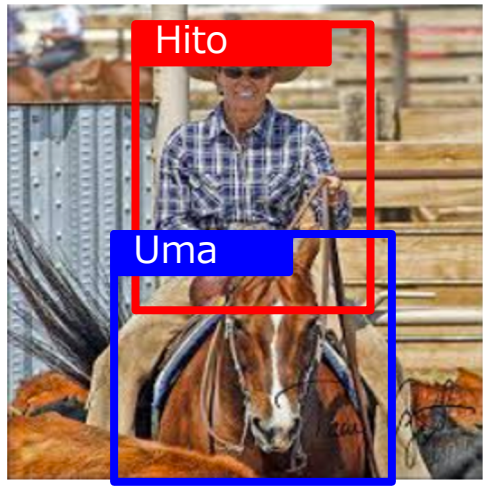
Latest Algorithm 精度重視，高速



Feature Pyramid
Networks (FPN)
[Lin+, CVPR17]



Mask R-CNN [He+, ICCV17]
RoI Align, Det+Seg



【物体１】
カテゴリ：Hito
位置：(x_1, y_1, w_1, h_1)

【物体２】
カテゴリ：Uma
位置：(x_2, y_2, w_2, h_2)

物体検出は物体カテゴリ識別と画像上の位置座標を回帰する問題

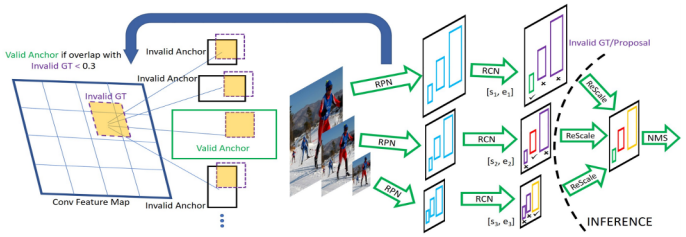
DNNの動向・CVのトレンド（12/42）

物体検出の流れ（2018～2020）

CNNによる工夫：
解像度による調整（上段），座標回帰による調整（下段）

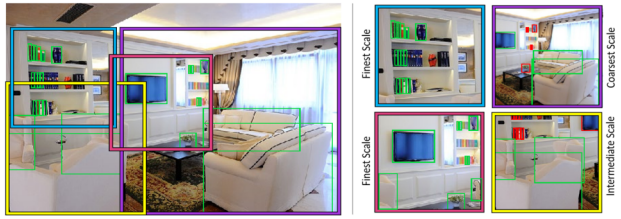
SNIP

B. Singh et al. “An Analysis of Scale Invariance in Object Detection,” in CVPR 2018.
https://arxiv.org/abs/1711.08189



SNIPER

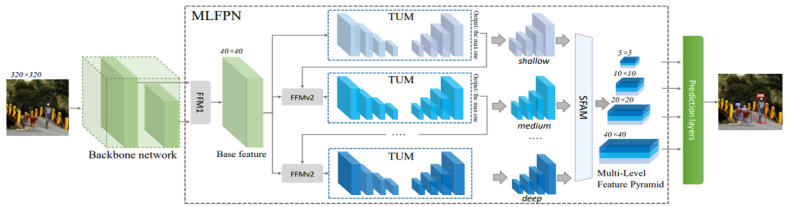
B. Singh et al. “SNIPER: Efficient Multi-Scale Training,” in NeurIPS 2018.
https://papers.nips.cc/paper/8143-sniper-efficient-multi-scale-training.pdf



物体候補が多く位置する領域を拡大

M2Det

Q. Zhao et al. “M2Det: A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network,” in AAAI 2019.
https://arxiv.org/abs/1811.04533



多重解像度の特徴マップ使用

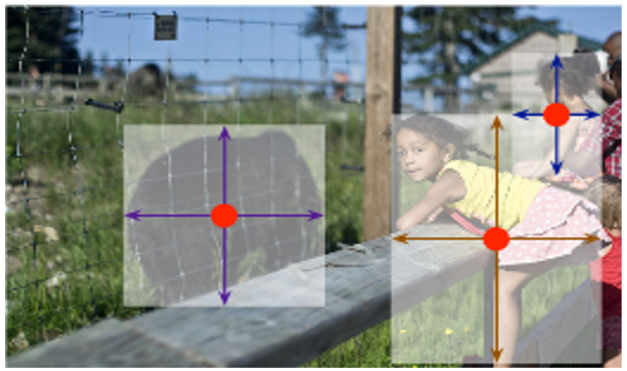
CornerNet

H. Law et al. “CornerNet: Detecting Objects as Paired Keypoints,” in ECCV 2018.
https://arxiv.org/abs/1808.01244

2点（左上, 右下）を回帰

CenterNet

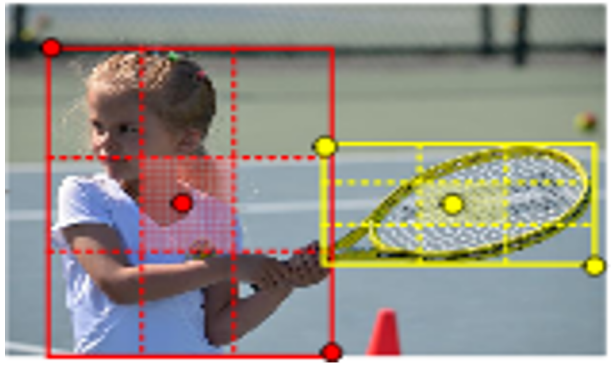
Z. Xingyi et al. “Objects as Points,” in arXiv 2019.
https://arxiv.org/abs/1904.07850



中央からxyサイズを回帰

CenterNet

K. Duan et al. “CenterNet: Keypoint Triplets for Object Detection,” in ICCV 2019.
https://arxiv.org/abs/1904.08189



中央+2点, xyサイズを回帰

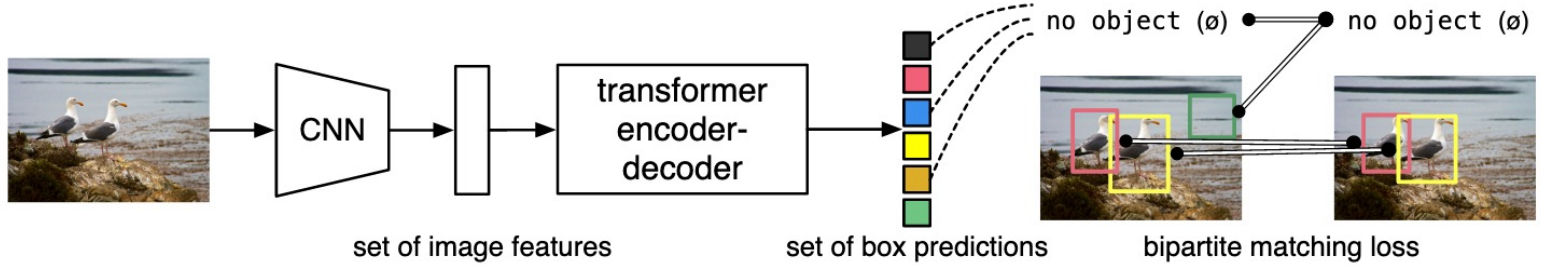
MSCOCO 基準で Mask R-CNN 41.7 -> CenterNet 47.0まで検出精度向上！

DNNの動向・CVのトレンド（13/42）

物体検出の流れ（2020～Present）

DETR

N. Carion et al. “End-to-End Object Detection with Transformers,” in ECCV 2020.
https://arxiv.org/abs/2005.12872

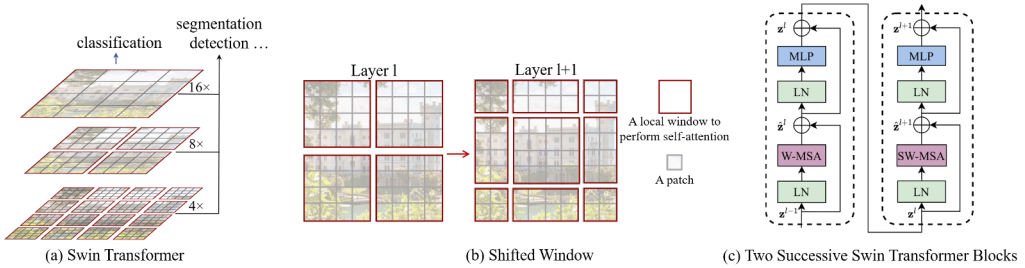


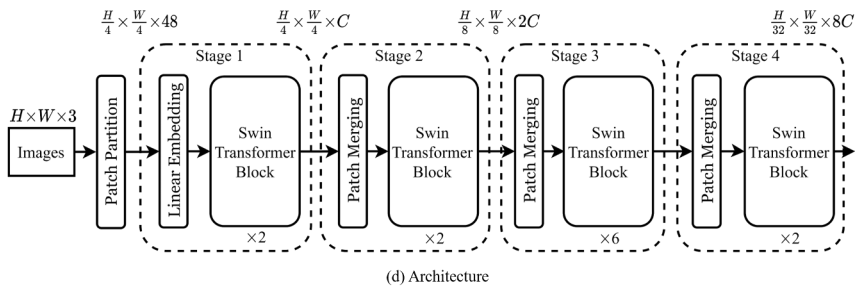
Transformerによる物体検出のパイオニア：Backboneには比較的うまくいくことが明らかなCNNも組み合わせている

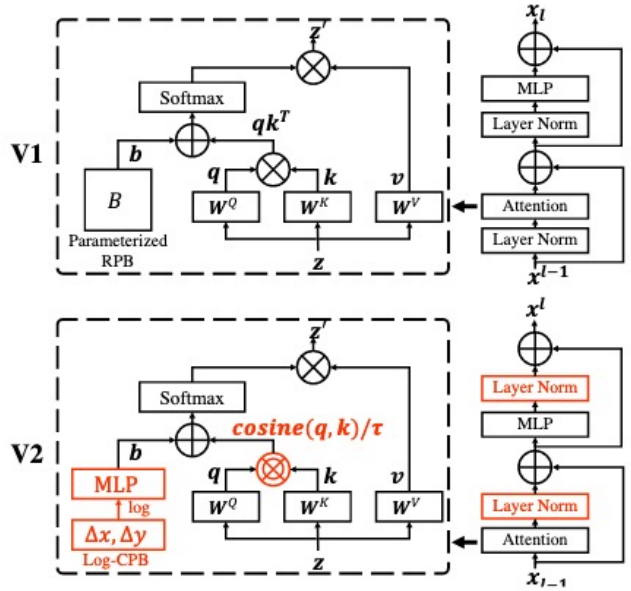
SwinTransformer V1/V2

Z. Liu et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in ICCV 2021.
https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html

Z. Liu et al. “Swin Transformer V2: Scaling Up Capacity and Resolution,” in ICCV 2021.
https://arxiv.org/abs/2111.09883







ViT特有のパッチ入力の問題点を，階層化&シフトにより改善，画像識別のみならず物体検出にも有効

DNNの動向・CVのトレンド（14/42）

セマンティック/インスタンスセグメンテーション（2006~2017）

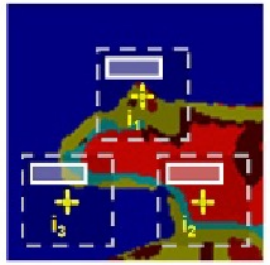
- BackBone（ベースとなるネットワーク）
- Head（ピクセルごとの回帰を出力）
  - 両者の組み合わせ
  - 文脈把握とスケール変動を考慮する傾向

【セマンティック/インスタンスセグメンテーション】

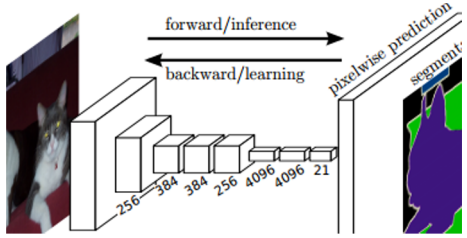
ピクセルごとに画像カテゴリを割り当てる問題設定

※下はセマンティック/インスタンスセグメンテーションを両方含む

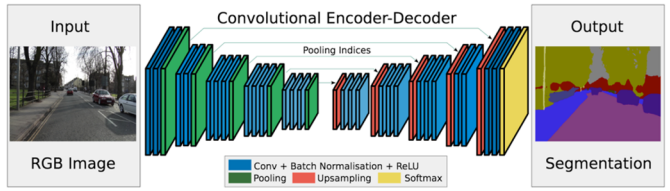
ベースアルゴリズム



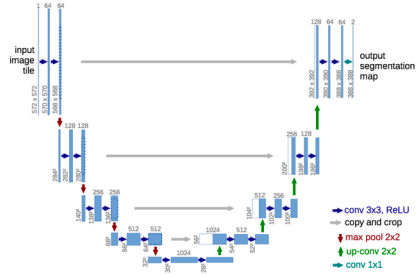
TextonBoost
[Shotton, ECCV06]



FCN [Long, CVPR15]
全層畳み込み，チャネル和

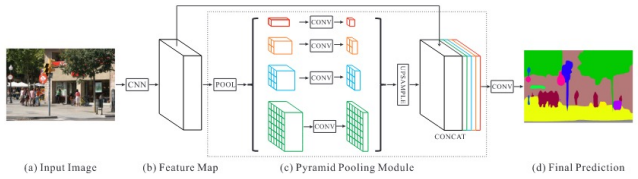


SegNet [Kendall, arXiv15]

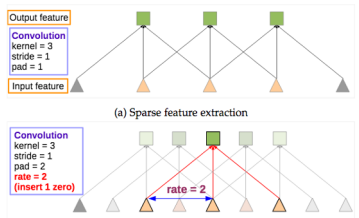


U-Net [Ronneberger, MICCAI15]
位置情報保持，チャネル連結

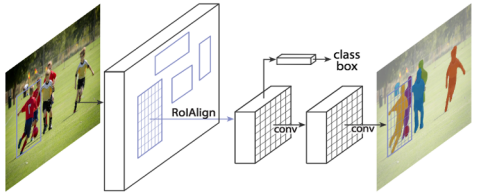
精度重視



PSPNet [Zhao, CVPR17]
特徴マップの階層化，コンテクスト情報



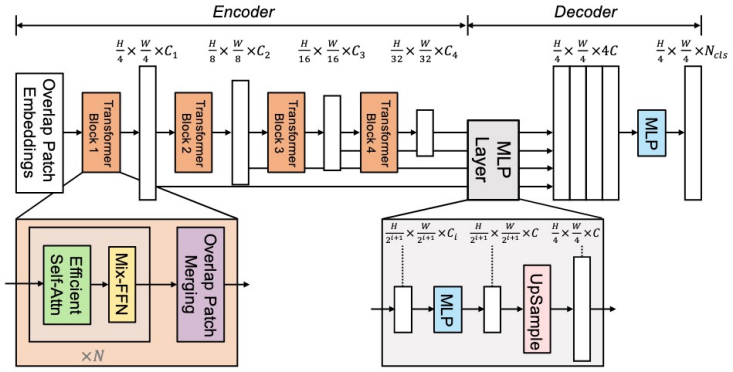
DeepLab(v1,v2,v3) [Chen, TPAMI17]
Dilated Conv, 特徴マップの並列化



Mask R-CNN [He, ICCV17]
RoI Align, Det+Seg

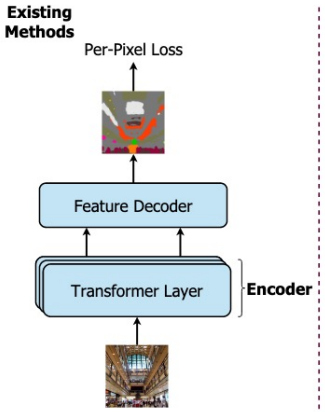
DNNの動向・CVのトレンド（15/42）

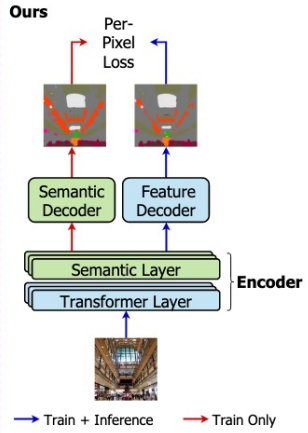
セマンティック/インスタンスセグメンテーションの流れ（2021～Present）



SegFormer

E. Xie et al. “SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers,” in arXiv 2105.15203, 2021.
https://arxiv.org/abs/2105.15203





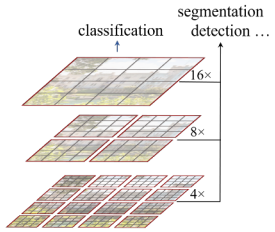
SeMask

J. Jain et al. “SeMask: Semantically Masked Transformers for Semantic Segmentation,” in arXiv 2112.12782 2021.
https://arxiv.org/abs/2112.12782

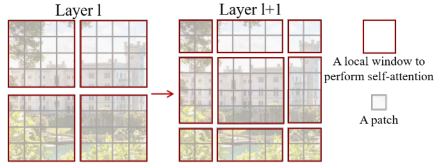
SwinTransformer V1/V2

Z. Liu et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” in ICCV 2021.
https://openaccess.thecvf.com/content/ICCV2021/html/Liu_Swin_Transformer_Hierarchical_Vision_Transformer_Using_Shifted_Windows_ICCV_2021_paper.html

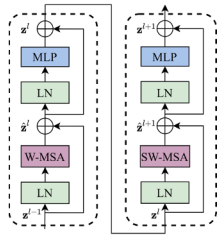
Z. Liu et al. “Swin Transformer V2: Scaling Up Capacity and Resolution,” in ICCV 2021.
https://arxiv.org/abs/2111.09883



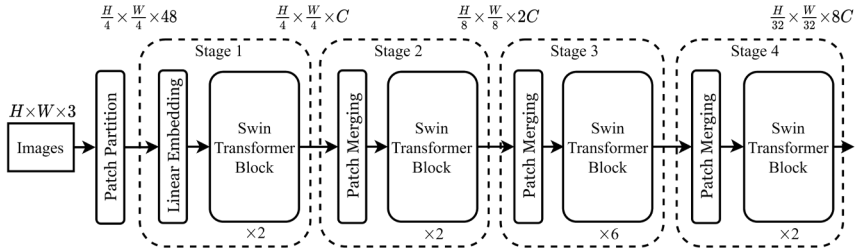
(a) Swin Transformer



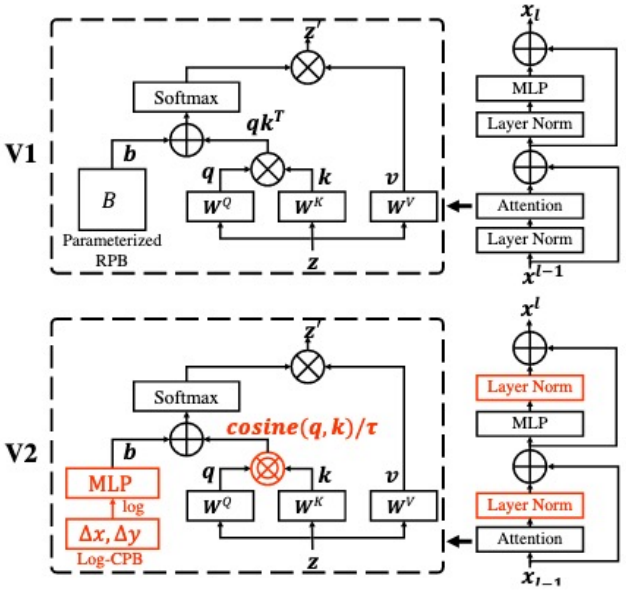
(b) Shifted Window



(c) Two Successive Swin Transformer Blocks



(d) Architecture

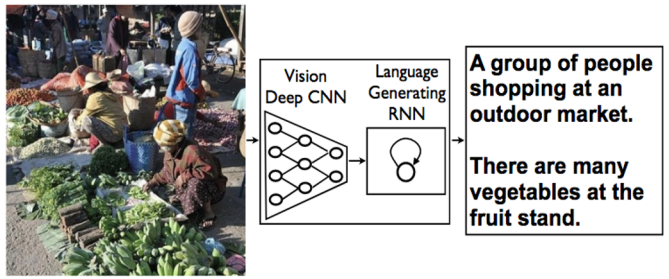


DNNの動向・CVのトレンド（16/42）

CVとNLP（自然言語処理）の融合分野

- 画像説明文（Image Captioning）
- 視覚的質問回答（Visual Question Answering; VQA）
- Visual Dialog

【Image Captioning】



画像を入力として文章を出力

Show and Tell
O. Vinyals et al., “Show and Tell: A Neural Image Caption Generator,” in CVPR 2015.
https://openaccess.thecvf.com/content_cvpr_2015/html/Vinyals_Show_and_Tell_2015_CVPR_paper.html

【VQA】

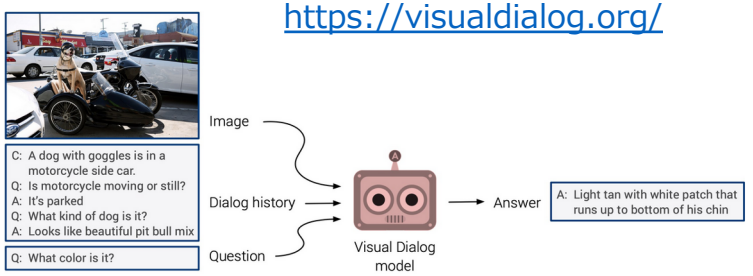
画像 (Visual)と質問文 (Q)を入力，回答 (A)を返却

https://visualqa.org/

VQA
A. Agrawal et al., “VQA: Visual Queation Answering,” in ICCV 2015.
https://arxiv.org/pdf/1505.00468.pdf

【Visual Dialog】
画像とそれに対する対話を繰り返しながら回答を行う

https://visualdialog.org/



Visual Dialog
A. Das et al., “Visual Dialog,” in CVPR 2017.
https://arxiv.org/abs/1611.08669

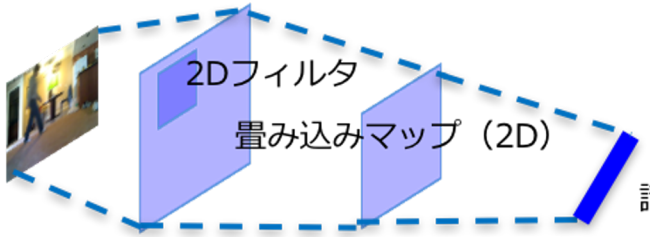
【その他のキーワード】
Text-to-Image（テキストから画像生成），Visual Navigation（画像からの道案内），Cross-modal Summarization（要約タスク）など多数

DNNの動向・CVのトレンド（17/42）

動画像認識のモデル（CNN; N-Dimension Convolution）

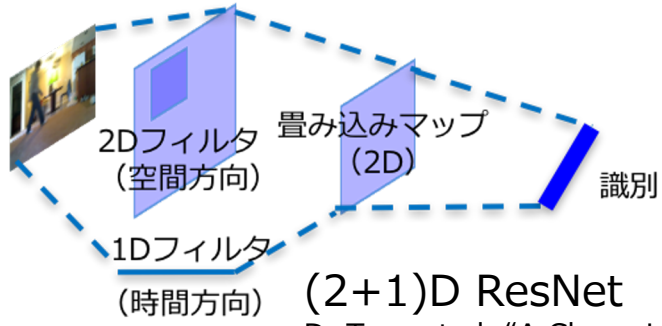
- 2D: Two-Stream ConvNets（フロー画像を同時に準備）
- 2D+再帰モデル: CNN+LSTM（最近ベンチマークとしてしか使われなくなった）
- 3D: 3D CNN（データが用意できればこれが本命）
- (2+1)D（少量データの場合のオプション）

2D畳み込み（TS ConvNet, CNN+LSTM）



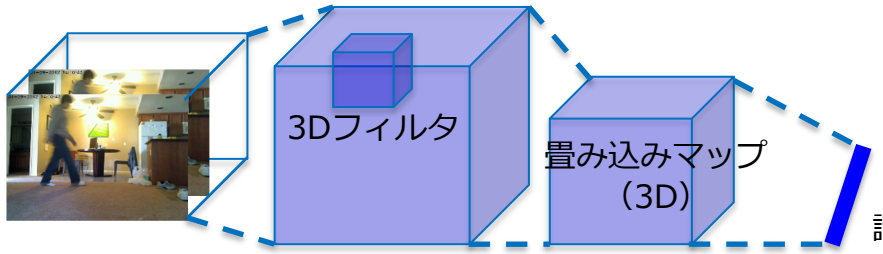
Two-stream ConvNets
K. Simonyan et al., “Two-Stream Convolutional Networks for Action Recognition in Videos,” in NIPS 2014.
https://www.robots.ox.ac.uk/~vgg/publications/2014/Simonyan14b/simonyan14b.pdf

(2+1)D畳み込み（Separable Conv.）



(2+1)D ResNet
D. Tran et al. “A Closer Look at Spatiotemporal Convolutions for Action Recognition,” in CVPR 2018.
https://arxiv.org/abs/1711.11248

3D畳み込み（C3D, 3D ResNet, I3D）

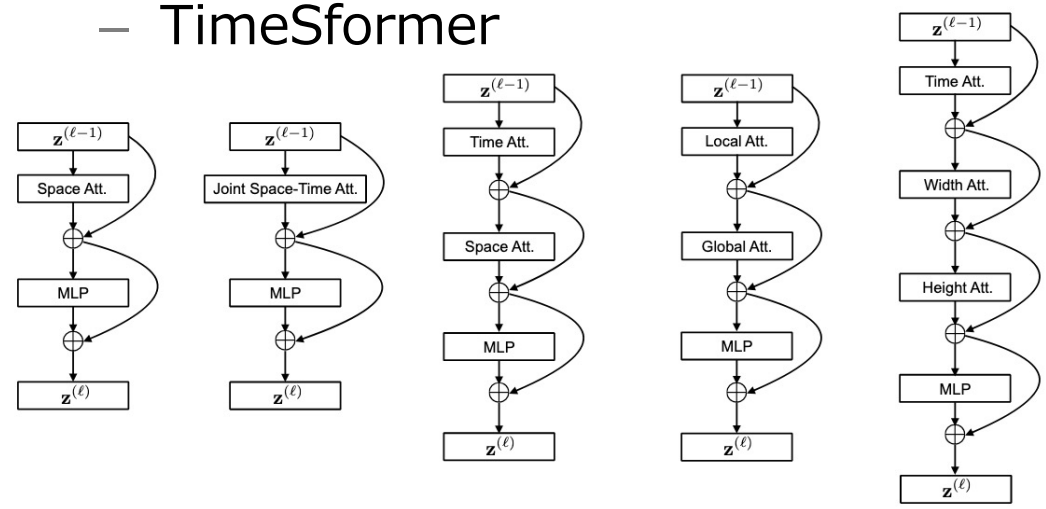


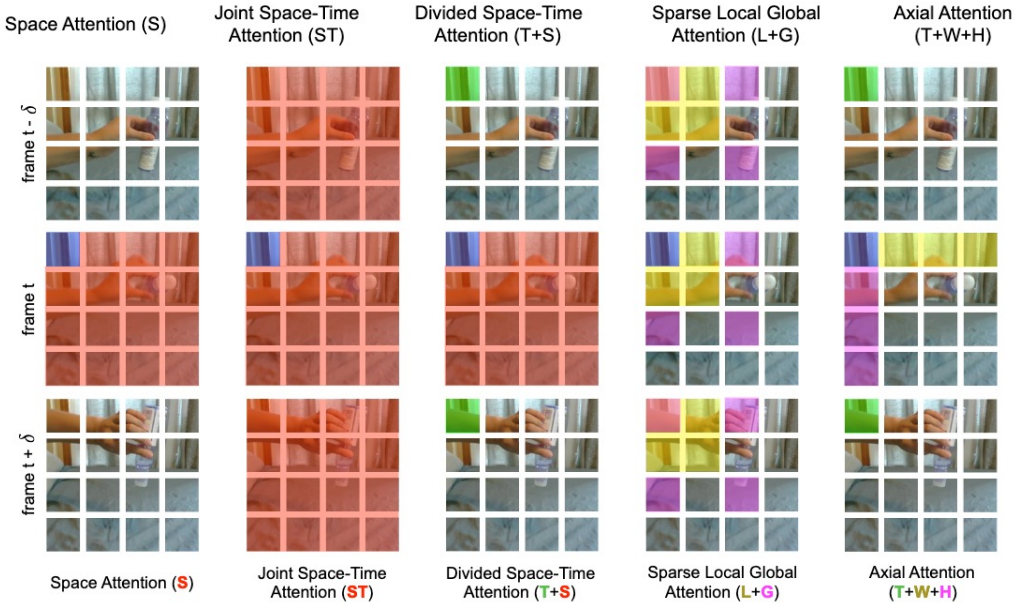
3D ResNet
K. Hara et al. “Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?,” in CVPR 2018.
https://arxiv.org/abs/1711.09577

DNNの動向・CVのトレンド（18/42）

動画像認識のモデル（Video Transformer）

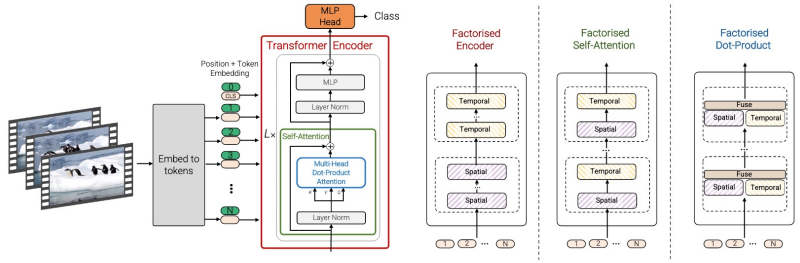
– TimeSformer





TimeSformer
G. Bertasius et al., “Is Space-Time Attention All You Need for Video Understanding?,” in ICML 2021.
https://arxiv.org/abs/2102.05095

ViViT
A. Arnab et al., “ViViT: A Video Vision Transformer,” in ICCV 2021.
https://arxiv.org/abs/2103.15691v2



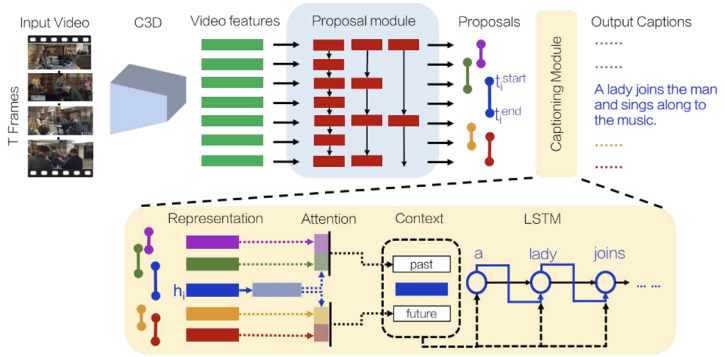
- 時間と空間の処理を実行するモジュールを複数種類比較実験
- 結果から，時間tと空間xyを分割して処理する方式（Divided Space-Time Attention）が有効
- 図中，中央のモジュールとアテンション

DNNの動向・CVのトレンド（19/42）

動画認識タスク

Dense Captioning Events

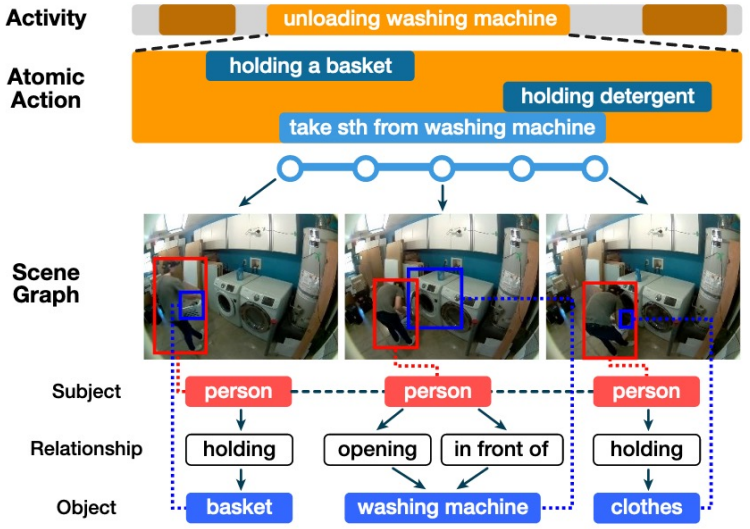
R. Krishna et al. “Dense Captioning Events in Videos,” in ICCV 2017.
https://arxiv.org/abs/1705.00754



動画像から説明文出力，動画像の検索にも用いることができる

Home Action Genome

N. Rai et al. “Home Actioin Genome: Contrastive Compositional Action Understanding,” in CVPR 2021.
https://homeactiongenome.org/



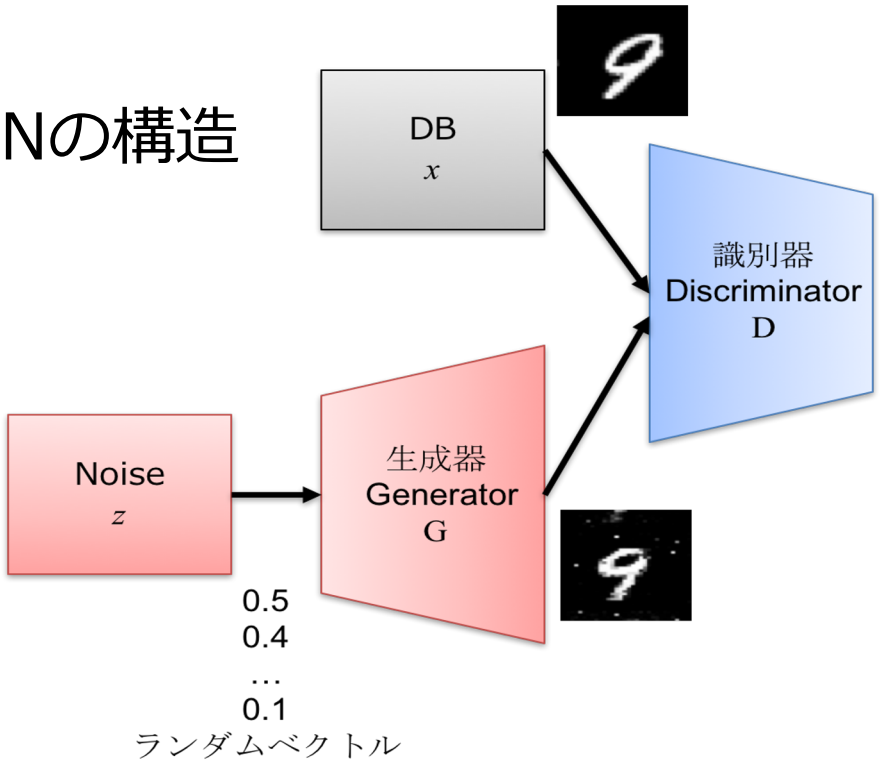
マルチモーダル・複数視点・シーングラフラベル付与により，包括的な動画の理解をサポート

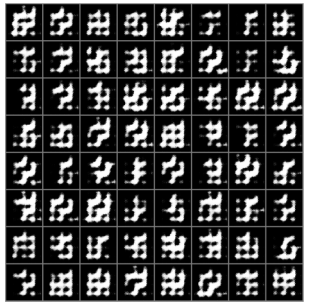
DNNの動向・CVのトレンド（20/42）

GAN: 画像生成を行うための構造として提案

- 現在，生成画像/実画像の分布を近づけることで生成モデルを学習
- 敵対的学習は超解像，異常検知，データ拡張 など多様な場面に応用

GANの構造





徐々に鮮明になるデータ

https://medium.com/@sunnerli/the-missing-piece-of-gan-d091604a615a

（注）下はGANにより生成された画像です

BigGAN
https://arxiv.org/pdf/1809.11096.pdf

DNNの動向・CVのトレンド（21/42）

GANの主要な流れ

1. GAN（オリジナルのGAN）
   - [Goodfellow, NIPS2014] https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
2. DCGAN（畳み込み層の使用）
   - [Radford, ICLR2016] https://arxiv.org/abs/1511.06434
3. Pix2Pix（pixel同士が対応付くという意味でConditionalなGAN）
   - [Isola, CVPR2017] https://arxiv.org/abs/1611.07004
4. CycleGAN（pix2pixの教師なし版）
   - [Zhu, ICCV2017] https://arxiv.org/pdf/1703.10593.pdf
5. ACGAN（カテゴリ識別も同時に実施してコンディションとした）
   - [Odera, ICML2017] https://arxiv.org/abs/1610.09585
6. WGAN/SNGAN（学習安定化）
   - [Arjovsky, ICML2017] http://proceedings.mlr.press/v70/arjovsky17a.html
   - [Miyato, ICLR2018] https://arxiv.org/abs/1802.05957
7. PGGAN（高精度化）
   - [Karras, ICLR2018] https://arxiv.org/abs/1710.10196
8. Self-Attention GAN（アテンション機構を採用）
   - [Zhang, arXiv 1805.08318] https://arxiv.org/abs/1805.08318
9. BigGAN（超高精細GAN）
   - [Brock, ICLR2019] https://arxiv.org/abs/1809.11096
10. StyleGAN（超高精細GAN）
    - [Karras, CVPR2019] https://arxiv.org/abs/1812.04948

2018年12月時点での調査

DNNの動向・CVのトレンド（22/42）

鮮明な画像生成: VAE, Diffusion Model！？

- キレイな画像生成はGANが優勢だった（～2019/05）
  - BigGAN, StyleGAN etc.
- VQVAEでVAEが，2021年はDiffusion Modelが話題になった
  - VAE: GANのようにMode Collapseを起こさないと主張
  - Diffusion Model: 学習時のサンプルノイズを徐々に取り除くことで生成の多様性と信頼性を向上（BigGAN よりも高精度であることから注目）

左：VQ-VAE-2，右：BigGAN

A. Razavi et al. “Generating Diverse High-Fidelity Imageswith VQ-VAE-2,” NeurIPS, 2019. https://arxiv.org/pdf/1906.00446.pdf

+Diffusion Model

P. Dhariwal et al. “Diffusion Models Beat GANs on Image Synthesis,” arXiv, 2105.05233, 2021. https://arxiv.org/abs/2105.05233

DNNの動向・CVのトレンド（23/42）

DALL·Eの衝撃

– テキストから画像を生成するText-to-Imageの問題設定で革新的な成果

- 「アボカド形の腰掛け椅子」のような見たことがない物体も概念を組み合わせて画像を生成
- 1200億パラメータのモデルをインターネットから集めた2.5億の画像・テキストのペアから学習

DALL·E: Creating Images from Text
https://openai.com/blog/dall-e/

DNNの動向・CVのトレンド（24/42）

教師なし/少量教師あり学習への拡がり

- キーワード
  - {Un-, Weak-, Semi-, Self-} supervision
  - {Zero-, One-, Few-} shot learning
  - Transfer Learning
  - Domain Adaptation
  - Reinforcement Learning
- 教師がない/間接的に教師を与える，ような仕組みに対する競争も激化
- 巨大IT企業のように大量のラベルを持たなくても学習を成功させる
  - アルゴリズム至上主義への回帰？

DNNの動向・CVのトレンド（25/42）

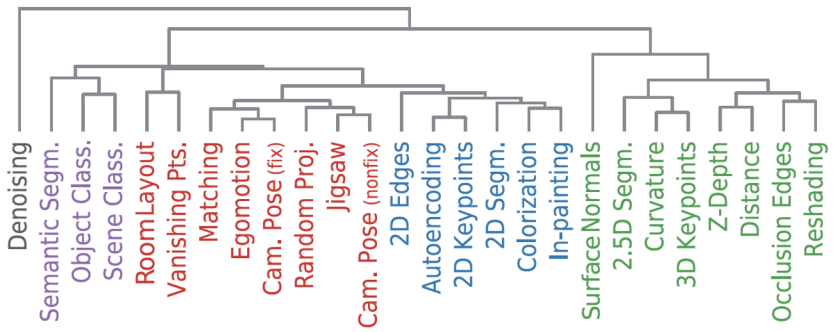
学習法の簡単な整理

- {Un-, Semi-, Weak-, Self-} supervision
  - Un-supervision（教師なし学習）
    アノテーションが一切ないデータで学習
  - Semi-supervision（半教師あり学習）
    アノテーションを持つデータと持たないデータで学習
  - Weak-supervision（弱教師付き学習）
    出力として必要な情報よりも拘束力の弱いデータを用いて学習
    - ex) 物体検出を行う際に画像ラベルのみを用いて学習
  - Self-supervision（自己教師あり学習）
    自ら教師を作り出して特徴表現を学習する「自己教師学習」
    - 特定タスクの前に自ら教師を作り出し特徴表現を学習するため，その後に特定タスクのためのファインチューニングを伴う
    - ex) 領域分割した画像でジグソーパズルを解く，回転を当てる

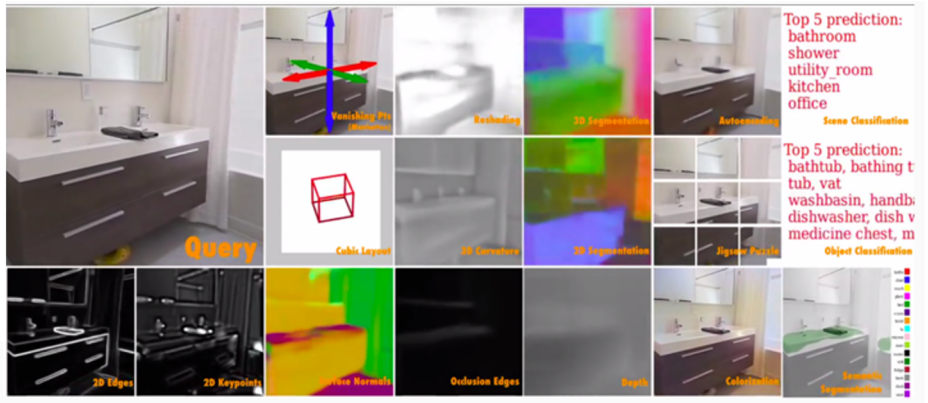
DNNの動向・CVのトレンド（26/42）

転移学習（Transfer Learning）

- 学習した知識を別の領域の学習に適用する技術
- 転移学習の網羅的探索：Taskonomy [Zamir, CVPR2018] http://taskonomy.stanford.edu/
  - CVPR 2018 Best Paper Award
  - 26種のタスク間の関連性を調べる
    - CVの歴史の中で別々に議論されていたサブタスクを繋げる
    - 効果を最大化する転移学習の関係性を明らかにした



Task Similarity Tree: 類似するタスク間の関係性を可視化

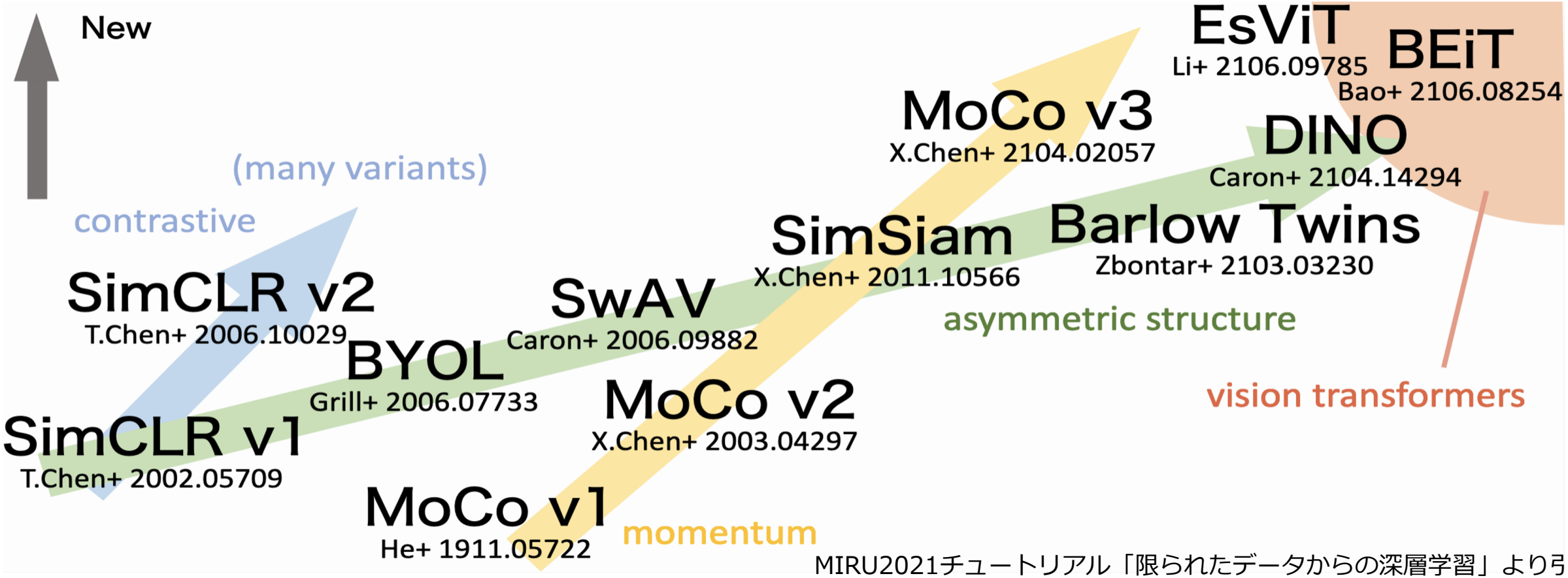


データセットは26タスクに対しラベル付け

DNNの動向・CVのトレンド（27/42）

最近の主流は自己教師あり学習

– 一部で人間の教師を置き換える性能まで到達



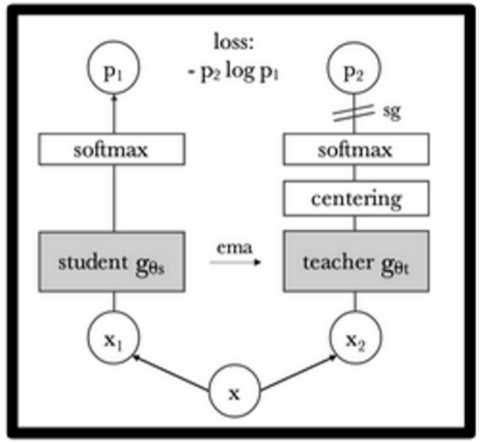
MIRU2021チュートリアル「限られたデータからの深層学習」より引用

https://raw.githubusercontent.com/hirokatsukataoka16/Formula-Driven-DataBase-Group/main/docs/material/MIRU21_Tutorial1.pdf

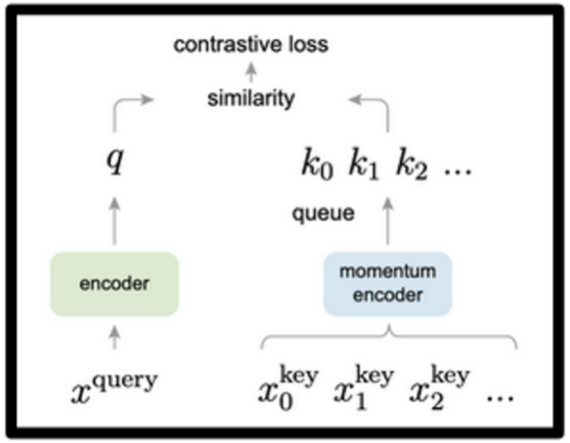
DNNの動向・CVのトレンド（28/42）

2022年 CV分野の激戦領域：ViTの自己教師あり学習

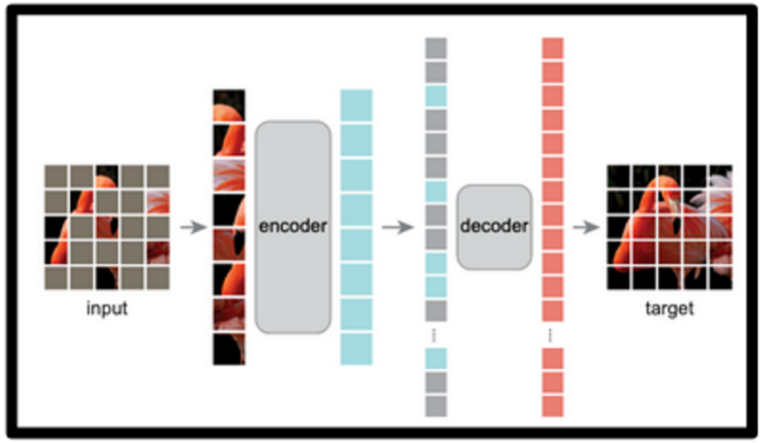
- 実画像に対して自動で一貫した教師を付与
  - ➤ DINO：ラベルなしでの自己蒸留 https://arxiv.org/abs/2104.14294
  - ➤ MoCoV3：対照学習MoCoと同様だが，ViT向けに改善 https://arxiv.org/abs/2104.02057
  - ➤ MAE：パッチの復元（BERTの画像版タスク） https://arxiv.org/abs/2111.06377



PT: DINO + ViT-B/16
FT: 78.2 @ ImageNet-1k val.
(ViT-B/8では80.1まで向上)



PT: MoCoV3 + ViT-B/16
FT: 83.2 @ ImageNet-1k val.



PT: MAE + ViT-B/16
FT: 83.6 @ ImageNet-1k val.
(ViT-H448では87.8まで向上)

マルチモダリティにて有効なData2Vecも登場（上記のスコアで84.2）
https://arxiv.org/abs/2202.03555

DNNの動向・CVのトレンド（29/42）

数式ドリブン教師あり学習 (FDSL; Formula-Driven Supervised Learning)

- 生成規則から画像パターンと教師ラベルを生成，大規模画像データセット構築
  - ➤ 下記はフラクタル幾何を生成規則としたFDSL
  - ➤ ベースとなるカテゴリ数/インスタンス数は1k/1kの100万画像

生成規則(式)

ラベルの生成 (ランダムサンプリング)

$$\Theta = \{(\theta_i, p_i)\}_{i=1}^{N}$$

データの生成 (IFS)

$$\mathrm{IFS} = \{\mathcal{X}; w_1, w_2, \cdots, w_N; p_1, p_2, \cdots, p_N\}$$

$$w_i(\boldsymbol{x}; \theta_i) = \begin{bmatrix} a_i & b_i \\ c_i & d_i \end{bmatrix} \boldsymbol{x} + \begin{bmatrix} e_i \\ f_i \end{bmatrix}$$

$$p_i = p(w^* = w_i) \qquad \boldsymbol{x}_{t+1} = w^*(\boldsymbol{x}_t)$$

FDSL / FractalDB

H. Kataoka et al., “Pre-training without Natural Images,” in IJCV 2022.
https://hirokatsukataoka16.github.io/Pretraining-without-Natural-Images/

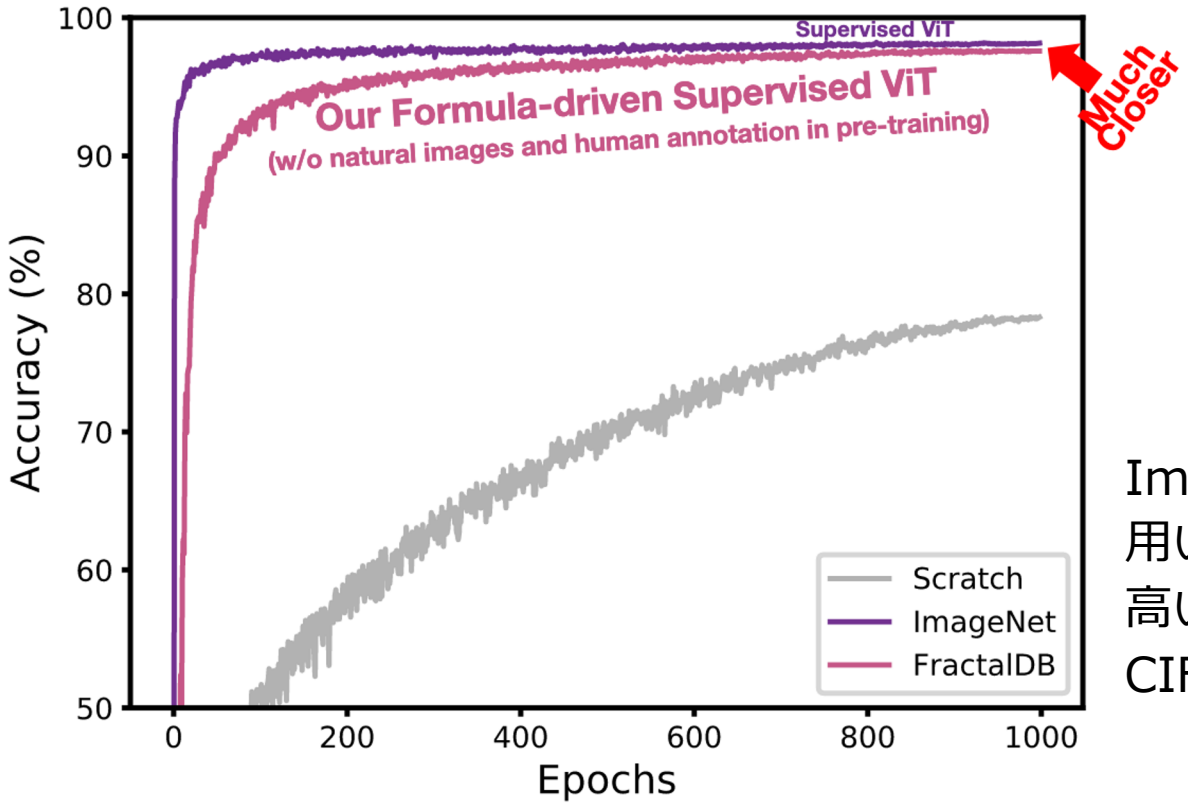
FractalDB Pre-trained ViT

K. Nakashima et al., “Can Vision Transformers Learn without Natural Images,” in AAAI 2022.
https://hirokatsukataoka16.github.io/Vision-Transformers-without-Natural-Images/

DNNの動向・CVのトレンド（30/42）

数式ドリブン教師あり学習 (FDSL; Formula-Driven Supervised Learning)

- 実画像+人間教師を用いない事前学習でImageNet事前学習に近接する精度に到達
- 理論上は無限に画像カテゴリ・インスタンスを生成可能
  - 今後，大規模分散学習によりデータセット・モデルサイズを巨大化予定

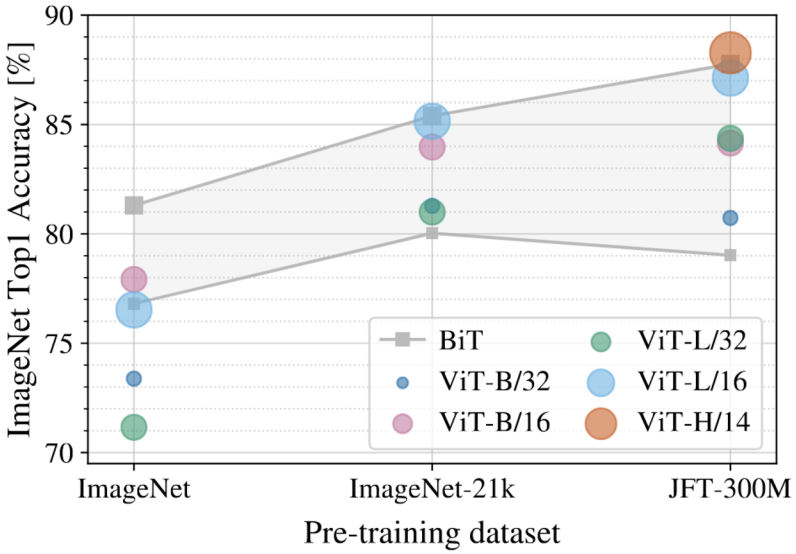


ImageNet vs. FractalDB：実画像も人間による教師ラベルも用いないFractalDBによる学習でScratch学習よりも圧倒的に高い精度，かつImageNetに近い精度まで到達．例えば，CIFAR-10ではImageNet 98.0 vs. FractalDB 97.8.

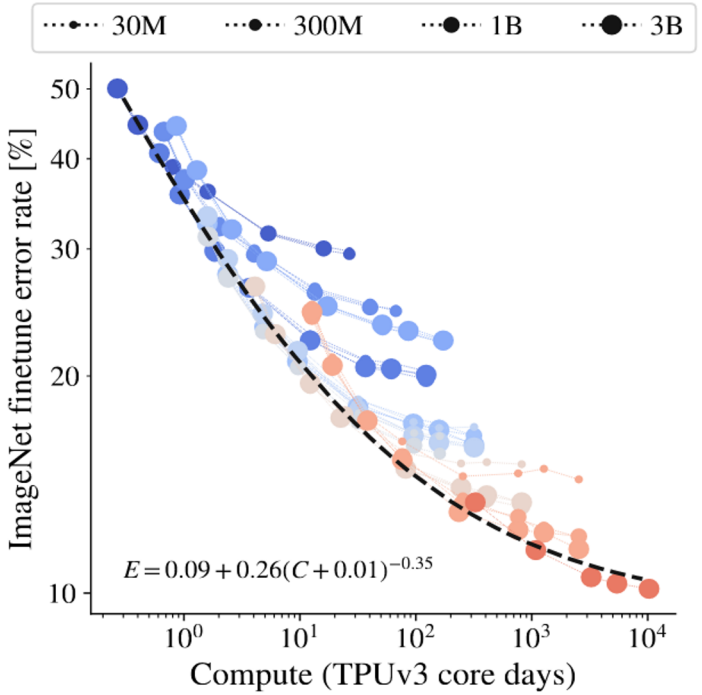
DNNの動向・CVのトレンド（31/42）

超大規模データセット JFT-300M/3B

- JFT-300M/3Bはそれぞれ３億/３０億画像含む画像データセット



Vision Transformers,
https://arxiv.org/pdf/2010.11929.pdf
ImageNet-1k Top-1 Accuracy: 88.55 (ViT-H/14)



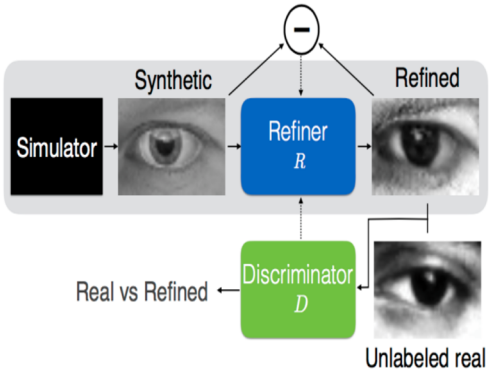
Scaling Vision Transformers,
https://arxiv.org/pdf/2106.04560.pdf
ImageNet-1k Top-1 Accuracy: 90.45 (ViT-G/14)

自己教師/数式教師は人間教師+超大規模画像データセットに匹敵するのか？

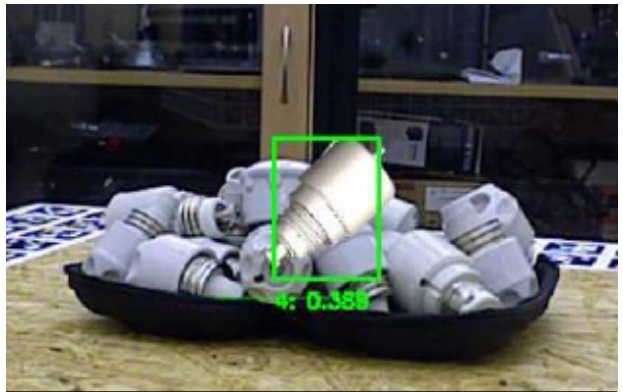
DNNの動向・CVのトレンド（32/42）

生成データ学習への期待

- CGを敵対的学習によりリアルに近づける（SimGAN）
- CGから大量の画像を生成（Domain Randomization）
- 切抜き画像をいかに自然に別画像に挿入するか（Cut-and-Paste学習）
- ノイズ画像と自己教師あり学習で視覚特徴を学習（Visual Representation）

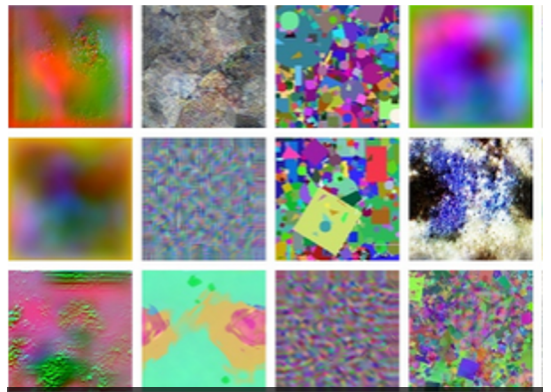


[Shrivastava, CVPR2017] BP
CGをより写実的に表現するRefinerを適用



[Sundermeyer, ECCV2018]Oral,BP
ラベル無しCGデータで実時間6D検出，さらに教師有りを倒した

[Remez, ECCV2018]Oral
Cut/Pasteで既存セグメントラベルを増加，教師有りに接近する精度

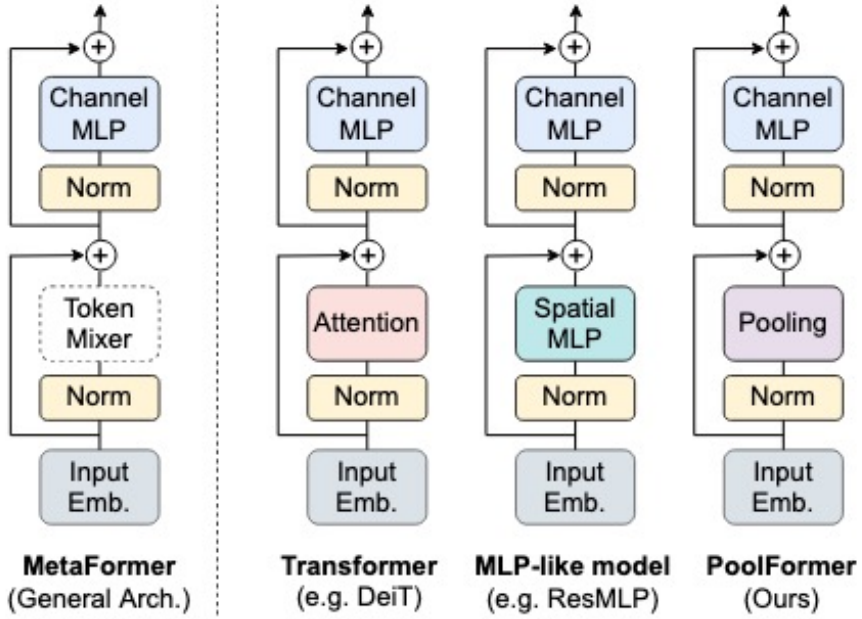


[Baradad, NeurIPS21]Spotlight
ノイズ画像と自己教師で視覚特徴を獲得可能

Keyword: Synthetic Data, Adversarial Learning, Data Augmentation, Domain Randomization, Formula-driven Supervised Learning

DNNの動向・CVのトレンド（33/42）

モデル / データの再考



(a)

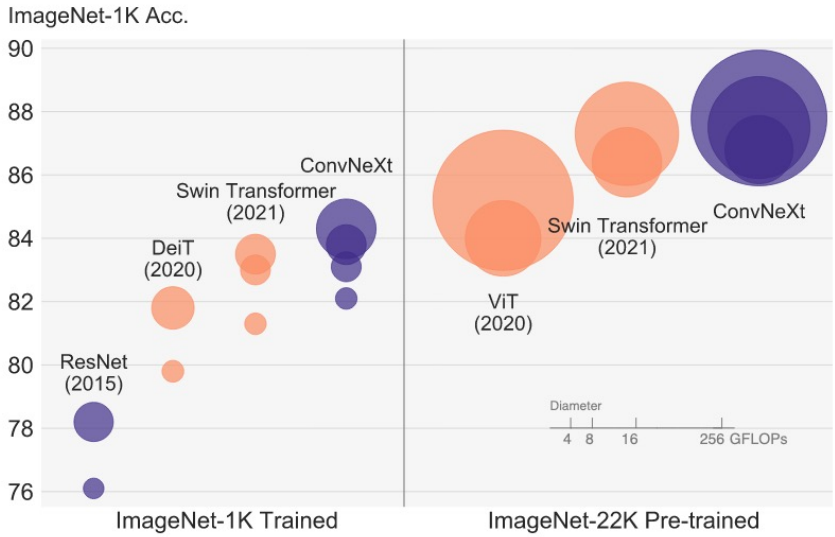
MLP-Mixer

I. Tolstikhin et al. “MLP-Mixer: An all-MLP Architecture for Vision,” in NeurIPS 2021.
https://arxiv.org/abs/2105.01601

MetaFormer

W. Yu et al. “MetaFormer is Actually What You Need for Vision,” in arXiv:2111.11418, 2021.
https://arxiv.org/abs/2111.11418

ViTの精度/機能は他のモジュールでも再現できるのではないか？多層パーセプトロン / Poolingで自己注意と同等のことができる。
→本質はどこにあるのか？

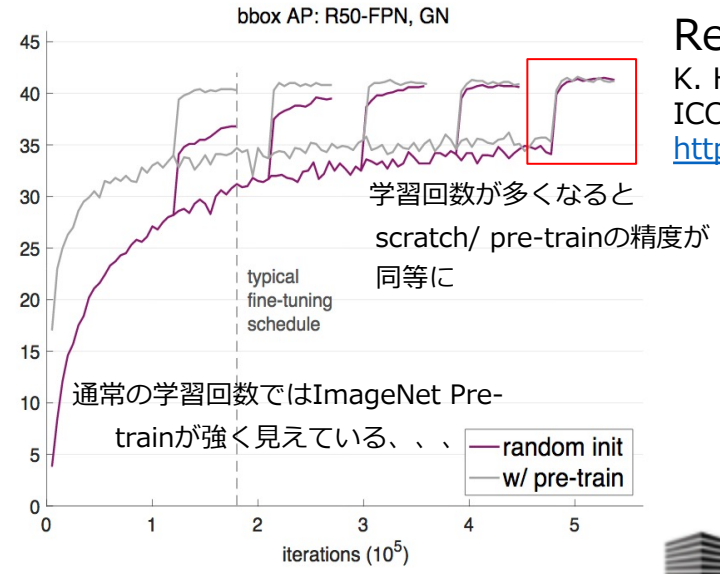


ConvNeXt

Z. Liu et al. “A ConvNet for the 2020s,” in arXiv:2201.03545, 2022.
https://arxiv.org/abs/2201.03545v1

CNNでも改善すればViTと同等以上にまで改善

- グループ化 / チャンネル数増加
- ボトルネック改善
- カーネルサイズ
- など



Rethinking ImageNet Pre-training

K. He et al. “Rethinking ImageNet Pre-training,” in ICCV, 2019.
https://arxiv.org/abs/1811.08883

ImageNetは他のタスクの精度向上に貢献する？

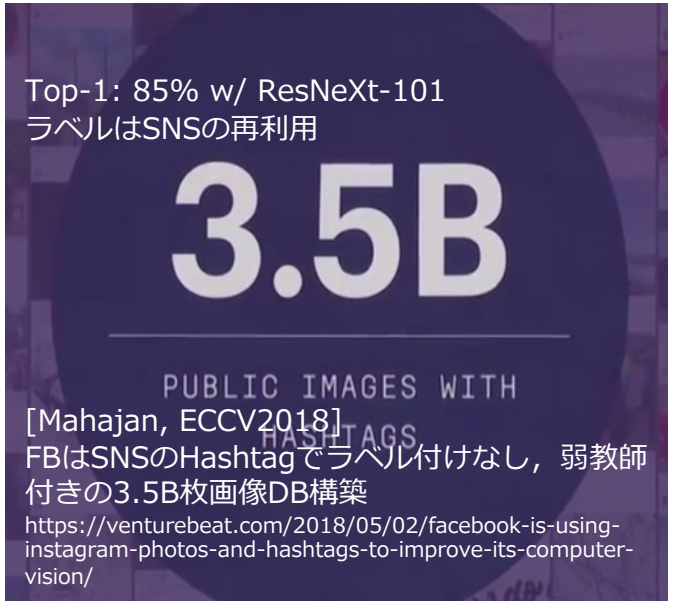
- 検出ではしない（左図参照）
- スクラッチで長く学習すれば同等の精度まで到達
  - » ただし，10K以上のラベルは必要
- 収束の高速化は分野に貢献

DNNの動向・CVのトレンド（34/42）

画像/動画DBの大規模化

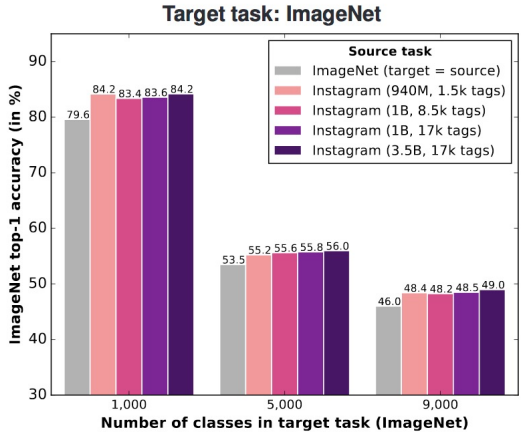
- 画像識別において，35億画像を含むデータセットが存在
- その他，物体検出や動画認識においても大規模化

【画像識別 / 動画認識】



[Mahajan, ECCV2018]
FBはSNSのHashtagでラベル付けなし，弱教師付きの3.5B枚画像DB構築
https://venturebeat.com/2018/05/02/facebook-is-using-instagram-photos-and-hashtags-to-improve-its-computer-vision/

無/弱/半教師付きの文脈で大量画像とその教師を与えられればモデルを強化できる

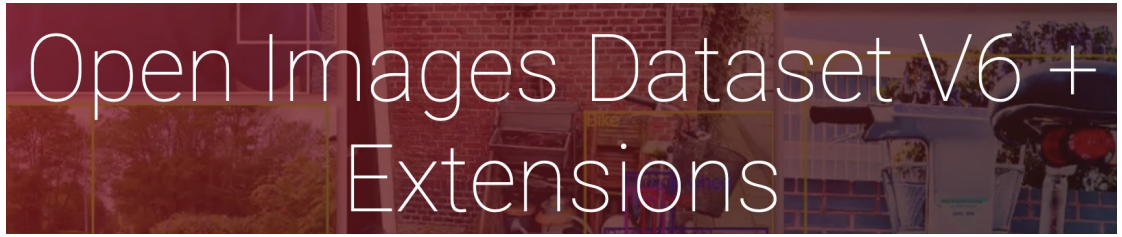


Instagram-3.5B / 65M

D. Mahajan et al. “Exploring the Limits of Weakly-Supervised Pretraining,” in ECCV, 2018.
https://arxiv.org/abs/1805.00932

D. Ghadiyaram et al. “Large-scale Weakly-Supervised Pre-Training for Video Action Recognition,” in ICCV, 2019.
https://openaccess.thecvf.com/content_CVPR_2019/html/Ghadiyaram_Large-Scale_Weakly-Supervised_Pre-Training_for_Video_Action_Recognition_CVPR_2019_paper.html

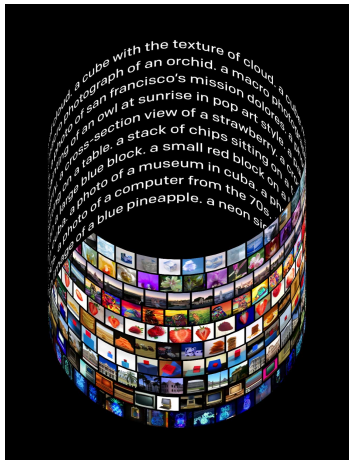
【物体検出】



OpenImages

I. Krasin et al. “OpenImages: A public dataset for large-scale multi-label and multi-class image classification,” 2017.
https://storage.googleapis.com/openimages/web/index.html

【画像生成】



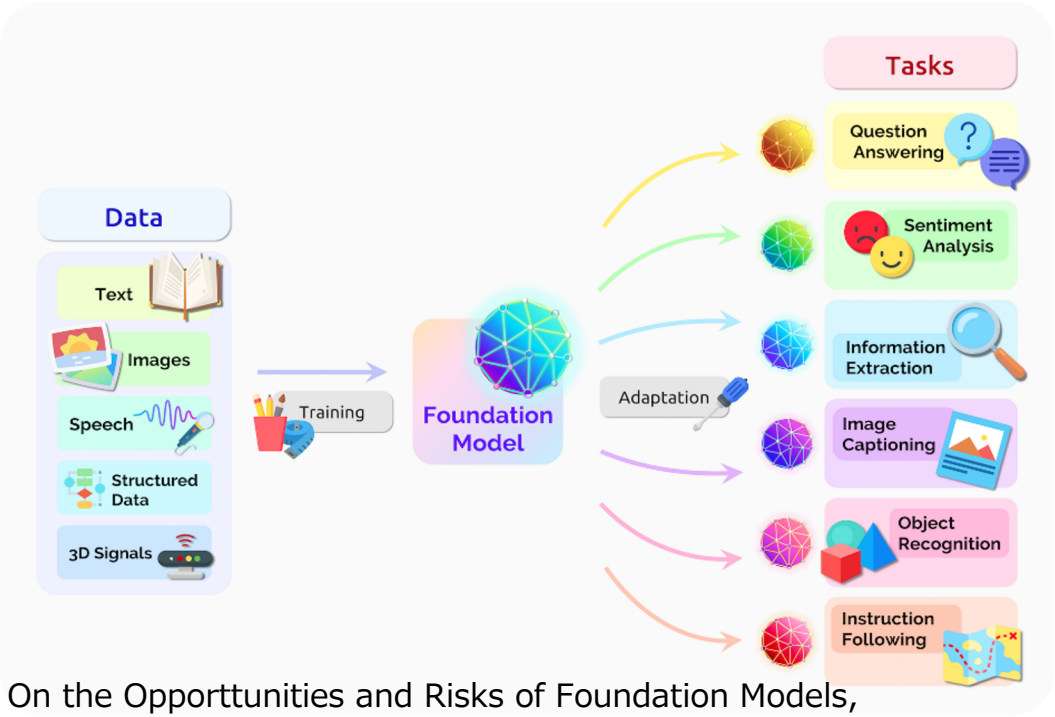
物体検出向けのOpenImageV6は15M画像・座標，DALL·E（前述）は250M画像・テキストを含むデータセットを構築

DALL·E: Creating Images from Text
https://openai.com/blog/dall-e/

DNNの動向・CVのトレンド（35/42）

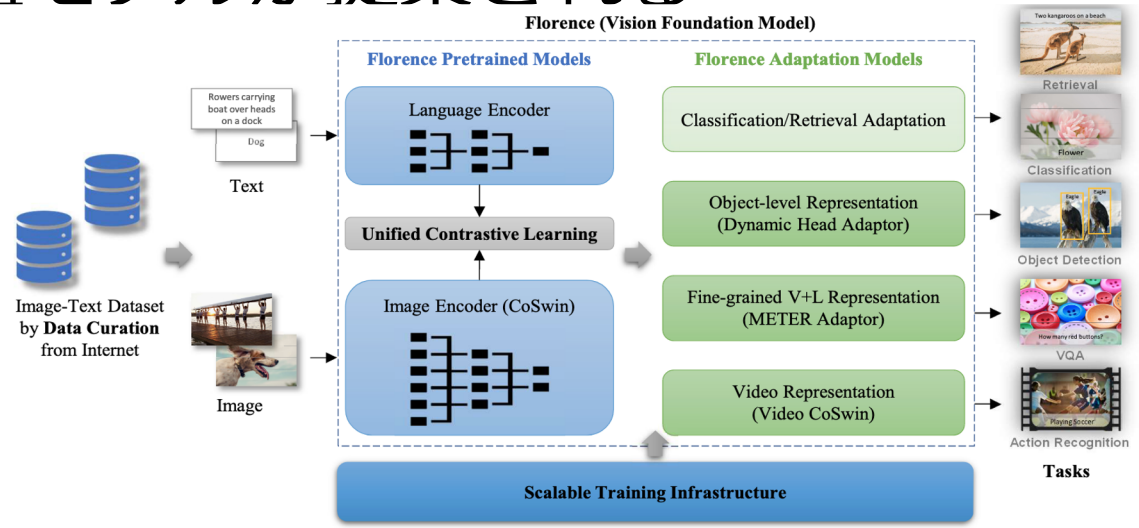
Foundation Model基盤モデル: 汎用適に適用可能なモデル(汎用人工知能への挑戦)

- NLP分野ではBERT, GPT-{1, 2, 3}, DALL-Eなど
- CV分野では最近Florenceなる基盤モデルが提案される



On the Opporttunities and Risks of Foundation Models, https://arxiv.org/abs/2108.07258

- 基盤モデルのホワイトペーパー的立ち位置
- FMの可能性と危険性を議論



Florence: A New Foundation Model for Computer Vision, https://arxiv.org/pdf/2111.11432.pdf

- 44タスクを処理可能
- 9億の画像・言語ペアデータを構築

→より最近ではPolyViT(2111.12993) / NuWA(2111.12417) / Omnivore (2201.08377)なるモデルが同時多発的に提案

DNNの動向・CVのトレンド（36/42）

High Performance Computing: 研究の加速

- AWS/Azure/Google Cloud, 日本でもTsubame3.0/ABCI/富岳
  - Tsubame3.0: 2,160 GPUs / 540 Nodes
    - NVIDIA Tesla P100 x 2,160
  - ABCI: 5,312 GPUs / 1,208 Nodes
    - NVIDIA V100 x 4,352 + NVIDIA A100 x 960
- ImageNet 世界最速記録の変遷
  - 29h > 1h > 30m > 15m > 6.6m > 1.8m > 2.0m > 1.2m

東工大TSUBAME 3.0

産総研ABCI

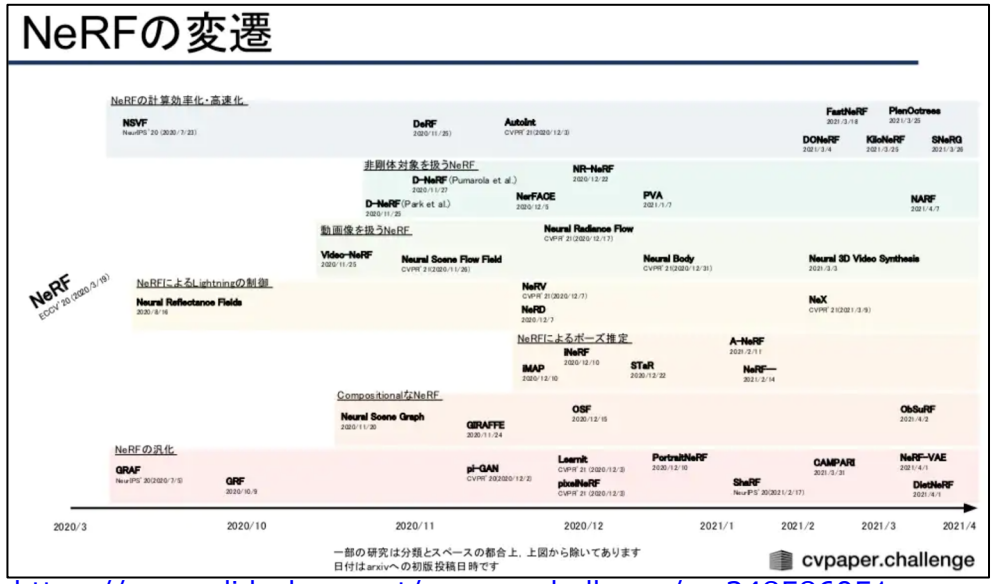
	Batch Size	Processor	DL Library	Time	Accuracy
He et al. [1]	256	Tesla P100 × 8	Caffe	29 hours	75.3 %
Goyal et al. [2]	8,192	Tesla P100 × 256	Caffe2	1 hour	76.3 %
Smith et al. [3]	8,192 → 16,384	full TPU Pod	TensorFlow	30 mins	76.1 %
Akiba et al. [4]	32,768	Tesla P100 × 1,024	Chainer	15 mins	74.9 %
Jia et al. [5]	65,536	Tesla P40 × 2,048	TensorFlow	6.6 mins	75.8 %
Ying et al. [6]	65,536	TPU v3 × 1,024	TensorFlow	1.8 mins	75.2 %
Mikami et al. [7]	55,296	Tesla V100 × 3,456	NNL	2.0 mins	75.29 %
This work	**81,920**	**Tesla V100 × 2,048**	**MXNet**	**1.2 mins**	**75.08%**

M. Yamazaki, et al. “Yet Another Accelerated SGD: ResNet-50 Trainingon ImageNet in 74.7 seconds,” arXiv pre-print, 1903.12650, 2019.
https://arxiv.org/pdf/1903.12650.pdf

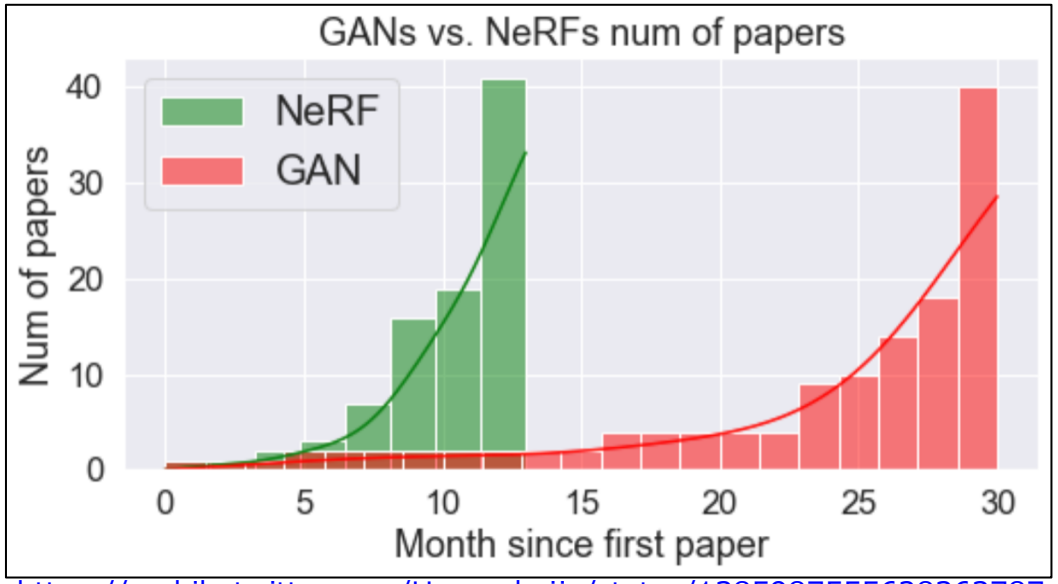
DNNの動向・CVのトレンド（37/42）

Neural Radiance Fields（NeRF）

- 三次元座標と視点方向を入力し，輝度と密度への関数として表現
- 2020年3月にarXivにて公開されて以降，急激に応用・派生研究が増加
- ECCV 2020 Best Paper Honorable Mentionを受け爆発的に拡がる



https://www.slideshare.net/cvpaperchallenge/ss-248586051



https://mobile.twitter.com/Hassanhaija/status/1385987555628363787

画像は相澤先生（広島大）のサーベイ https://www.slideshare.net/cvpaperchallenge/ss-248586051 より

なお発表者はNeRFについては疎いので目下勉強中です

DNNの動向・CVのトレンド（38/42）

人工知能は現在も進化の一途を辿り，社会実装が進む

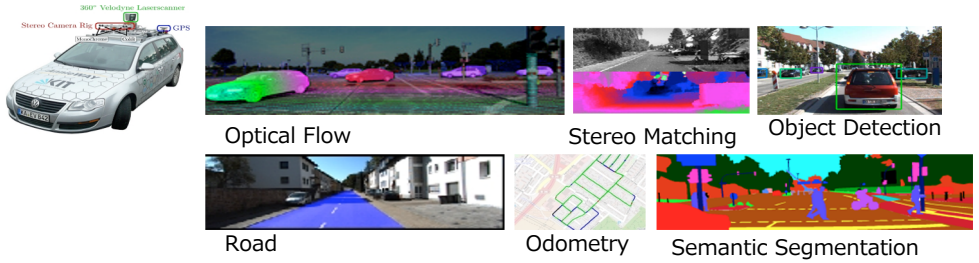
- 自動運転/ADAS
- ロボティクス
- ファッション
- 画像/動画検索
- XR（VR, AR）, Metaverse
- 等

研究者としては「こんなこともできる」を世に出したい

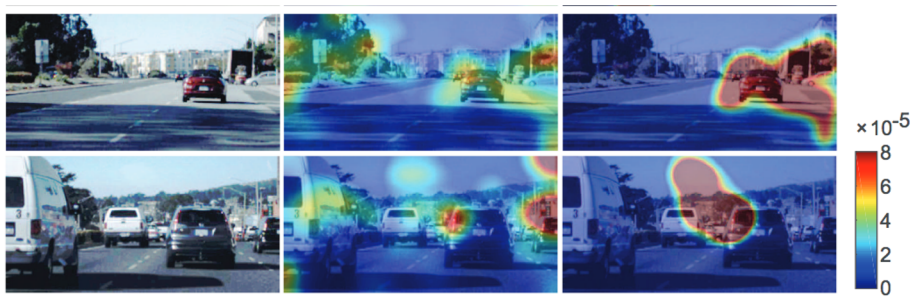
DNNの動向・CVのトレンド（39/42）

自動運転/ADAS（Self-Driving Cars/ADAS）

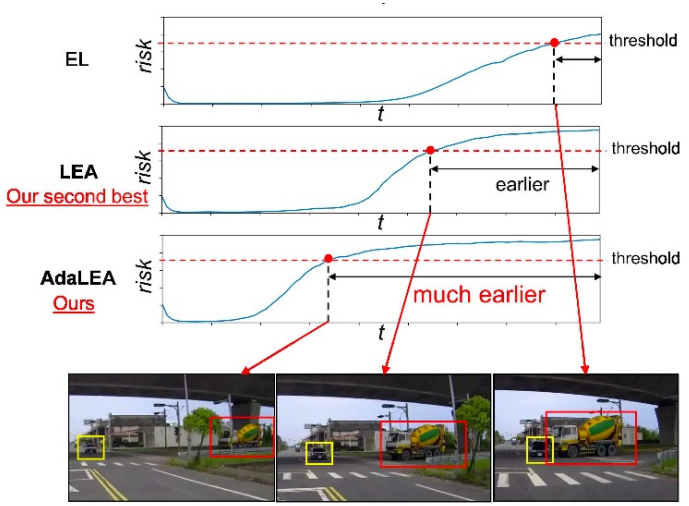
- 国際会議の研究（検知など単純タスク）は減少傾向，実利用に向け開発？
- 数年前まではKITTI datasetに対しての精度競争が盛ん
- 現在は自動運転の解釈性，ニアミスシーンの解析等



KITTI
A. Geiger et al. “Are we ready for autonomous driving? The KITTI vision benchmark suite,” in CVPR 2012.
https://ieeexplore.ieee.org/document/6248074
物体検出，ステレオ視，セグメンテーション問題を提供



[Kim, ICCV2017]
自動運転時の解釈性：物体検知の際にどこを参照



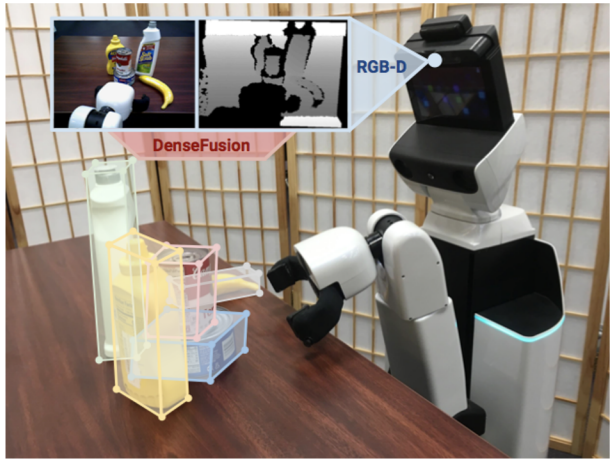
[Suzuki, CVPR2018]
ニアミス・事故シーンを認識，予測

自動運転業界も続々スタートアップ企業が立ち上がり盛り上がりを見せている

DNNの動向・CVのトレンド（40/42）

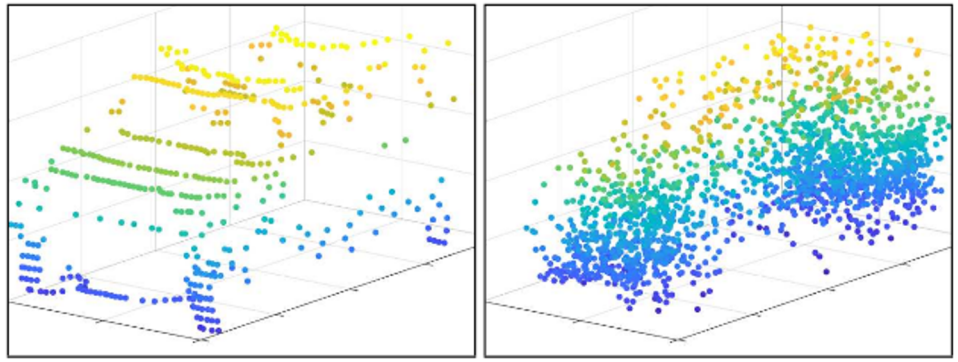
ロボティクスへの応用

- 距離マップ，点群等の利用など3D Visionが盛んに議論
- マニピュレーション：DenseFusionでは6D Det.と把持を実施（左図）
- 3D点群による形状復元・物体検出・追跡（右図）



C. Wang et al. “DenseFusion: 6D Object Pose Estimation by Iteratibve 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

点群トラッキング+形状復元（下の例は車両の復元）



S. Giancola et al. “Leveraging Shape Completion for 3D Siamese Tracking,” in CVPR 2019.
http://openaccess.thecvf.com/content_CVPR_2019/papers/Giancola_Leveraging_Shape_Completion_for_3D_Siamese_Tracking_CVPR_2019_paper.pdf

ロボティクスも発展，物流業界などに応用されている

DNNの動向・CVのトレンド（41/42）

ファッション分野への応用

- ファッション分野のデータ整備が進展
- DeepFashion2（左図）
  - DeepFashionの強化版，服装画像により詳細なラベル付与
- FCDB（右図）
  - 世界のファッショントレンド解析向けのデータ

Y. Ge et al. “DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation,Segmentation and Re-Identification of Clothing Images,” in CVPR 2019.
http://openaccess.thecvf.com/content_CVPR_2019/papers/Ge_DeepFashion2_A_Versatile_Benchmark_for_Detection_Pose_Estimation_Segmentation_and_CVPR_2019_paper.pdf

H. Kataoka, K. Abe, M. Minoguchi, A. Nakamura, Y. Satoh, "Ten-million-order Human Database for World-wide Fashion Culture Analysis", in CVPR 2019 Workshop on FFSS-USAD.
http://openaccess.thecvf.com/content_CVPR_2019/papers/Ge_DeepFashion2_A_Versatile_Benchmark_for_Detection_Pose_Estimation_Segmentation_and_CVPR_2019_paper.pdf

DNNの動向・CVのトレンド（42/42）

動画認識の応用

- 動画共有サイト，見守り，料理行動解析，インタラクションなど
- 動画DBは群雄割拠（下図）

Kinetics

W. Kay et al. “The Kinetics Human Action Video Dataset,” in arXiv:1705.06950 2017.
https://deepmind.com/research/open-source/open-source-datasets/kinetics/

EPIC KITCHENS

D. Damen et al. “Scaling Egocentric Vision: The EPIC-KITCHENS Dataset,” in ECCV 2018.
https://epic-kitchens.github.io/2018

HACS Dataset

H. Zhao et al. “HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization,” in arXiv pre-print 1712.09374 2017.
http://hacs.csail.mit.edu/

AVA

C. Gu et al. “AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions,” in CVPR 2018.
https://research.google.com/ava/download.html

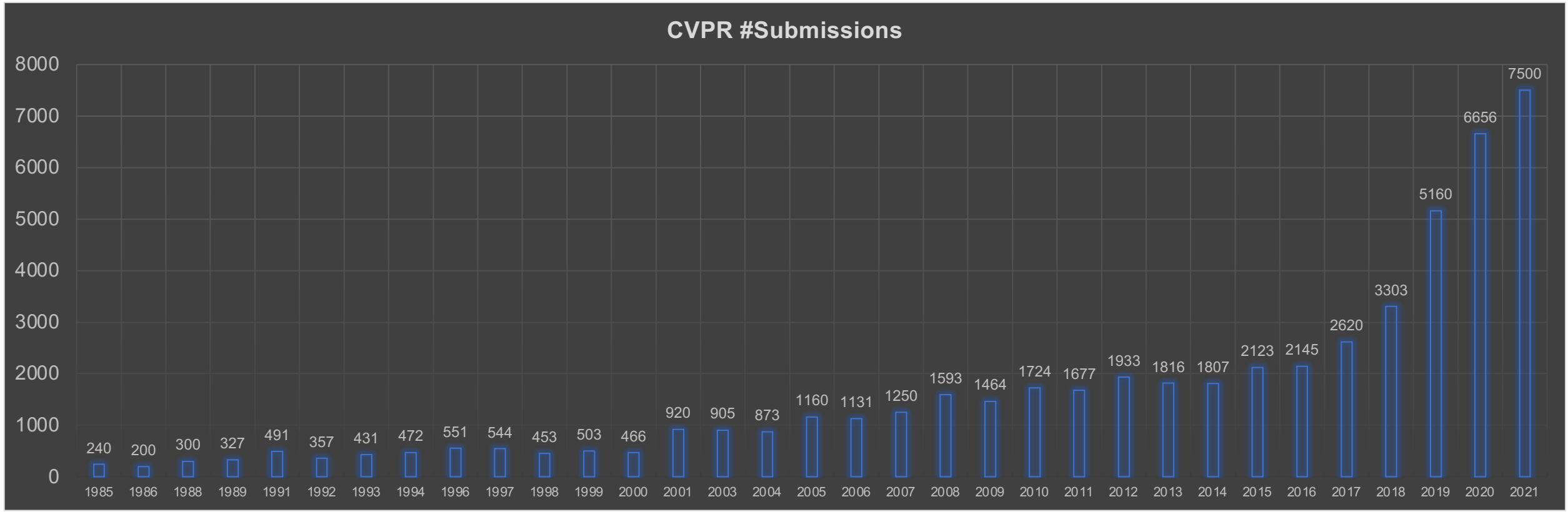
M. Monfort et al. “Moments in Time Dataset: one million videos for event understanding,” in arXiv pre-print 1801.03150, 2018.
http://moments.csail.mit.edu/

Something-Something v2 dataset
https://20bn.com/datasets/something-something

AI時代における論文数/CV人口の爆発！

投稿数，参加者数等，爆発的増加傾向

- CVPRの投稿数は最近も過去最高を記録
- CVPR'19の参加者数 約9,000⁺人と増加傾向（オンラインではやや減少気味）



主要国際会議の採択論文数

1,660 papers!

https://cvpr2021.thecvf.com/

1,612 papers!

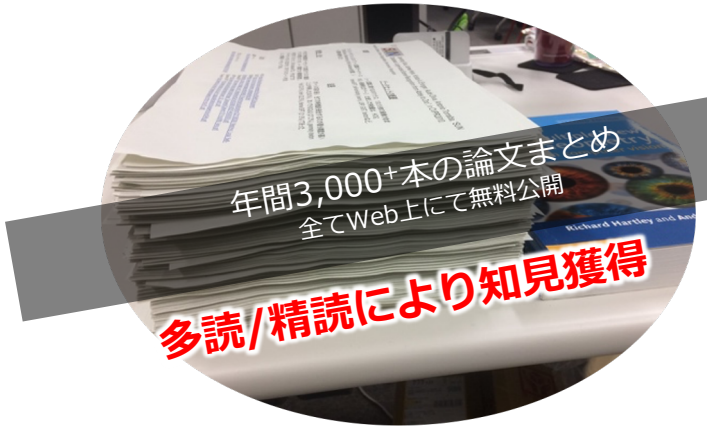
https://iccv2021.thecvf.com/

- 年間，主要２会議だけで3,000論文を超える
- このような状況でいかに情報を捉え研究するか？

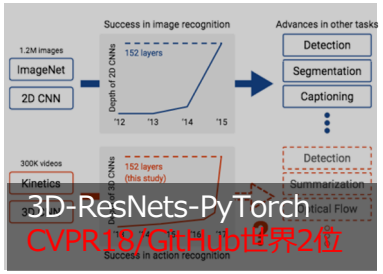
CV分野の研究コミュニティ cvpaper.challenge

今を映す（網羅的サーベイ），トレンドを創る（連携して研究）

◆論文読破・まとめ・発想・議論・実装・論文執筆に至るまで取り組む



cvpaper.challengeの研究プロジェクト例



その他多数のProj.が進行中

Survey Member: 1,000⁺名

Research Member: 65⁺名

（産総研/筑波大/電大/早大/慶大/東工大/東大/広島大/奈良先端大/大阪大/九州大/中部大/横国大/芝浦工大/会津大/YSFH/AI-SCHOLAR/ユースコミュニケーションズ）

HP: http://xpaperchallenge.org/

Twitter: @CVpaperChalleng

日本のCV分野を強くするチャレンジ