Convergence of Architectures and Learning Methods in Computer Vision with NLP and Other AI fields

Han Hu

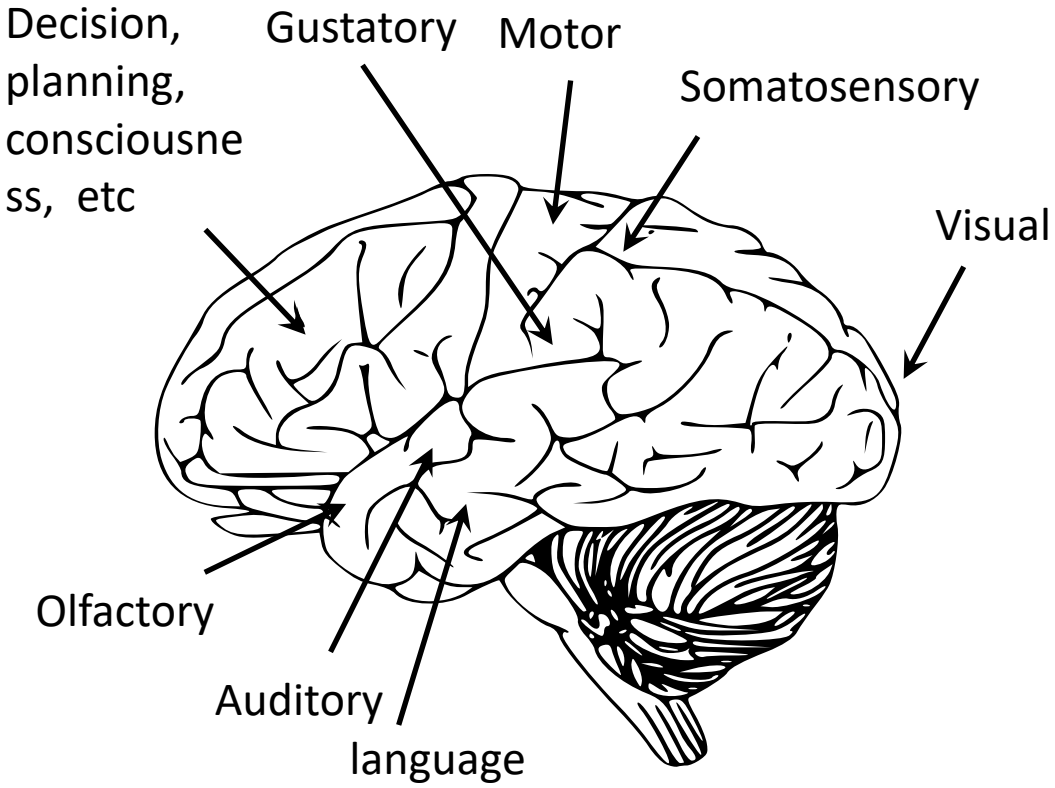
Microsoft Research Asia

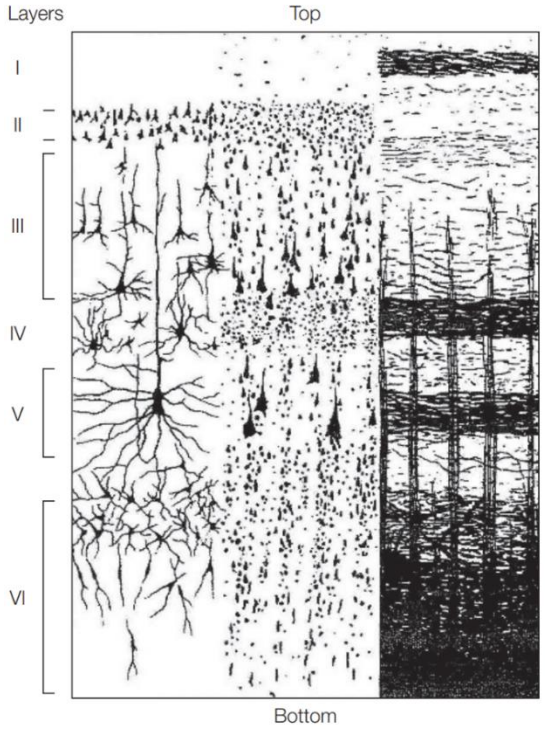
CVPR2022 OpenMMLab Tutorial

June 20th , 2022

Human brain: A universal architecture and universal learning machine

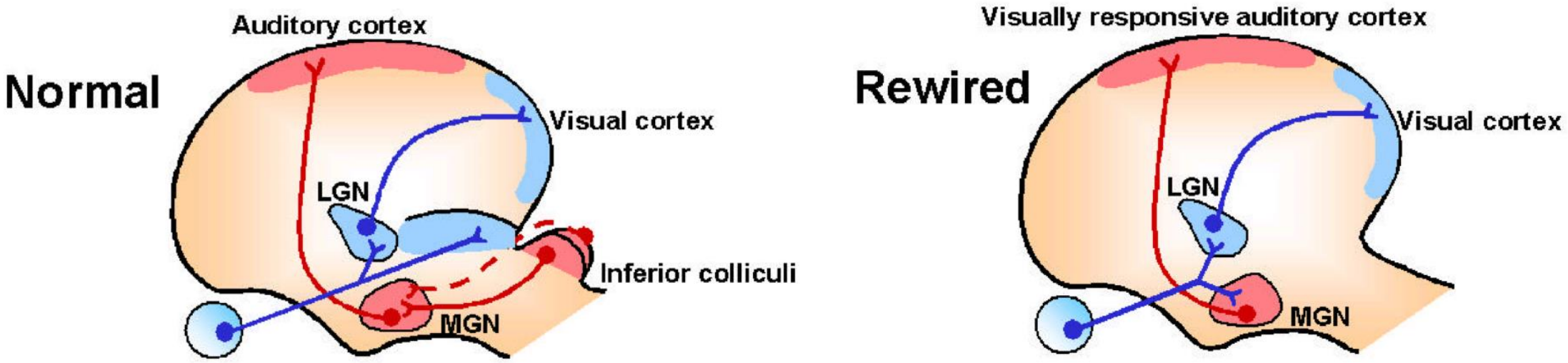
- Universal architecture (innate)： neocortex





Human brain: A universal architecture and universal learning machine

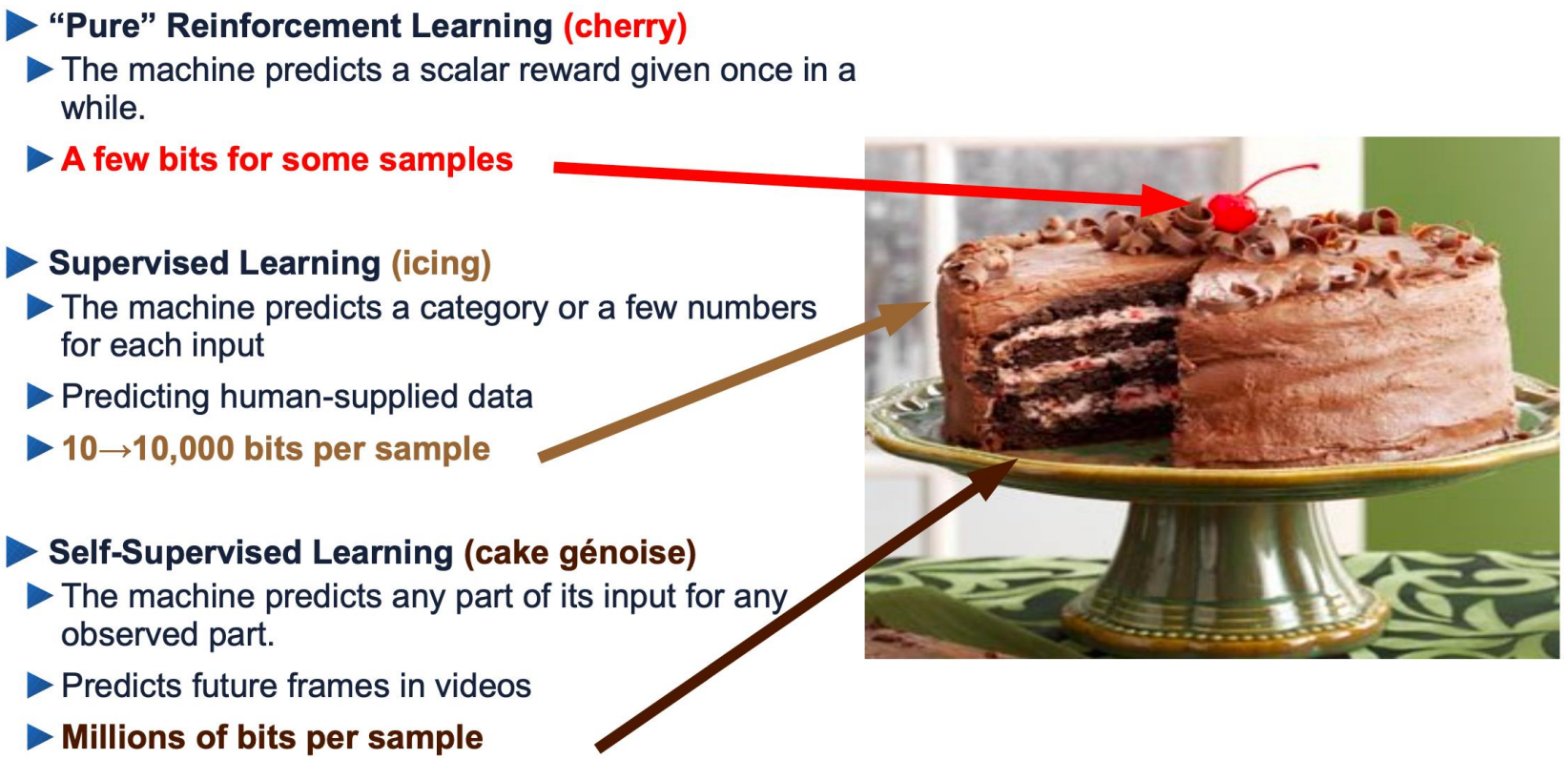
- Plasticity of neocortex: a mouse brain experiment



J. Sharma et al. Induction of visual orientation modules in auditory cortex. Nature 2000

Human brain: A universal architecture and universal learning machine

- Learning (nurture): Yann LeCun's cake



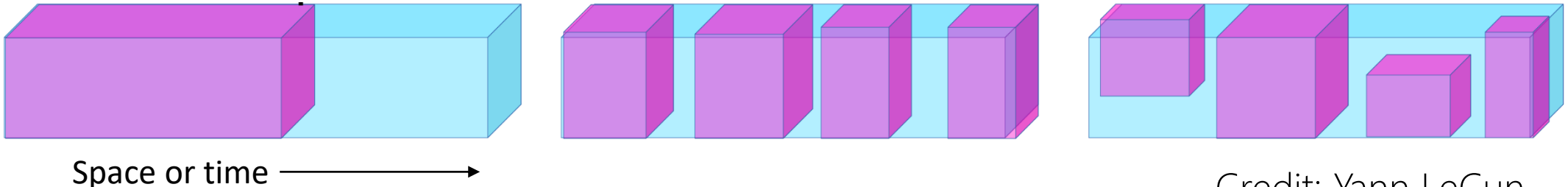
Credit: Yann LeCun

Human brain: A universal architecture and universal learning machine

- Learning (nurture) : baby's learning through observation (self-supervised)

Human brain: A universal architecture and universal learning machine

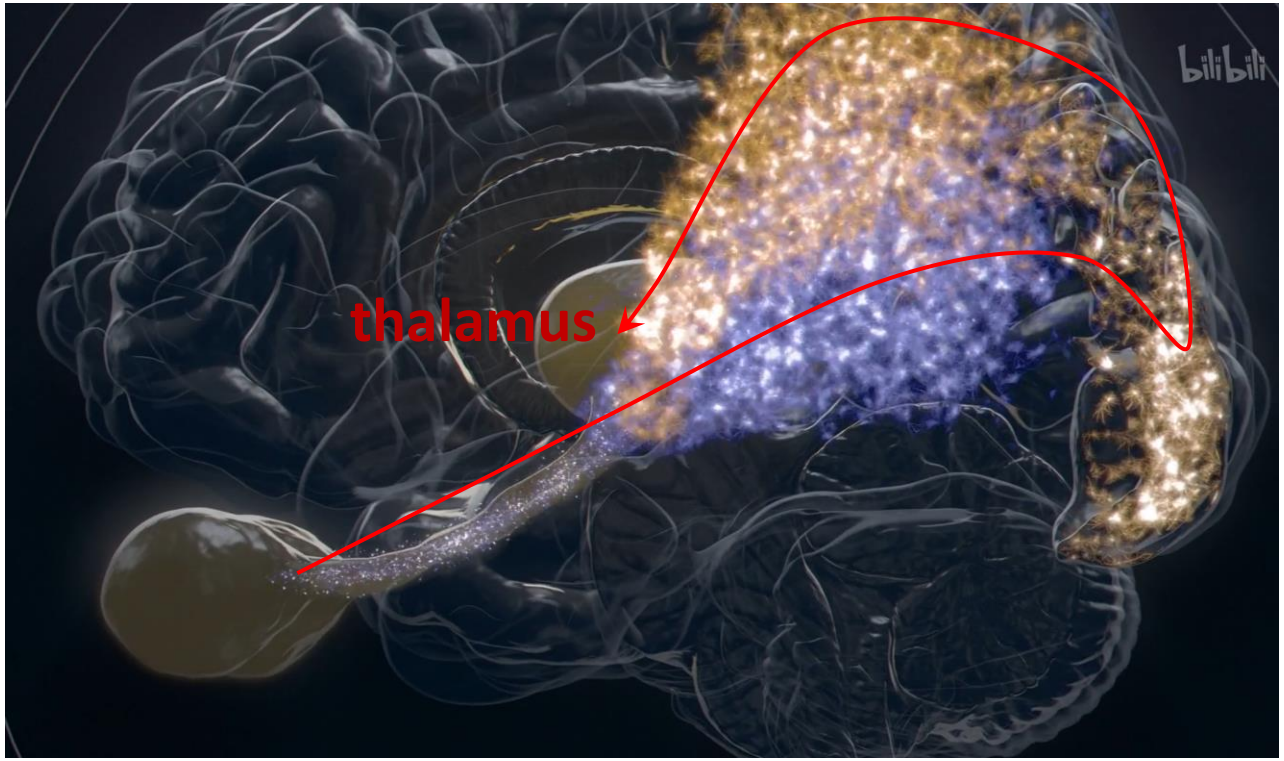
- Unified learning approach: Compare difference between prediction and ground truth input
  - Vision, touch, sound, language, etc



Credit: Yann LeCun

Human brain: A universal architecture and universal learning machine

- Unified learning approach: Compare difference between prediction and ground truth input
  - Thalamus plays a key role

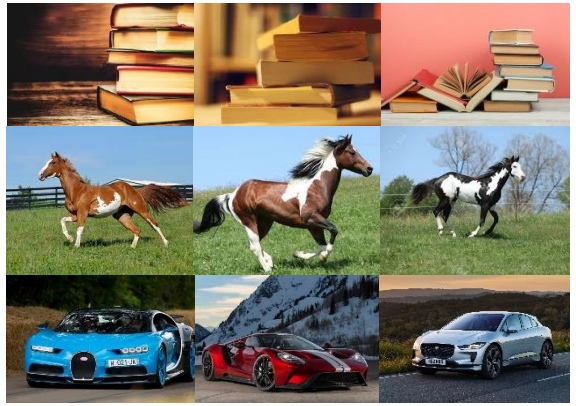


Credit: David Eagleman

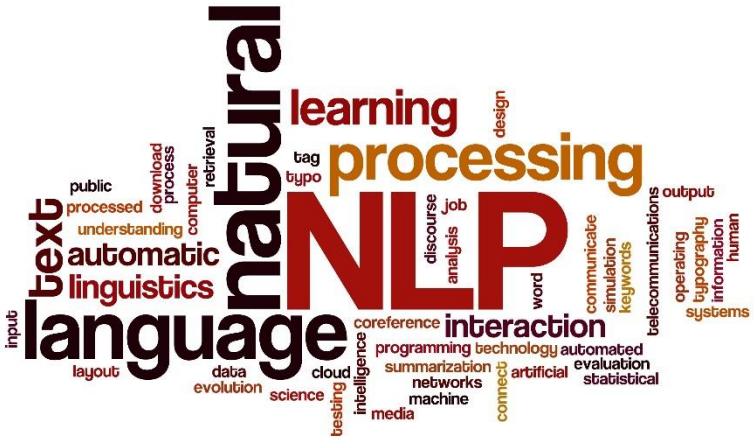
Towards Converged AI Architectures

(Swin) Transformer

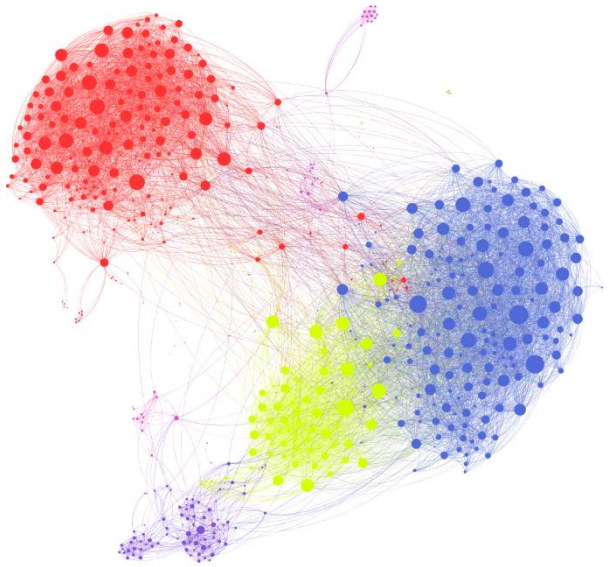
Mainstream models of AI sub-fields



vision-CNN

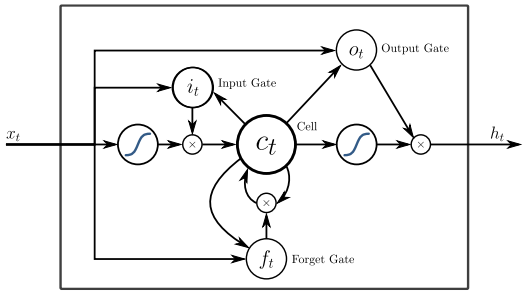


language-Transformer



social-Graph Networks

Model evolution for NLP or sequential data



1995

LSTM

Jürgen Schmidhuber

2014

GRU

Google

2017

Transformer

Google

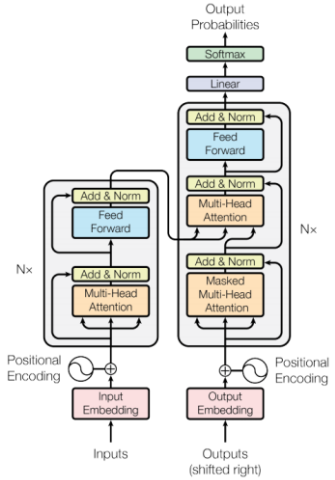


Figure 1: The Transformer - model architecture.

2013

Deep LSTM

Geoffrey Hinton
& Microsoft

2015

RNN+attention

CIFAR

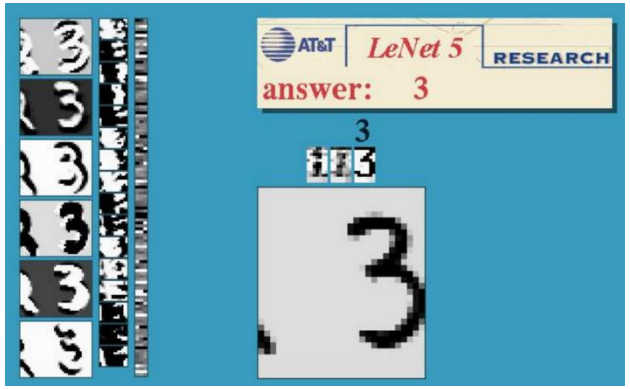
Ashish Vaswani et al. Attention is all you need. NeurIPS 2017

Model evolution for CV

1980

CNN

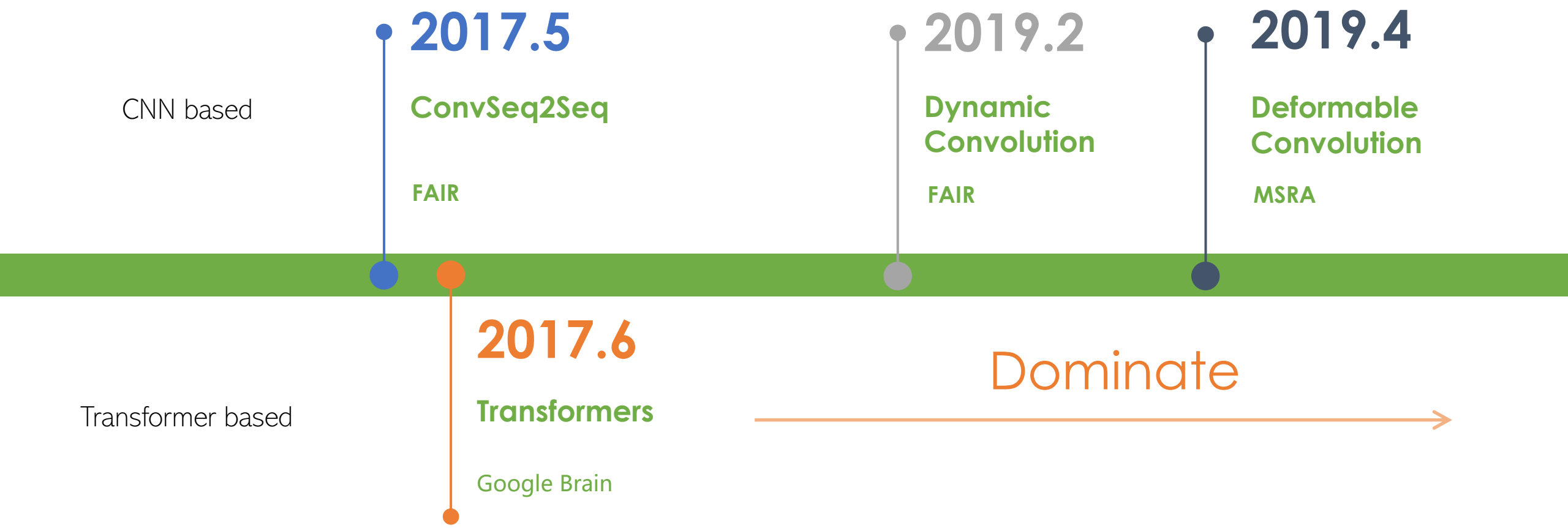
Kunihiko Fukushima
Yann LeCun



AlexNet, GoogleNet, VGGNet, ResNet, DenseNet, …

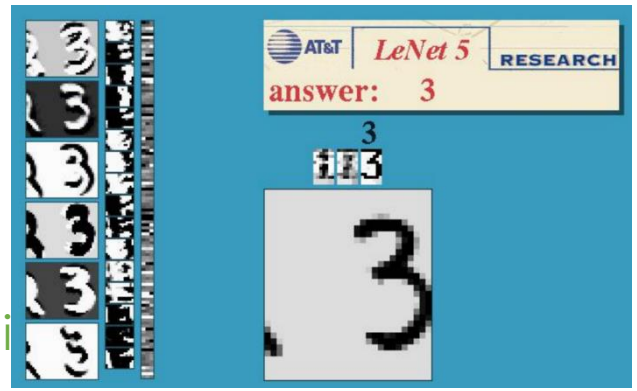
Can NLP and CV use the same architectures?

- Adapt CNN or convolution for NLP



Can NLP and CV use the same architectures?

- Adapt Transformers or attention for CV



1980

CNN

Kunihiko Fukushi
Yann LeCun

AlexNet, GoogleNet, VGGNet, ResNet, DenseNet, …

2020.10

ViT (Google)

2021.03

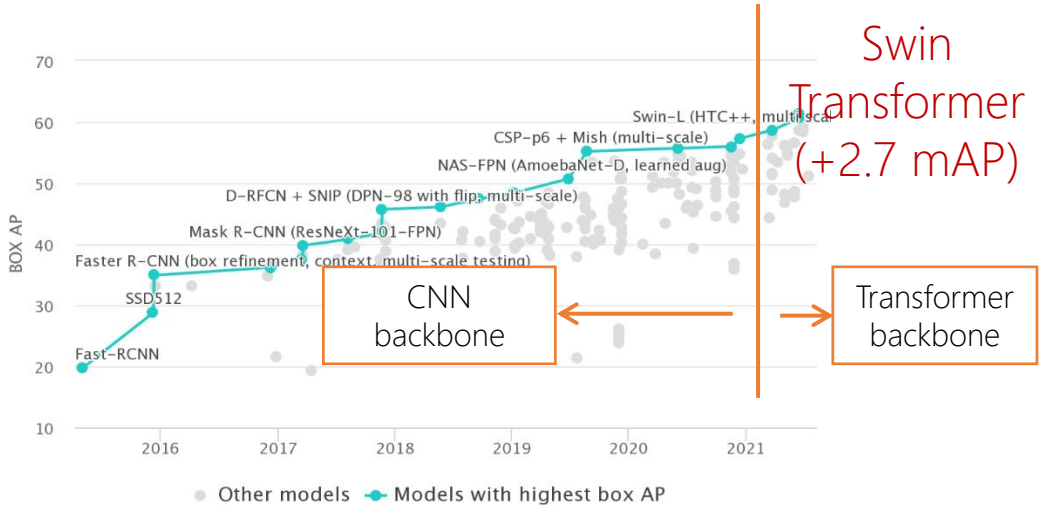
Swin Transformer (MSRA)

CNN ← → Vision Transformers

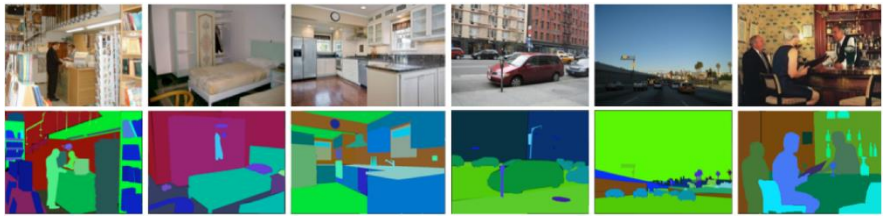
Swin Transformer: A general-purpose vision backbone

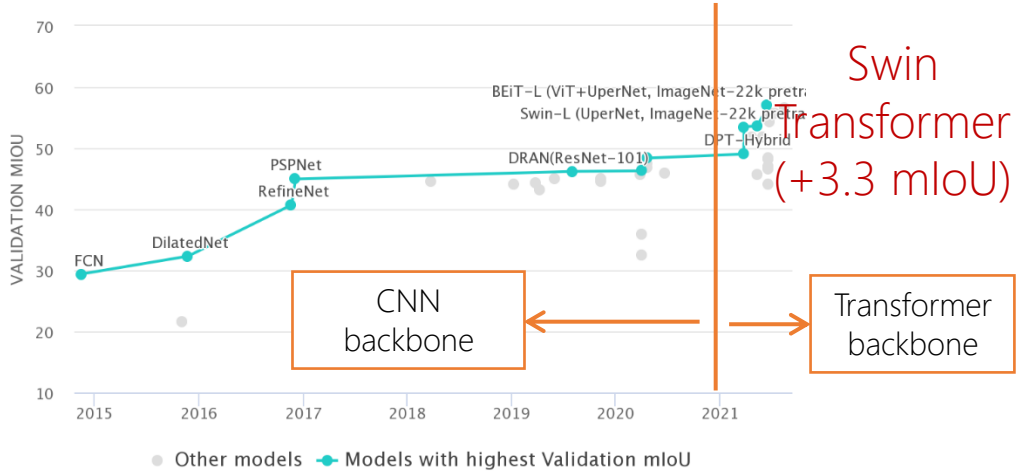
- The first time to significantly surpass previous best CNN on broad vision tasks

COCO object detection



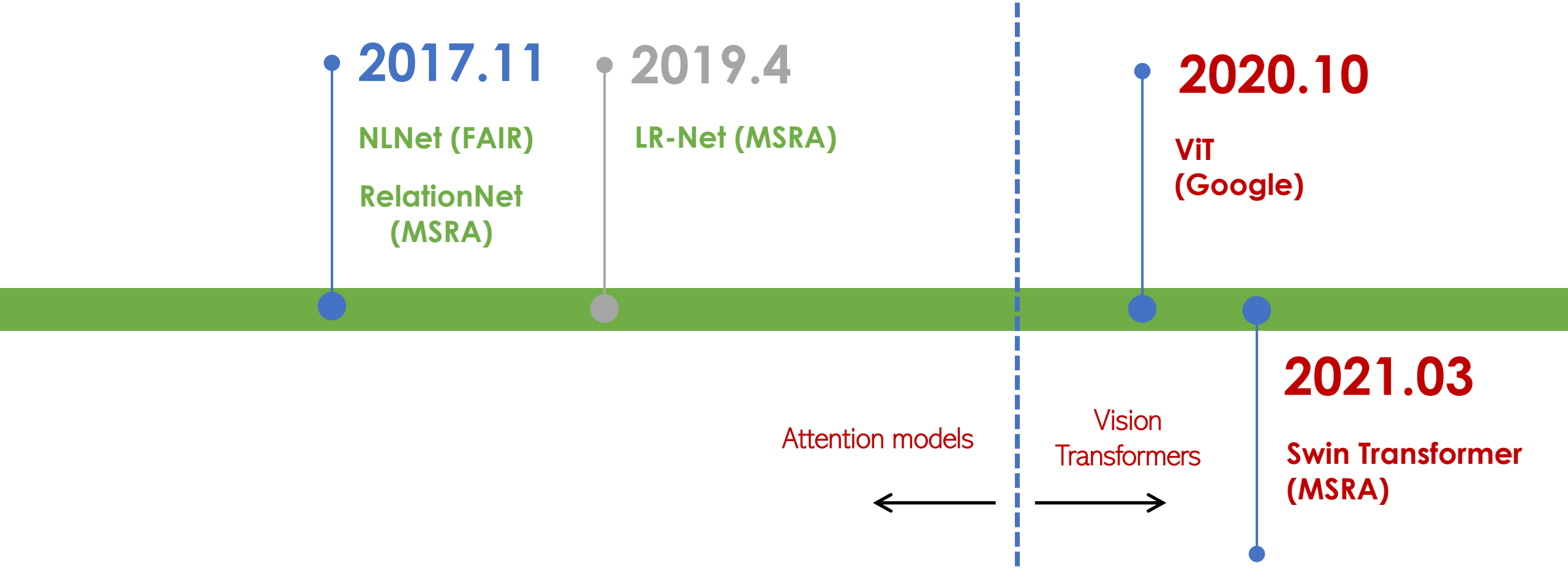
ADE20K semantic segmentation





Can NLP and CV use same architectures?

- Vision Transformers and attention models



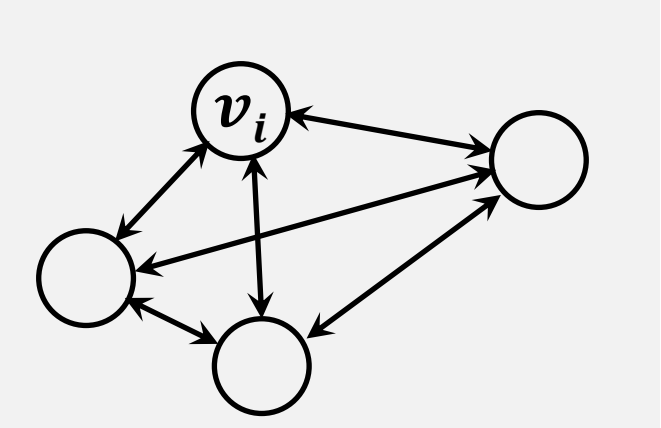
It is a long-term belief

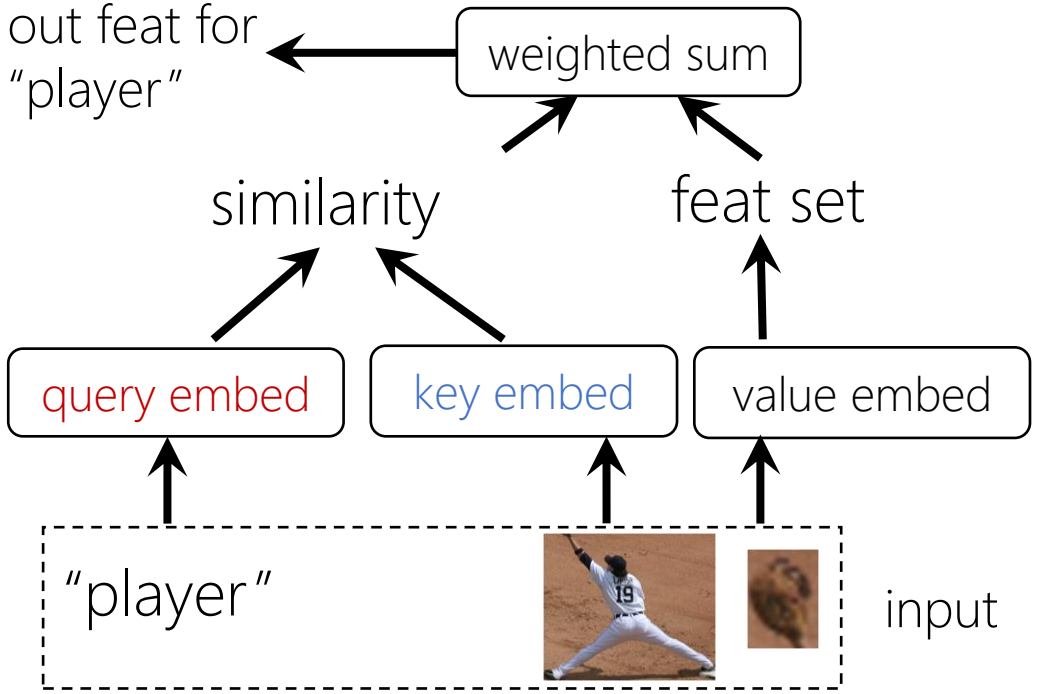
Towards Universal Learning Machine: Self-Attention for Visual Modeling

2019.7 @valse webinar

Why is it Transformer or attention?

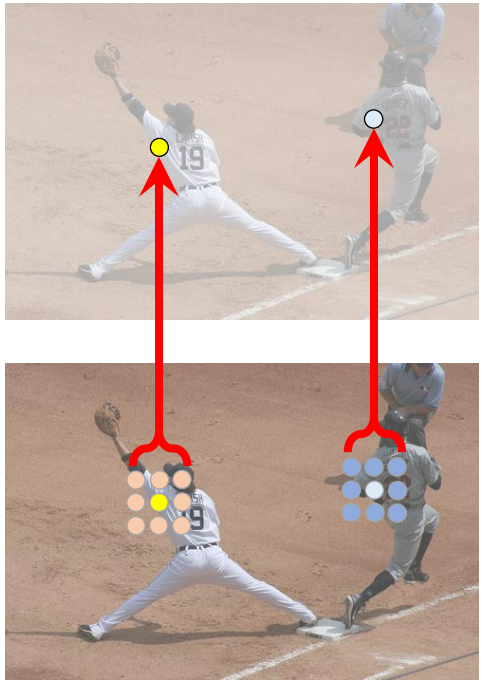
- General modeling capability
  - All concepts (concrete or abstract) and their relationships can be modeled by a graph
  - Modeling arbitrary relationship via verification, which is hard for CNN



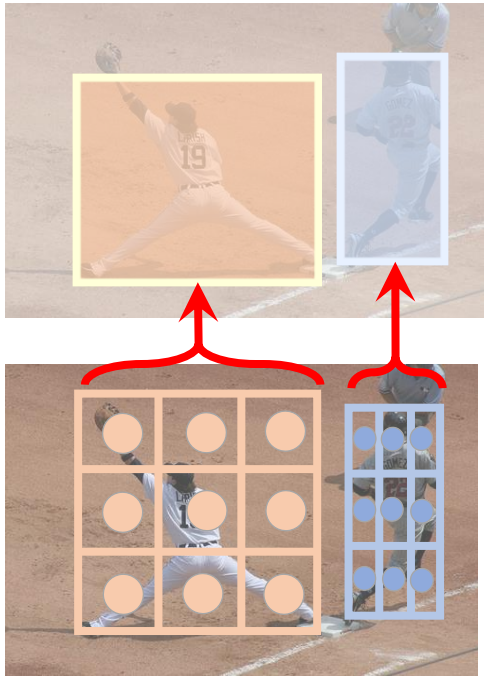


Transformer is so general

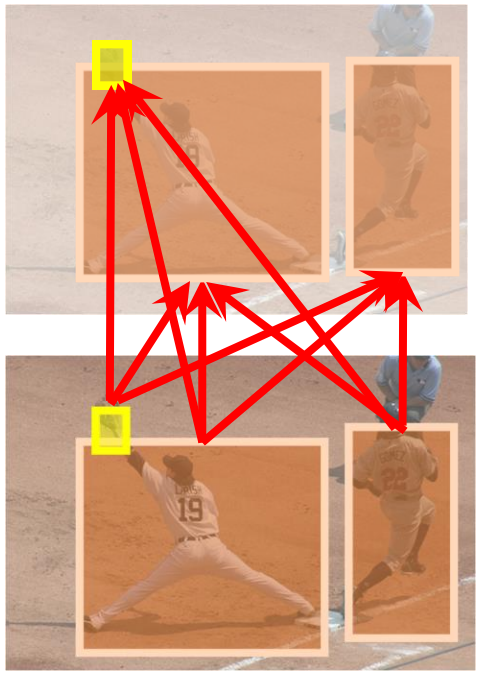
- Can model all of pixel-to-pixel, object-to-pixel, object-to-object relationships, which covers almost all vision applications



pixel-to-pixel



object-to-pixel



object-to-object

Our works (2017-2021)

- Pixel-to-pixel

  [1] GCNet: Non-local networks meet squeeze-excitation networks and beyond, ICCVW 2019
  [2] Local Relation Networks for Image Recognition. ICCV 2019 (**the first full-attention visual backbone**)
  [3] Disentangled Non-local Neural Networks. ECCV 2020
  [4] Swin Transformer. ICCV 2021 (**best paper award**)
  [5] Swin Transformer V2. CVPR 2022

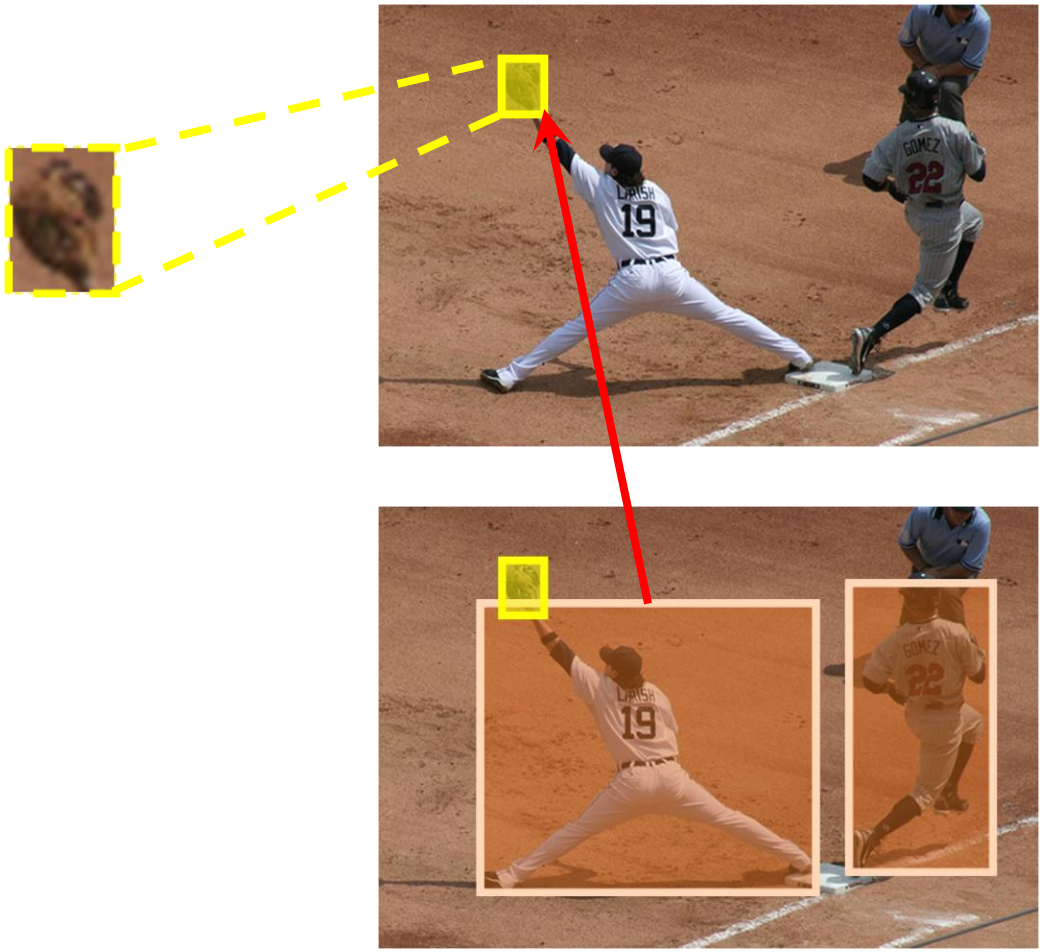
- Object-to-pixel

  [6] Learning Region Features for Object Detection. ECCV 2018
  [7] RelationNet++: Bridging Visual Representations for Object Detection via Transformer Decoder, NeurIPS 2020

- Object-to-object

  [8] Relation Networks for Object Detection, CVPR2018 (**the first end-to-end object detector**)
  [9] Spatial-temporal relation networks for multi-object tracking. ICCV 2019
  [10] Memory Enhanced Global-Local Aggregation for Video Object Detection. CVPR 2020

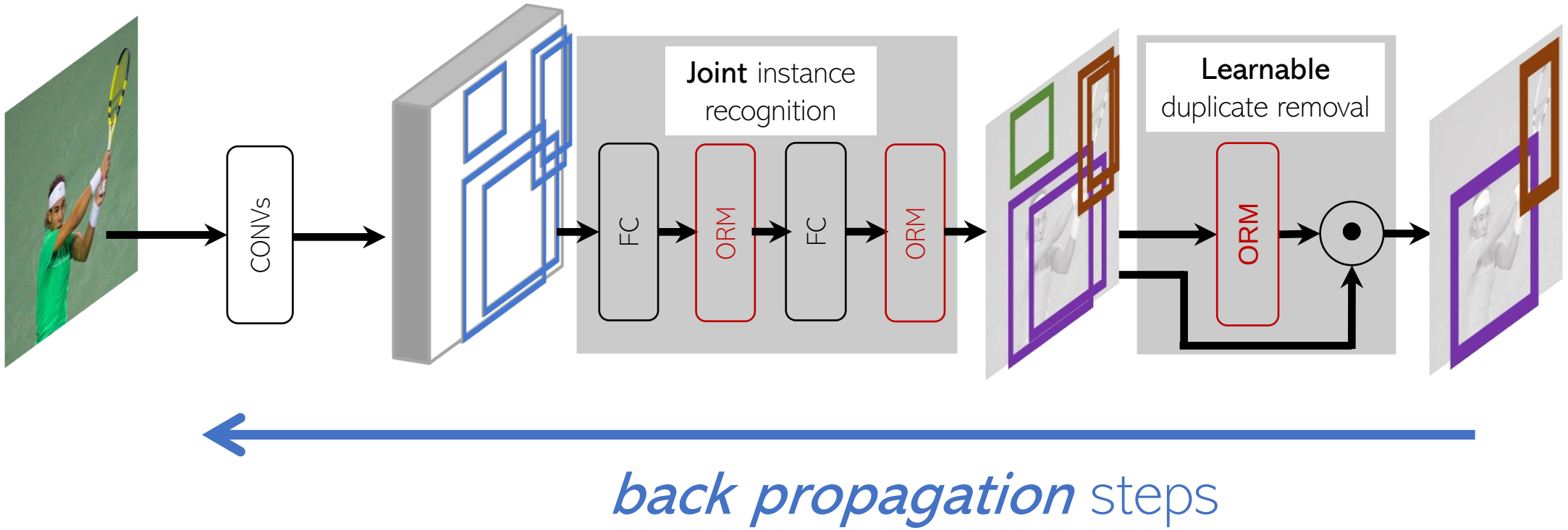
Relation Networks for Object Detection (CVPR'2018)



It is much easier to detect the ***glove*** if we know there is a ***baseball player***.

Relation Networks for Object Detection (CVPR'2018)

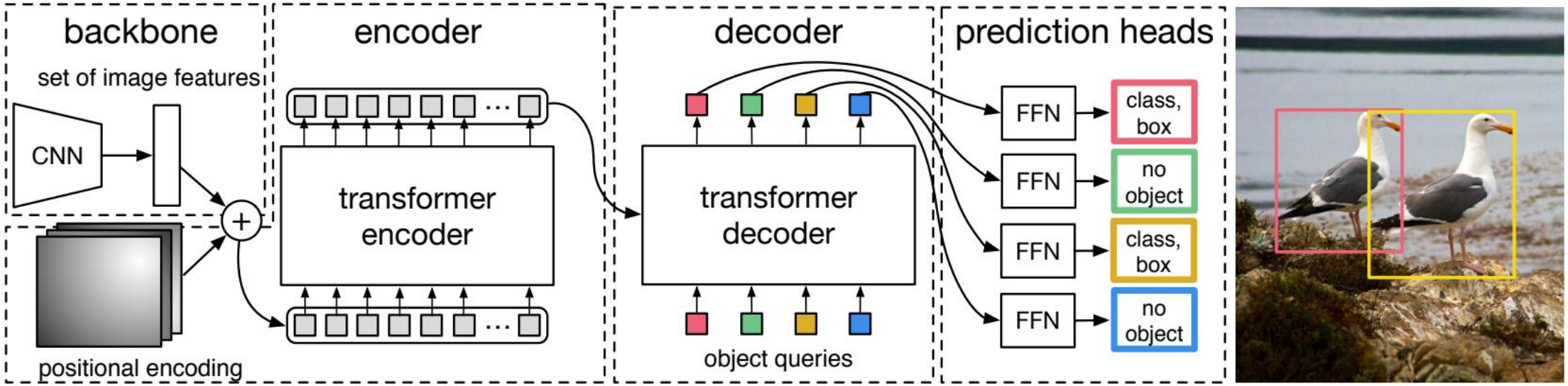
- The first fully end-to-end object detector



Han Hu et al. *Relation Networks for Object Detection*. CVPR 2018

DETR (ECCV'2020)

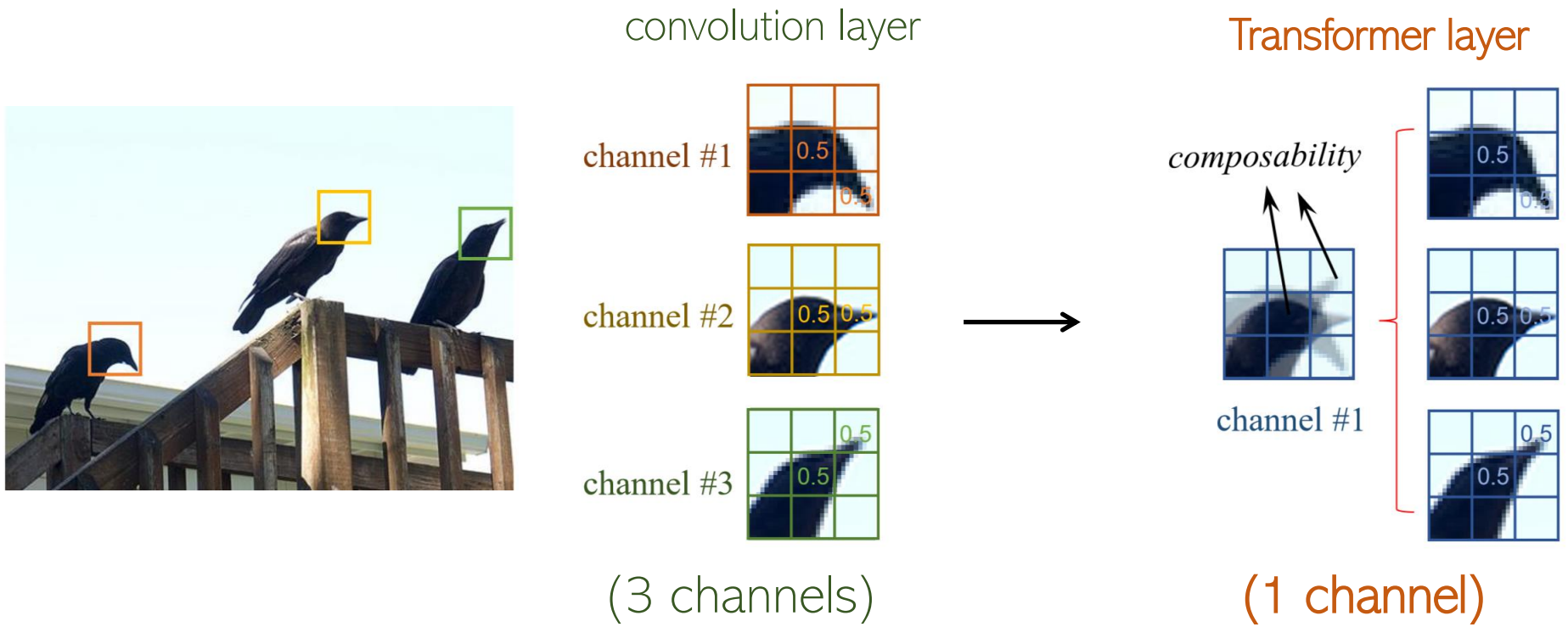
- Another end-to-end object detector



Nicolas Carion et al. End-to-End Object Detection with Transformers. ECCV 2020

Local relation networks (ICCV'2019)

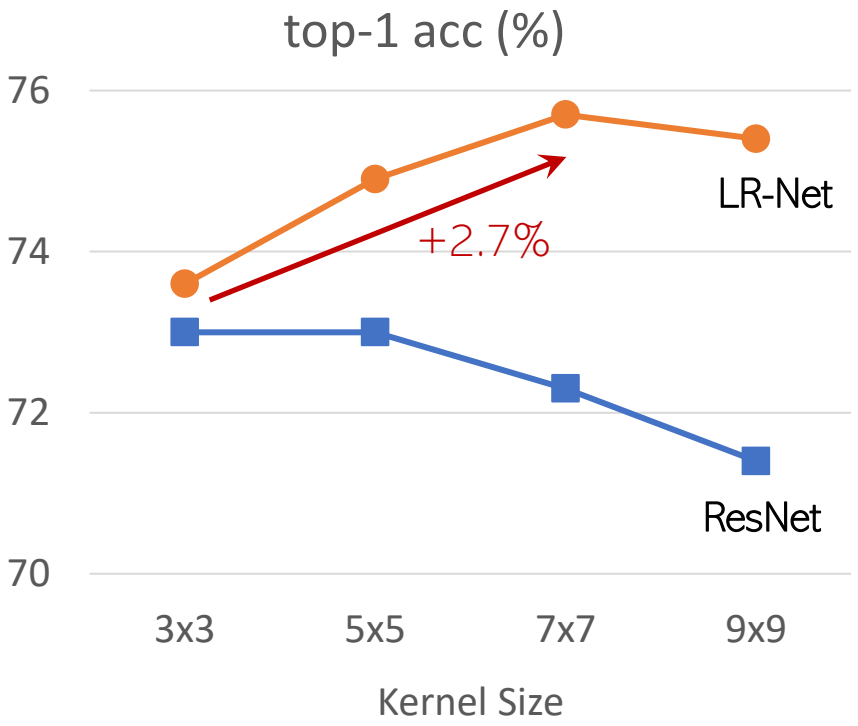
- Full-attention visual backbone
  - "Convolution is exponentially inefficient!"



Local relation networks (ICCV'2019)

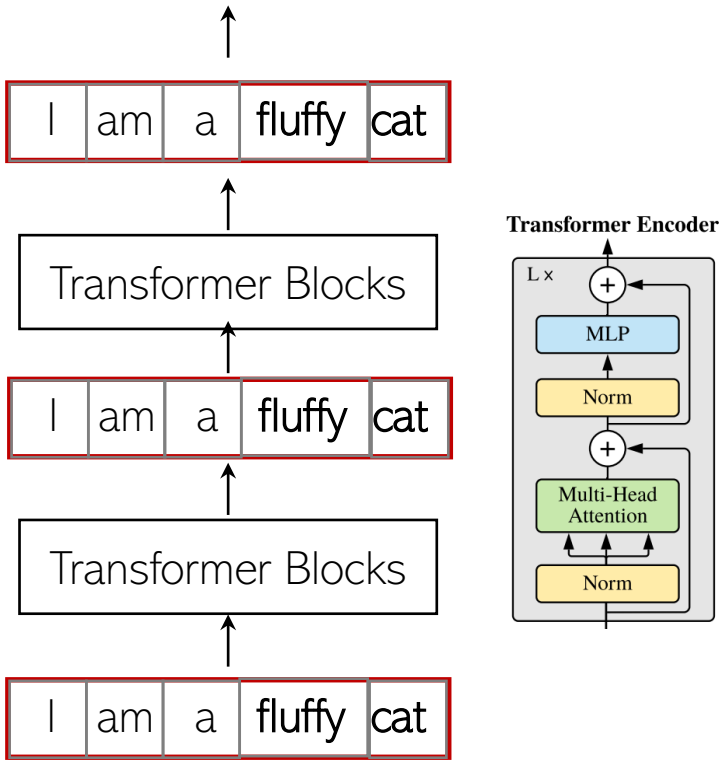
- The first full-attention visual backbone

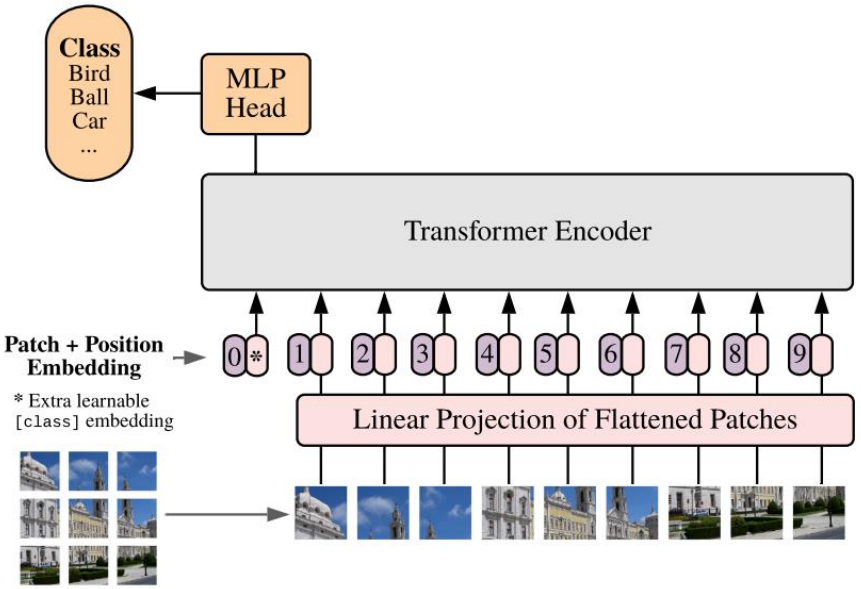
stage	output	ResNet-50	LR-Net-50 (7×7, m=8)
res1	112×112	7×7 conv, 64, stride 2	**1×1, 64** **7×7 LR, 64, stride 2**
res2	56×56	3×3 max pool, stride 2	3×3 max pool, stride 2
		[1×1, 64 3×3 conv, 64 1×1, 256] ×3	[1×1, 100 **7×7 LR, 100** 1×1, 256] ×3
res3	28×28	[1×1, 128 3×3 conv, 128 1×1, 512] ×4	[1×1, 200 **7×7 LR, 200** 1×1, 512] ×4
res4	14×14	[1×1, 256 3×3 conv, 256 1×1, 1024] ×6	[1×1, 400 **7×7 LR, 400** 1×1, 1024] ×6
res5	7×7	[1×1, 512 3×3 conv, 512 1×1, 2048] ×3	[1×1, 800 **7×7 LR, 800** 1×1, 2048] ×3
	1×1	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax
# params		**25.5**×10^6	**23.3**×10^6
FLOPs		**4.3**×10^9	**4.3**×10^9



Han Hu et al. *Local Relation Networks for Visual Recognition. ICCV 2019*

Vanilla Vision Transformer (ViT)



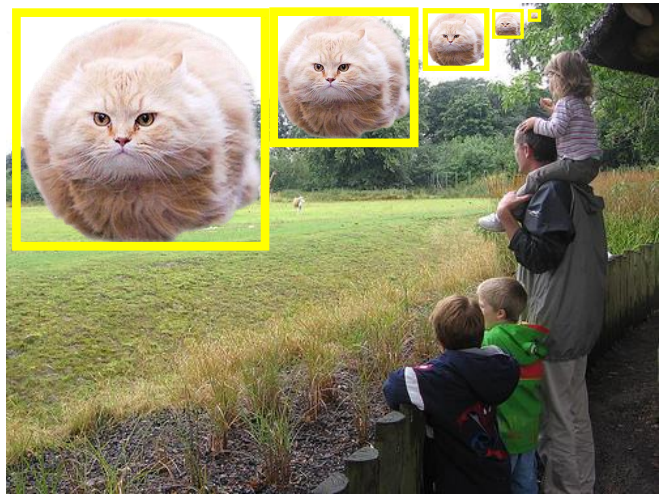


Ashish Vaswani et al, Attention is all you need, NeurIPS 2017

Alexey Dosovitskiy et al, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR2021

Key difference between visual and text signals

hierarchy
(scale invariance)

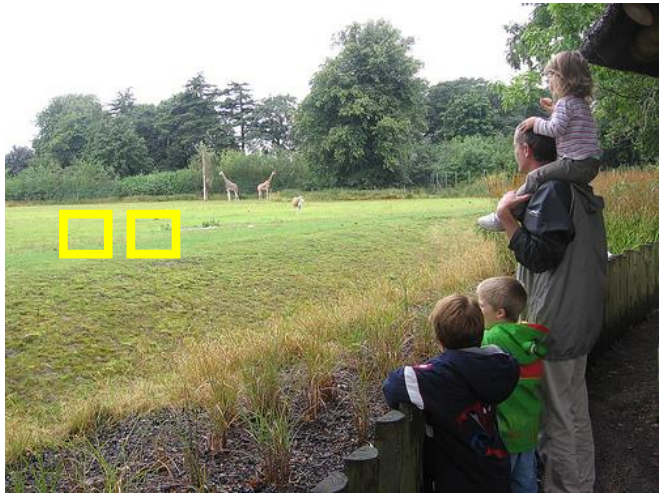


I am a **fluffy cat.**

I am a **fluffy fluffy cat cat.** (invalid)

No scale variation

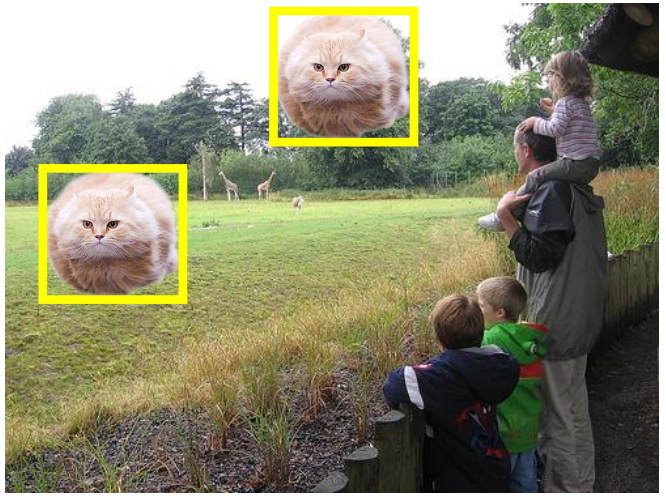
locality
(spatial smoothness)



I like the **green grass**.

No spatial smoothness

translation invariance



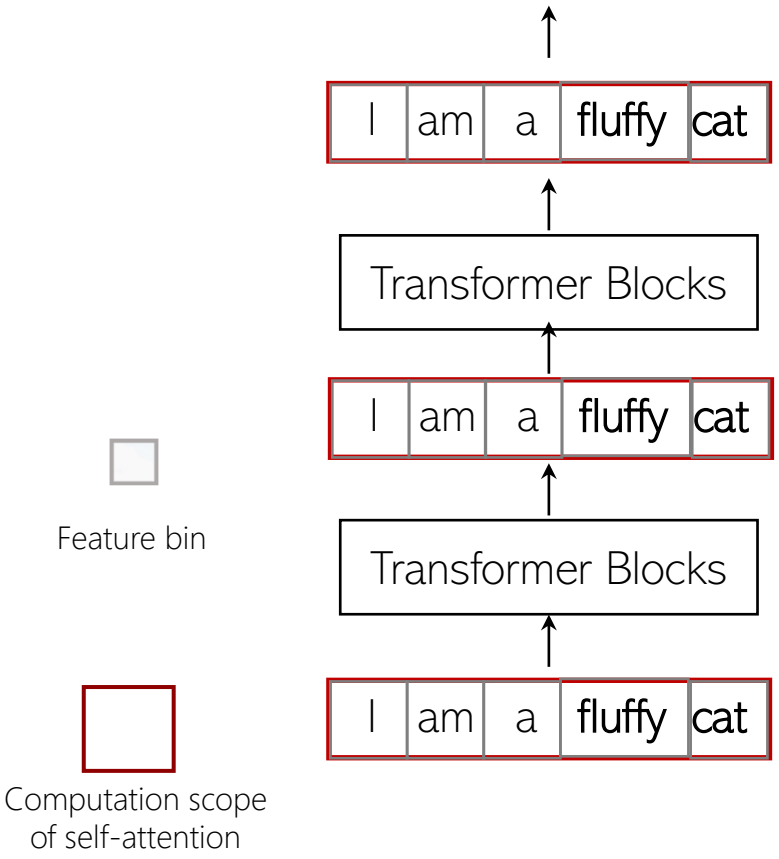
I am a **fluffy cat**.

Fluffy cat is me.

Sensitive to absolute locations

Computer vision priors: the bridge from NLP to CV

Standard Transformer (2017.6)

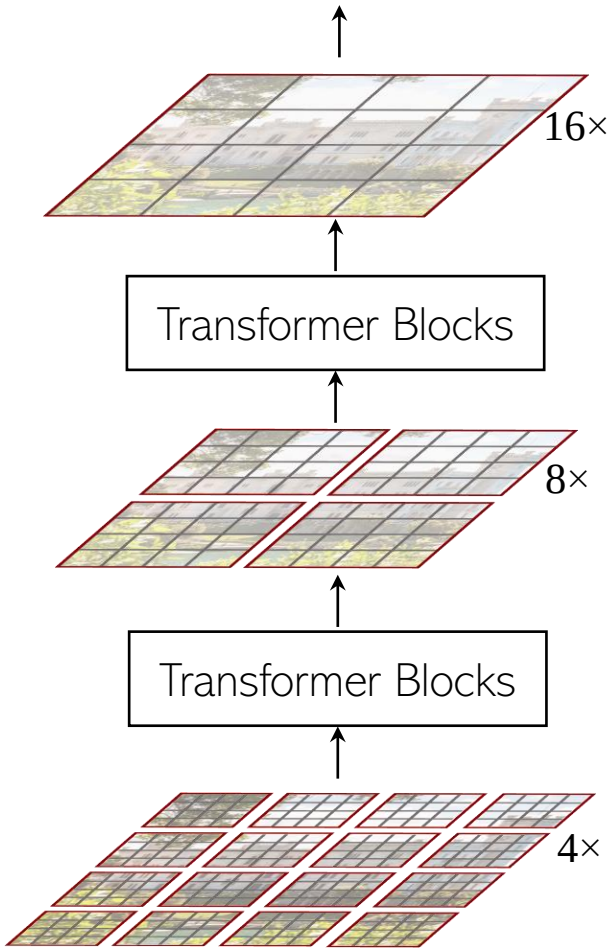


plain → hierarchy

global → locality

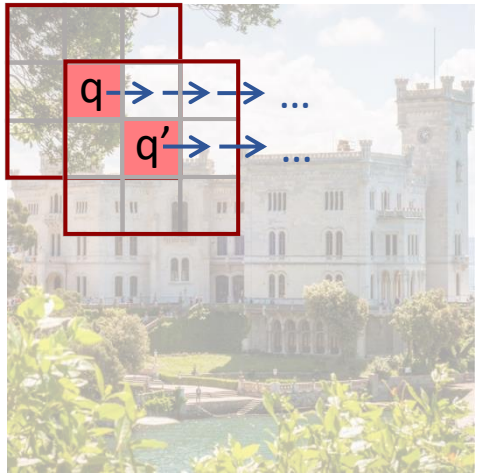
absolute position → translation invariance

Swin Transformer (2021.3)

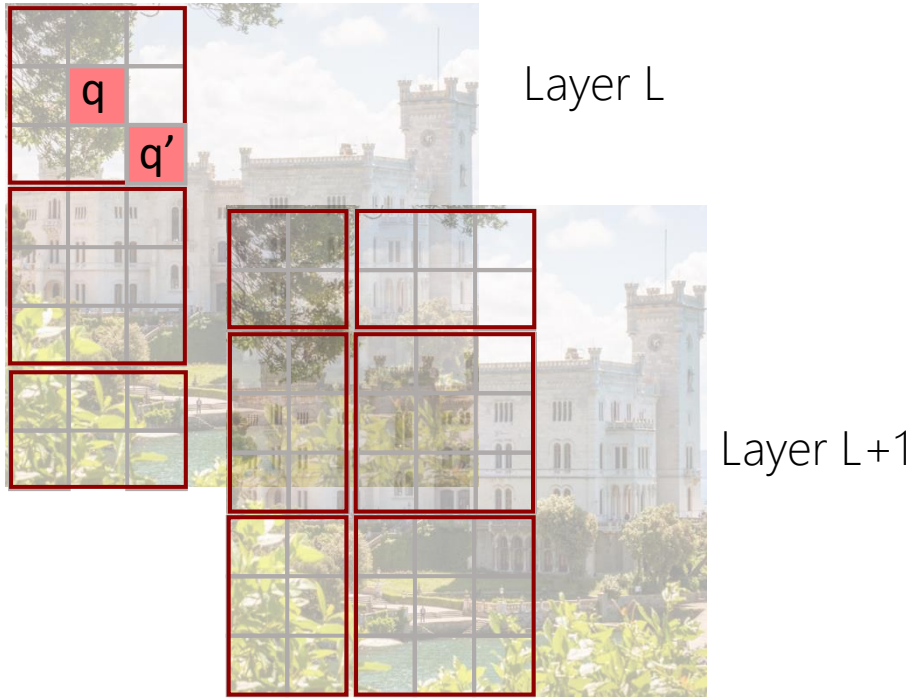


From sliding windows to shifted windows

- Non-overlapping windows (3x speed up in latency)
- Shifted window configurations in the next layer



Sliding window
(CNN/LR-Net)



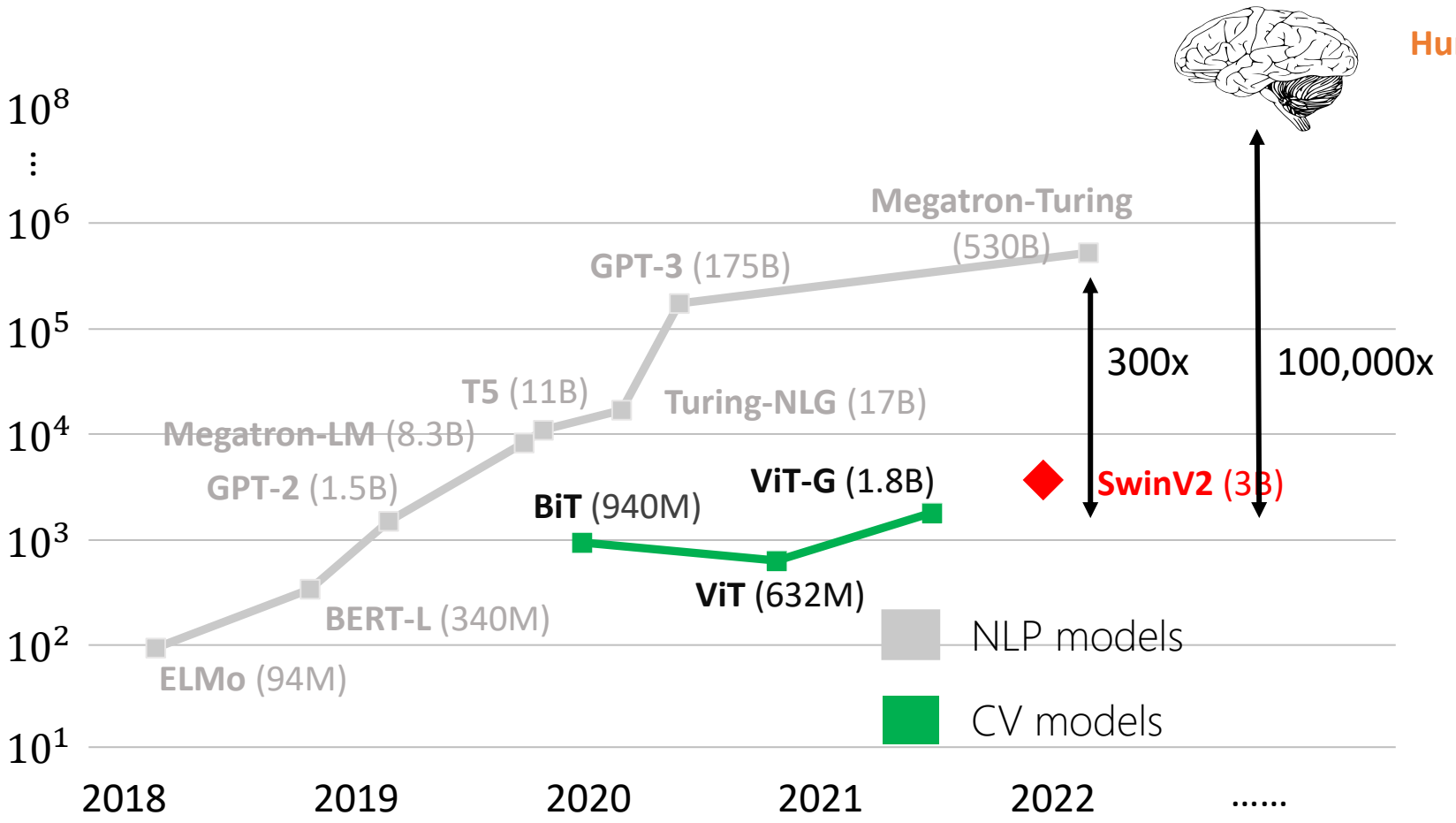
Shifted window (Swin, ICCV'2021)

Scaling up visual models

- Swin Transformer V2 (CVPR'2022)
  - The world's largest dense visual model (3B parameters)
  - Applicability of billion-scale visual models beyond image classification
  - Requiring 40 times less labelled data than Google's

Benchmark	ImageNetV2	COCO test-dev	ADE20K val	Kinetics-400
Swin V1	77.5	58.7	53.5	84.9
Previous SoTA	83.3 (Google, July 2021)	61.3 (Microsoft, July 2021)	58.4 (Microsoft, October 2021)	85.4 (Google, October 2021)
Swin V2 (Giant) (November 2021)	84.0 (+0.7)	63.1 (+1.8)	59.9 (+1.5)	86.8 (+1.4)

Computer vision models are still relatively small

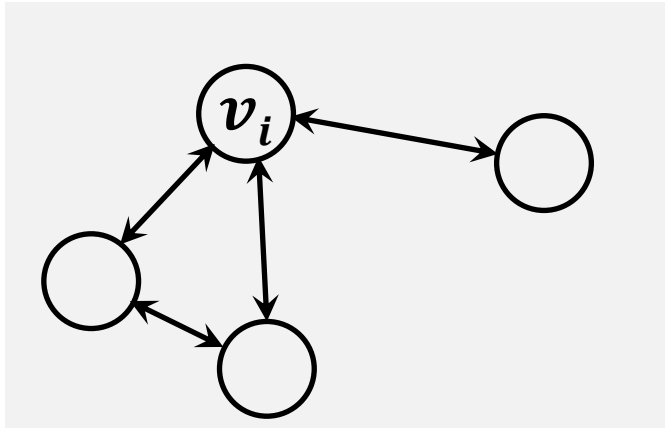


A general principle to apply Transformer

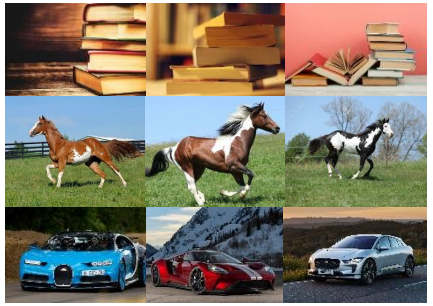
Generic Transformer Unit

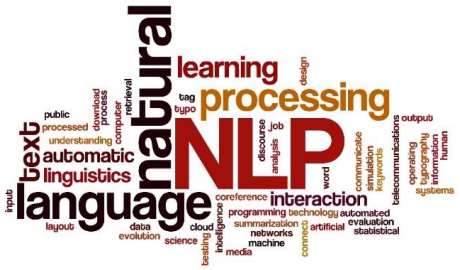
+

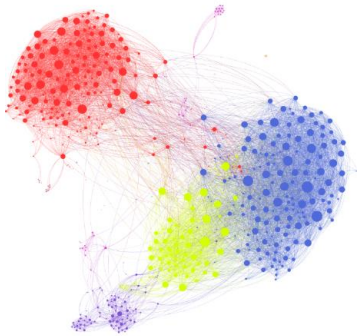
Domain-Specific Priors
(Defined by graphs)



→







OGB-LSC @ KDD Cup 2021

Large-Scale Challenge

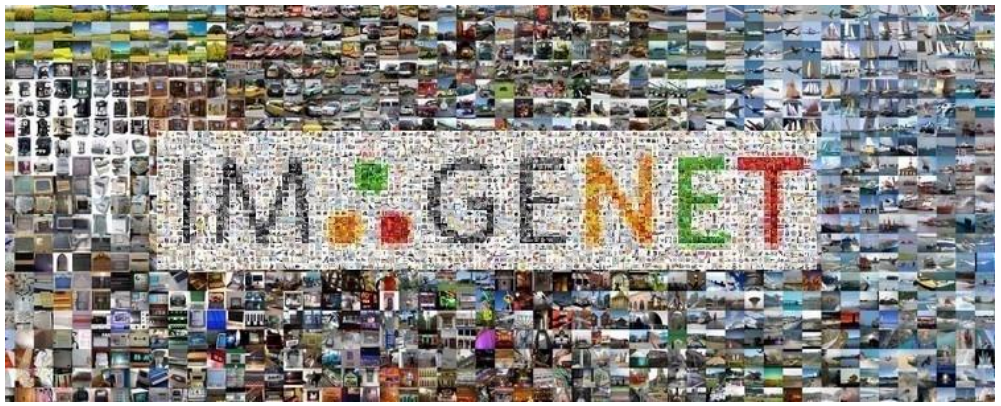
Dataset: Predicting a quantum property of molecular graphs

Graphormer, NeurIPS 2021

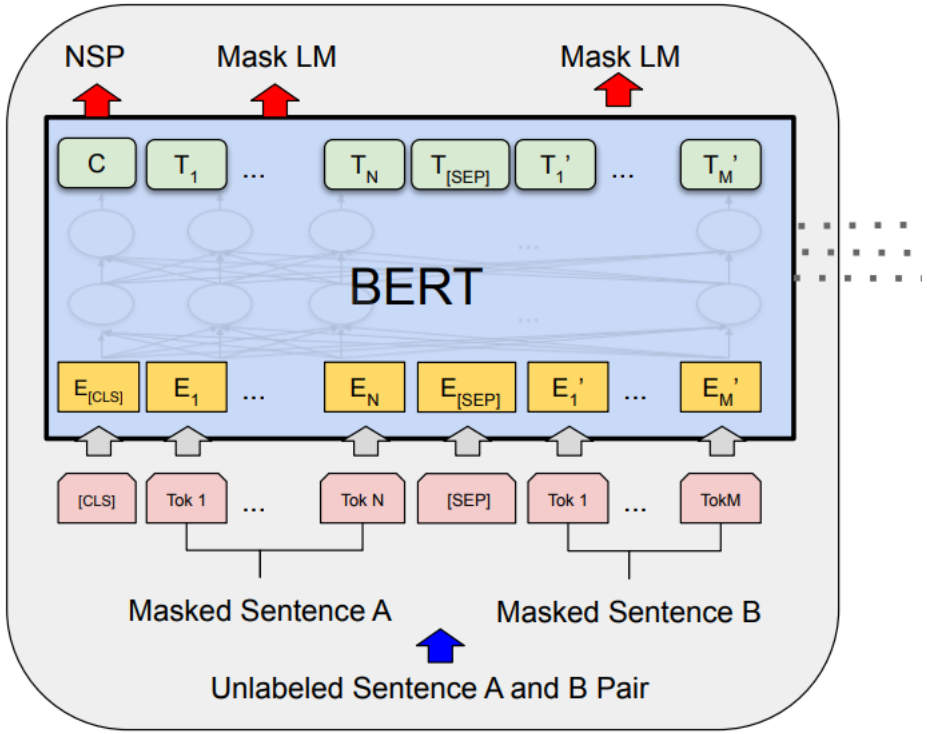
Towards converged learning methods

Masked signal modeling (BERT-like)

Mainstream pre-training methods for AI subfields

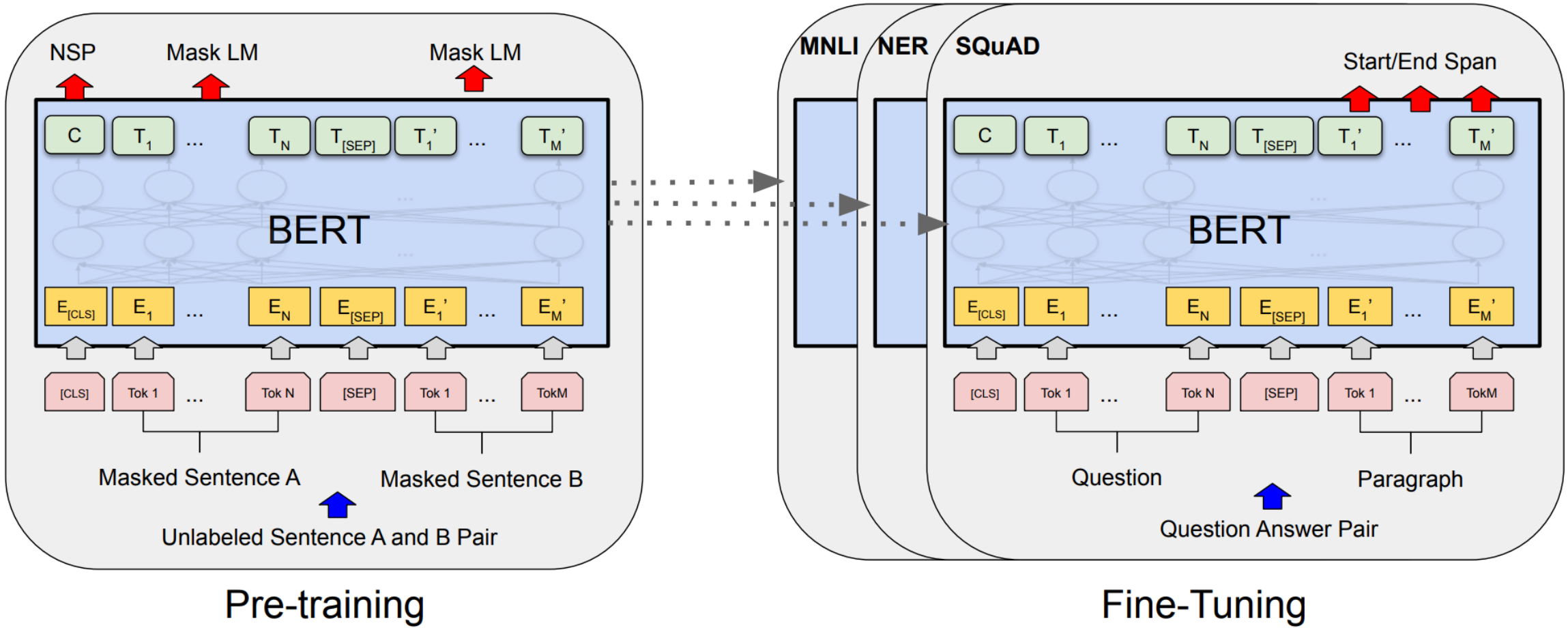


Supervised learning
(Computer vision)



Self-supervised learning
(Nature language processing)

BERT in NLP: Predicting unknown input



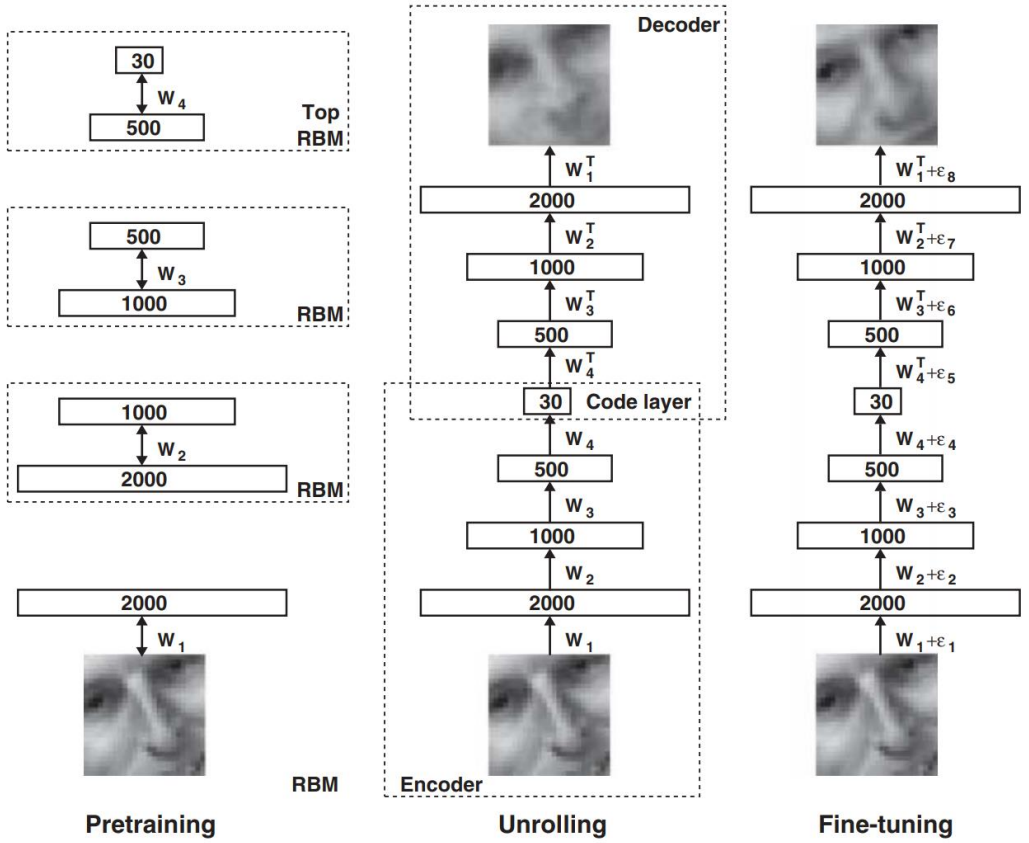
Self-supervised learning in computer vision

- Pioneer of deep learning

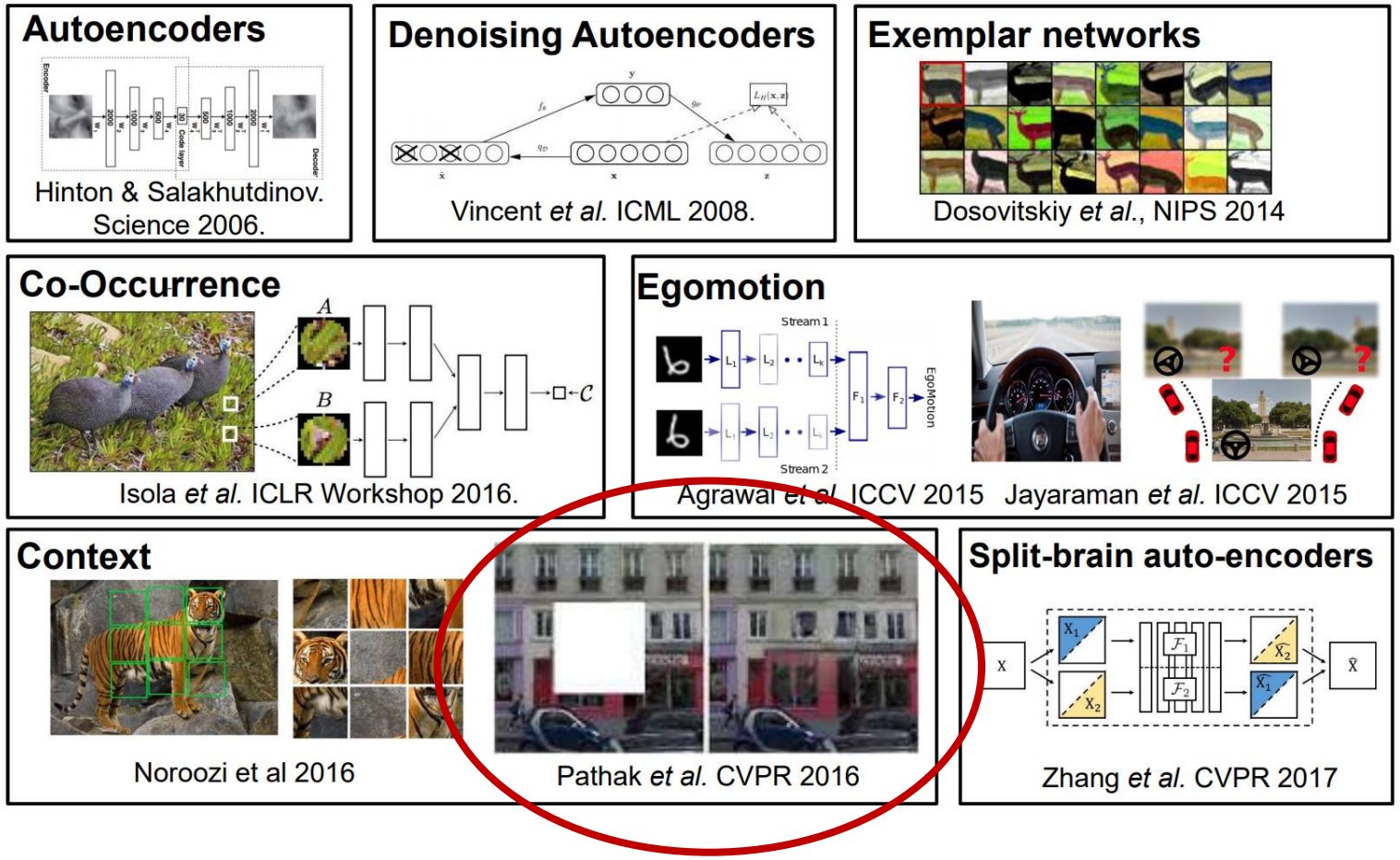
Science, 2006

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov



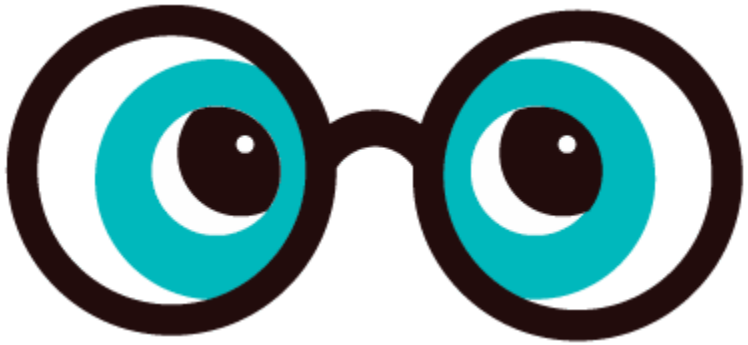
Self-supervised learning in computer vision

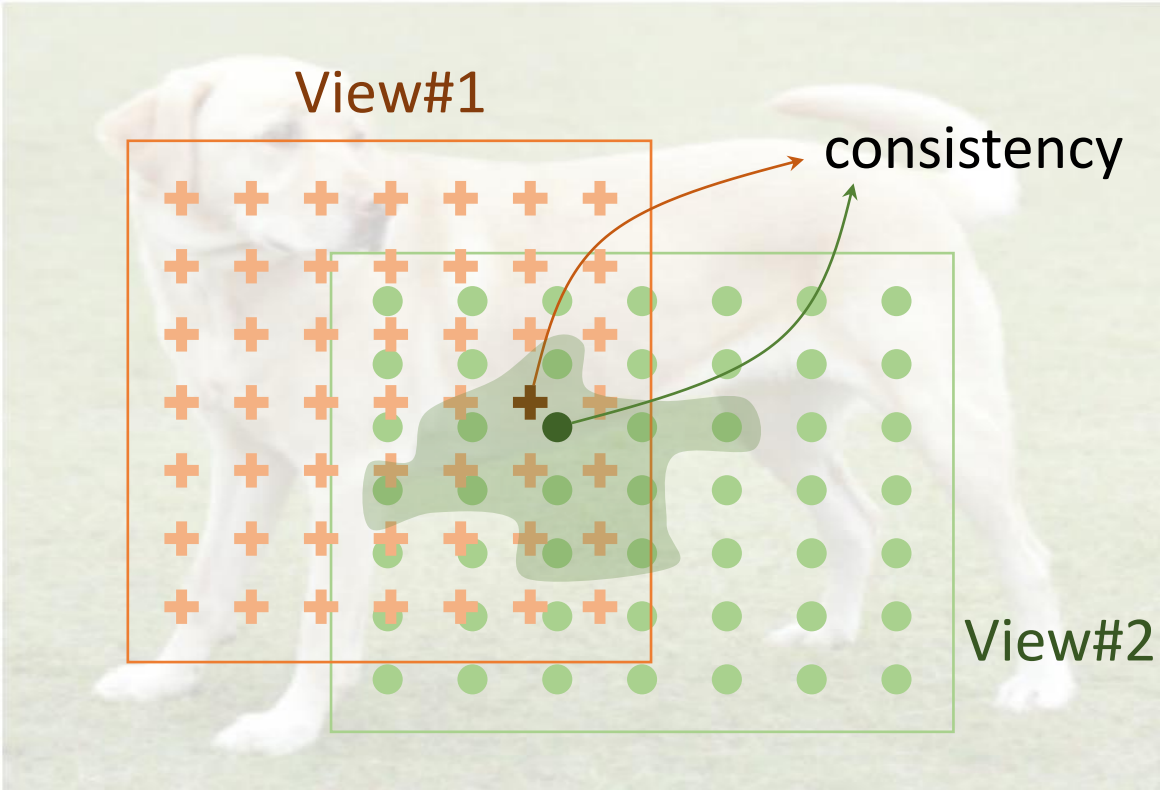


From Andrew Zisserman

Self-supervised learning in CV: Prediction of unknown input or using contrast

- MoCo (CVPR'2020)
- PixPro (CVPR'2021)
  - Inspired by eye movements

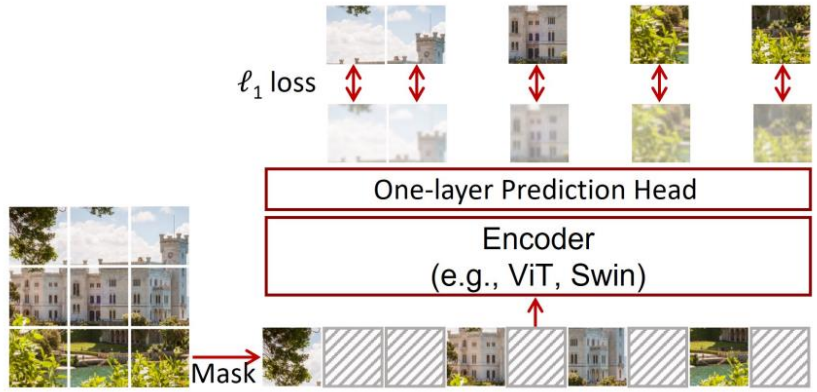




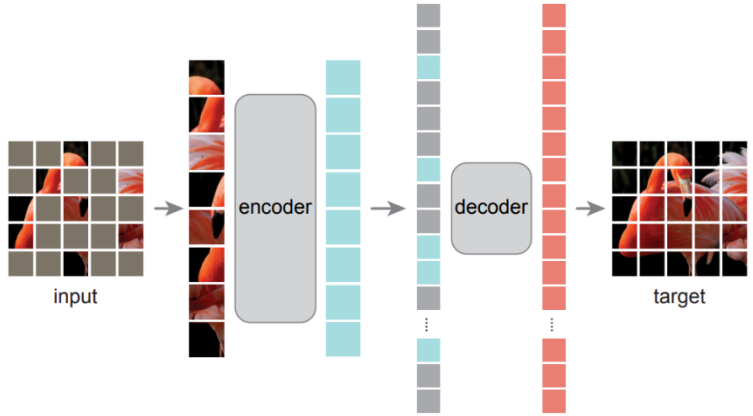
PixPro method

Self-supervised learning in CV: Prediction of unknown input

- GPT/BERT-like approaches
  - Image-GPT (OpenAI)
  - BEiT (MSRA)
  - SimMIM (MSRA)
  - MAE (Meta)
  - PeCo (MSRA)
  - data2vec (Meta)
  - CAE (Baidu)
  - …



SimMIM: A Simple Framework for Masked Image Modeling



MAE: Masked Autoencoders Are Scalable Vision Learners

BERT-like approach: differences between CV and NLP

- SimMIM (CVPR'2022)

I am a fluffy cat

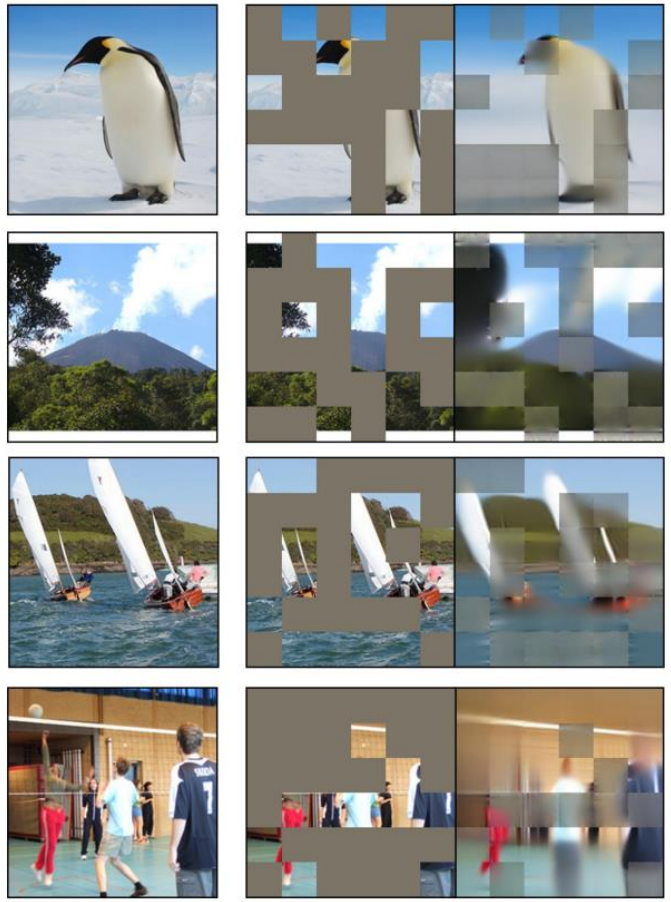
↑

I am a [mask] cat

- ✓ No spatial smoothness
- ✓ High-level
- ✓ discrete

→

- ✓ Spatial smoothness
- ✓ Raw and low-level
- ✓ continuous



Low mask ratio (15%), token classification

Large patch size and mask ratio (ps=32, ratio=60%), direct pixel value regression

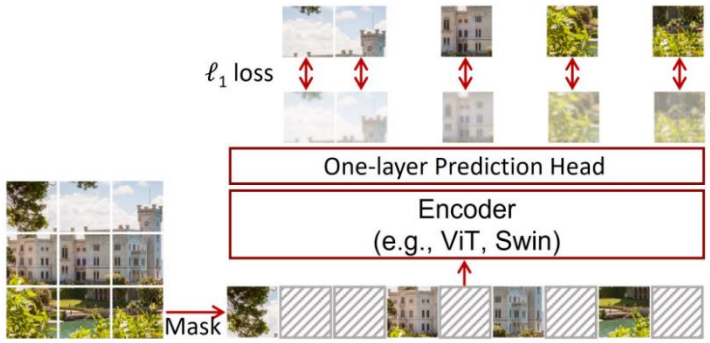
Take-home message

- Embrace converged architectures and learning methods in AI
  - Converged architectures
    - Transformer + domain-specific priors
      - NLP: Transformer
      - CV: Swin Transformer V1/V2
  - Converged pre-training methods
    - Prediction of unknown input
      - NLP: BERT/GPT/XLNet
      - CV: ImageGPT/BEiT/SimMIM/MAE/PeCo/data2vec/CAE …

Generic Transformer Unit

+

Domain-specific priors (by graph)



Acknowledgements

- Swin Transformer series benefit a lot from OpenMMLab