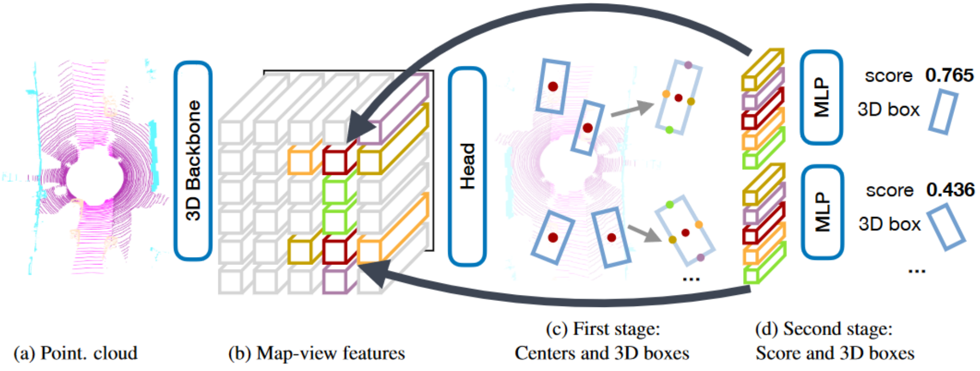
“Attention” and “Transformer” Architectures

James Hays

Recap – Lidar processing

- Popular CNN backbones aren't a direct fit for 3D point processing tasks.
- It's not clear how best to use deep learning on 3D data
  - Use a truly permutation invariant representation (PointNet)
  - Use a voxel representation (VoxelNet)
  - Use a bird's a view representation (PointPillars)
  - Create a range image (LaserNet)
- Multi-modal approaches (adding images, radar) help surprisingly little compared to lidar-only approaches.
- These alternate representations might be applicable more broadly, e.g. reasoning about depth estimates might be easier in bird's eye view (PseudoLidar)



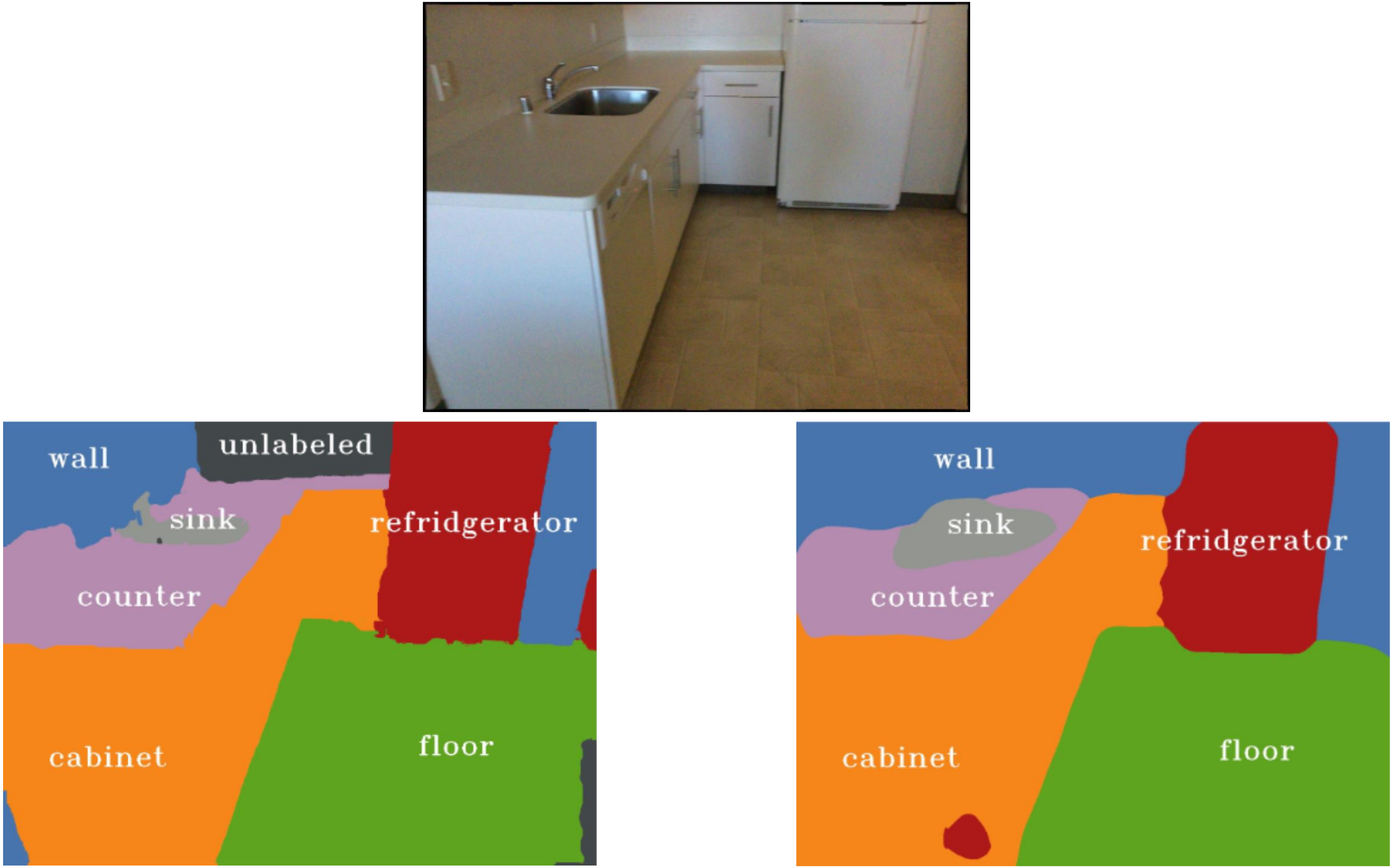
CenterPoint, which is near state of the art for 3D object detection when using VoxelNet backbone.

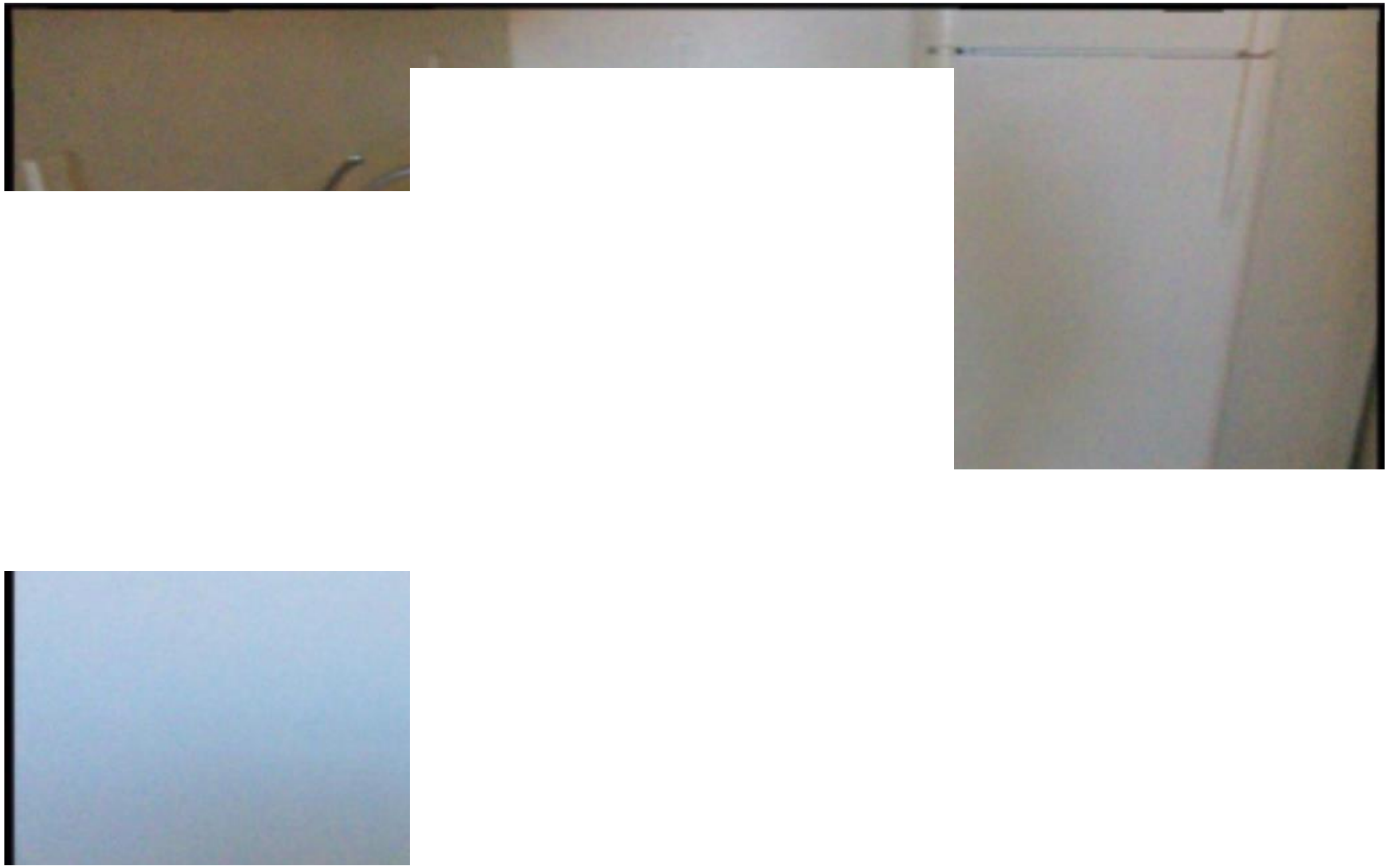
Center-based 3D Object Detection and Tracking. Tianwei Yin, Xingyi Zhou, Philipp Krähenbühl. CVPR 2021

https://paperswithcode.com/sota/3d-object-detection-on-nuscenes

Outline

- Context and Receptive Field
- Going Beyond Convolutions in…
  - Text
  - Point Clouds
  - Images


wall
unlabeled
sink
refridgerator
counter
cabinet
floor
wall
sink
refridgerator
counter
cabinet
floor



Language understanding

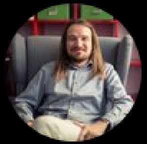
… serve …

Language understanding

… great serve from Djokovic …

Language understanding

… be right back after I serve these salads …



Brendan Dolan-Gavitt
@moyix

The latest generation of adversarial image attacks is, uh, somewhat simpler to carry out openai.com/blog/multimoda...

Attacks in the wild

We refer to these attacks as *typographic attacks*. We believe attacks such as those described above are far from simply an academic concern. By exploiting the model's ability to read text robustly, we find that even *photographs of hand-written text* can often fool the model. Like the Adversarial Patch,[22] this attack works in the wild; but unlike such attacks, it requires no more technology than pen and paper.

Attack text label iPod

Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%

Granny Smith	0.1%
iPod	99.7%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.0%

When we put a label saying "iPod" on this Granny Smith apple, the model erroneously classifies it as an iPod in the zero-shot setting.

3:38 PM · Mar 4, 2021 · Twitter Web App

Mark O. Riedl
@mark_riedl

Replying to @mark_riedl

In case of AI uprising...

6:42 PM · Mar 4, 2021 · Twitter for iPad

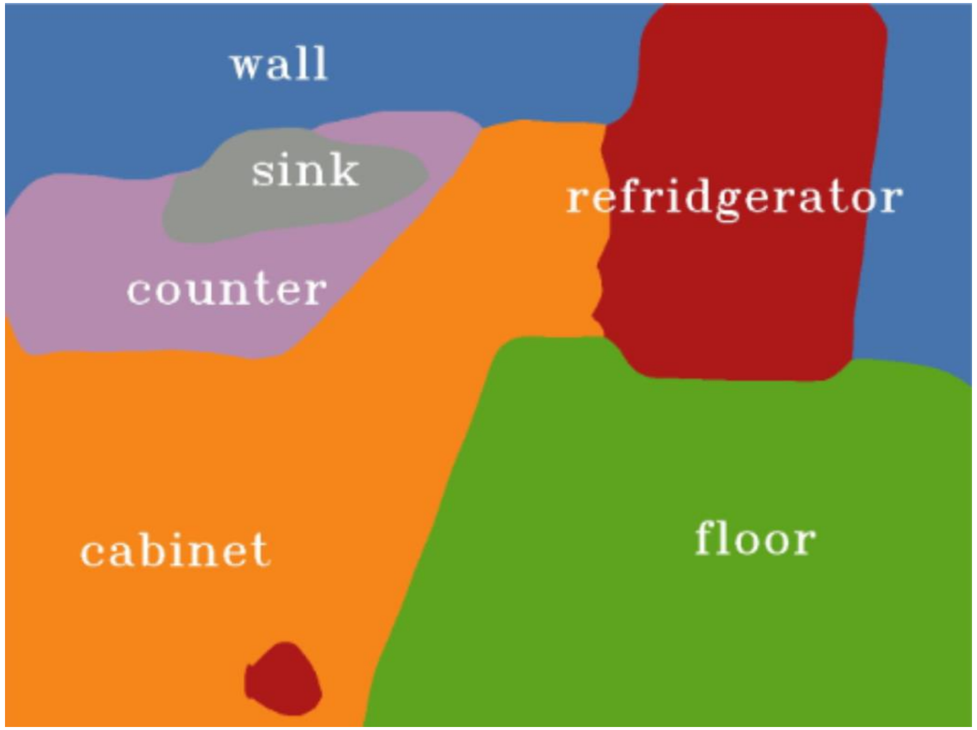
Mark O. Riedl
@mark_riedl

Replying to @mark_riedl

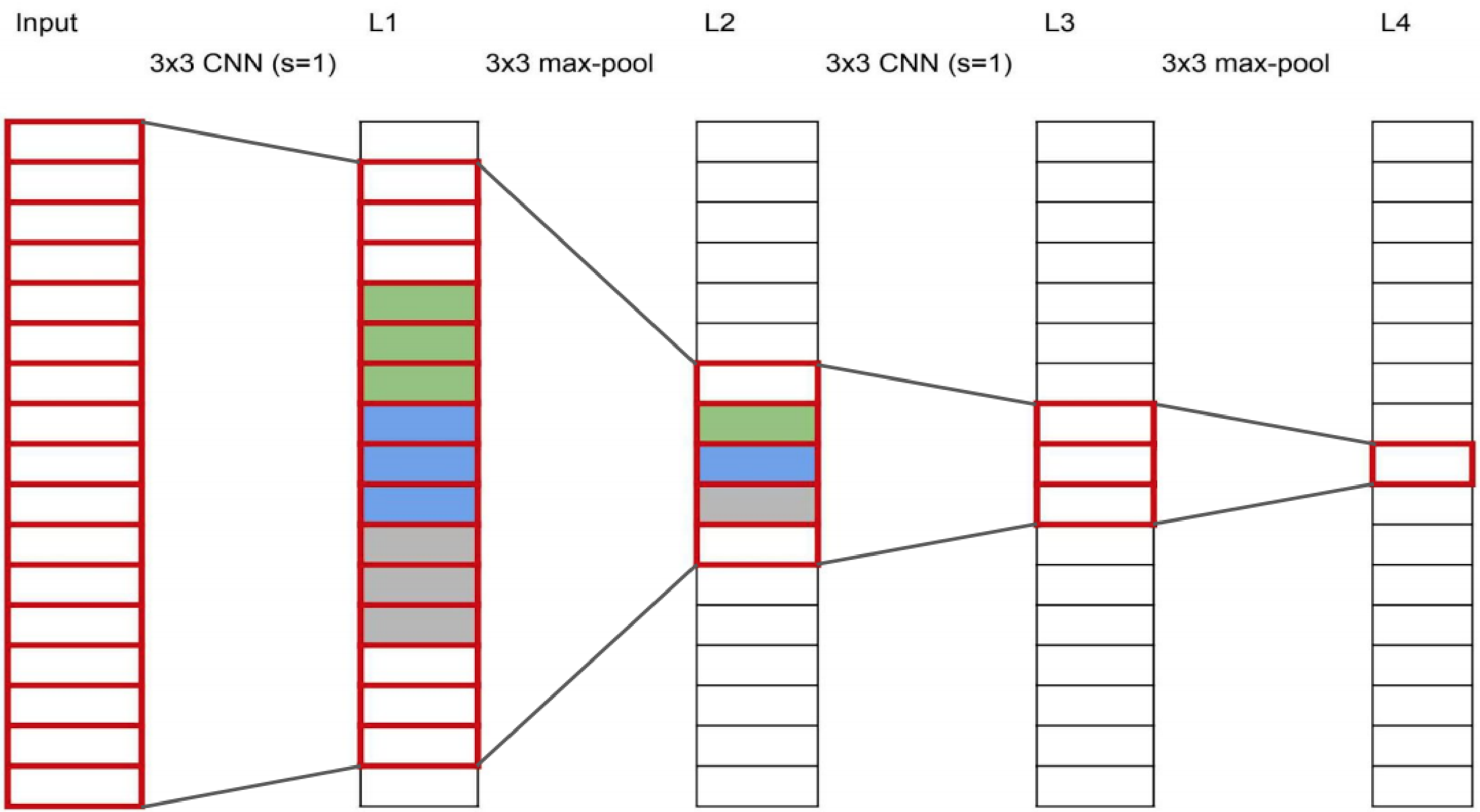
Upon further reflection, neural language models aren't always so good with negations. I recommend this instead

9:28 PM · Mar 4, 2021 · Twitter for iPad

So how do we fix these problems?

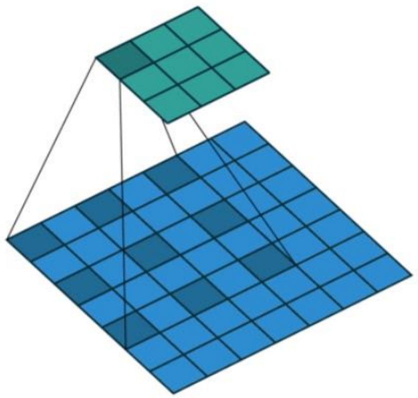


Receptive field

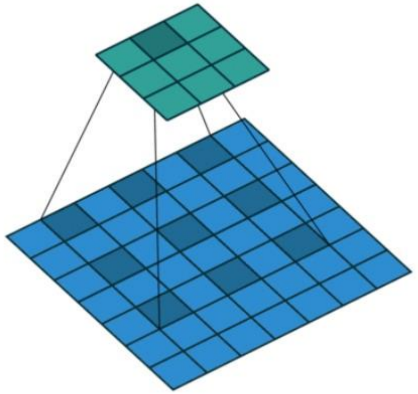


Slide Credit: Frank Dellaert https://dellaert.github.io/19F-4476/resources/receptiveField.pdf

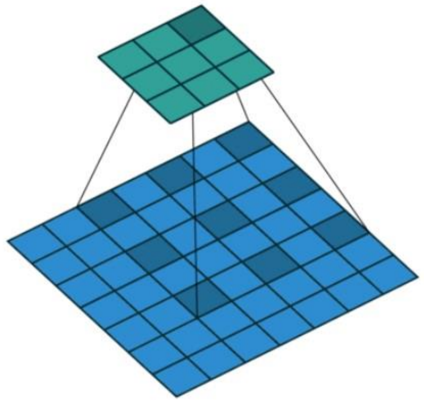
Dilated Convolution



No padding, no stride, dilation



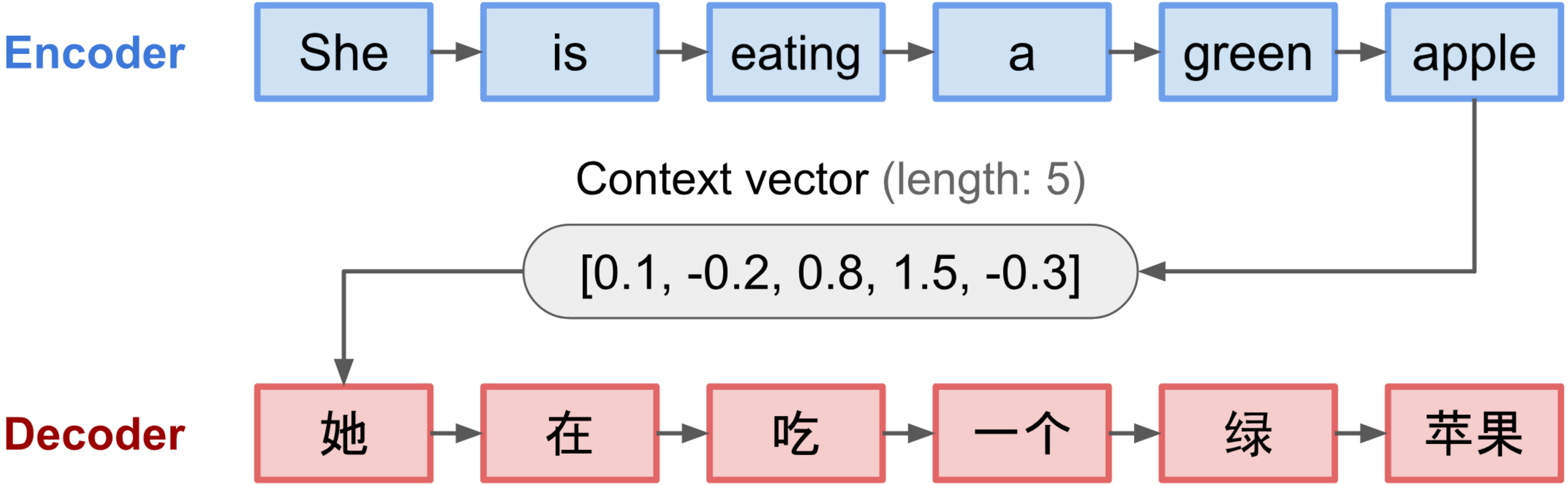
No padding, no stride, dilation



No padding, no stride, dilation

Figure source: https://github.com/vdumoulin/conv_arithmetic

Sequence 2 Sequence models in language



Source: https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

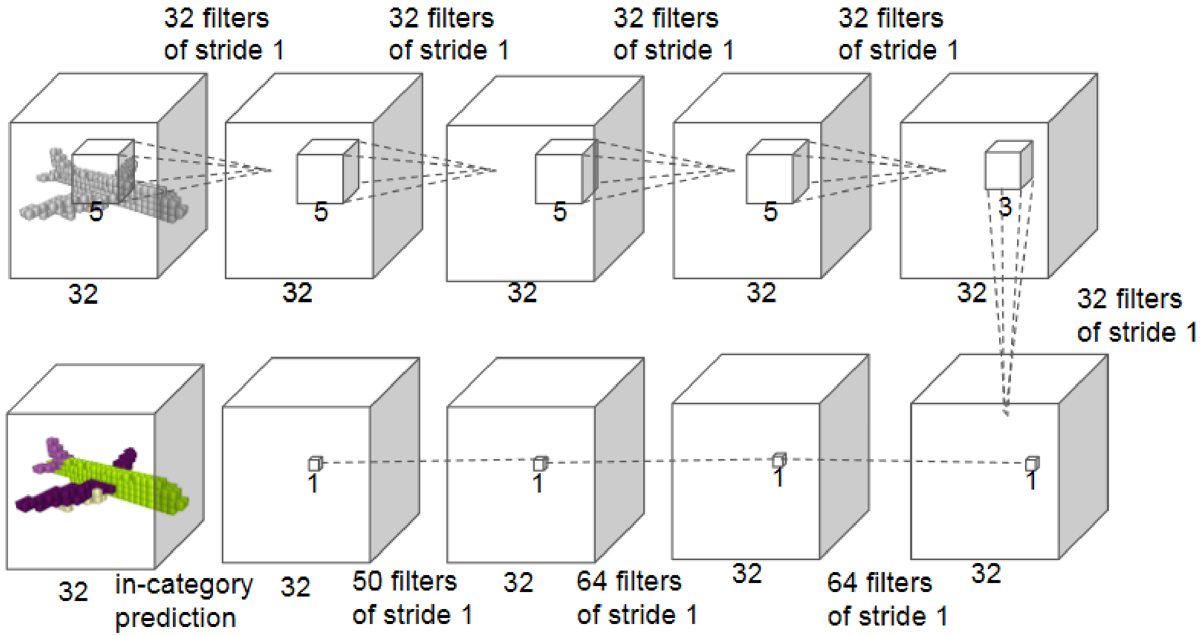


Figure 10. **Baseline 3D CNN segmentatic**
network is fully convolutional and predicts p
voxel.

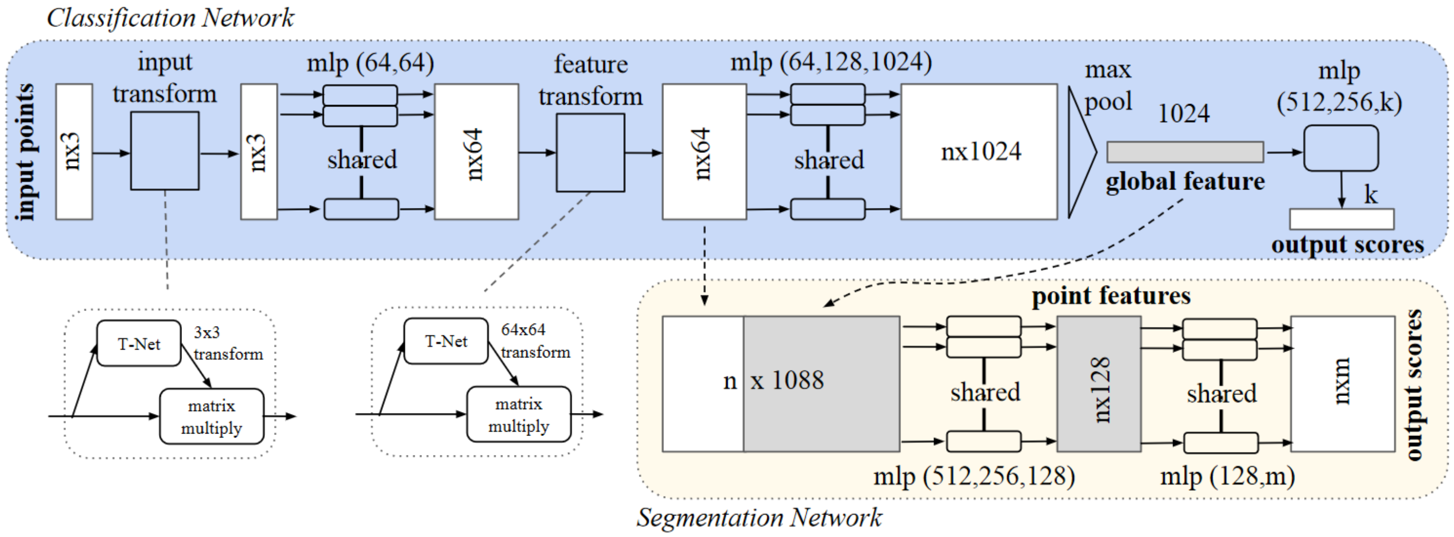


Figure 2. **PointNet Architecture.** The classification network takes n points as input, applies input and feature transformations, and then aggregates point features by max pooling. The output is classification scores for k classes. The segmentation network is an extension to the classification net. It concatenates global and local features and outputs per point scores. "mlp" stands for multi-layer perceptron, numbers in bracket are layer sizes. Batchnorm is used for all layers with ReLU. Dropout layers are used for the last mlp in classification net.

Outline

- Context and Receptive Field
- Going Beyond Convolutions in…
  - Text
  - Point Clouds
  - Images

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

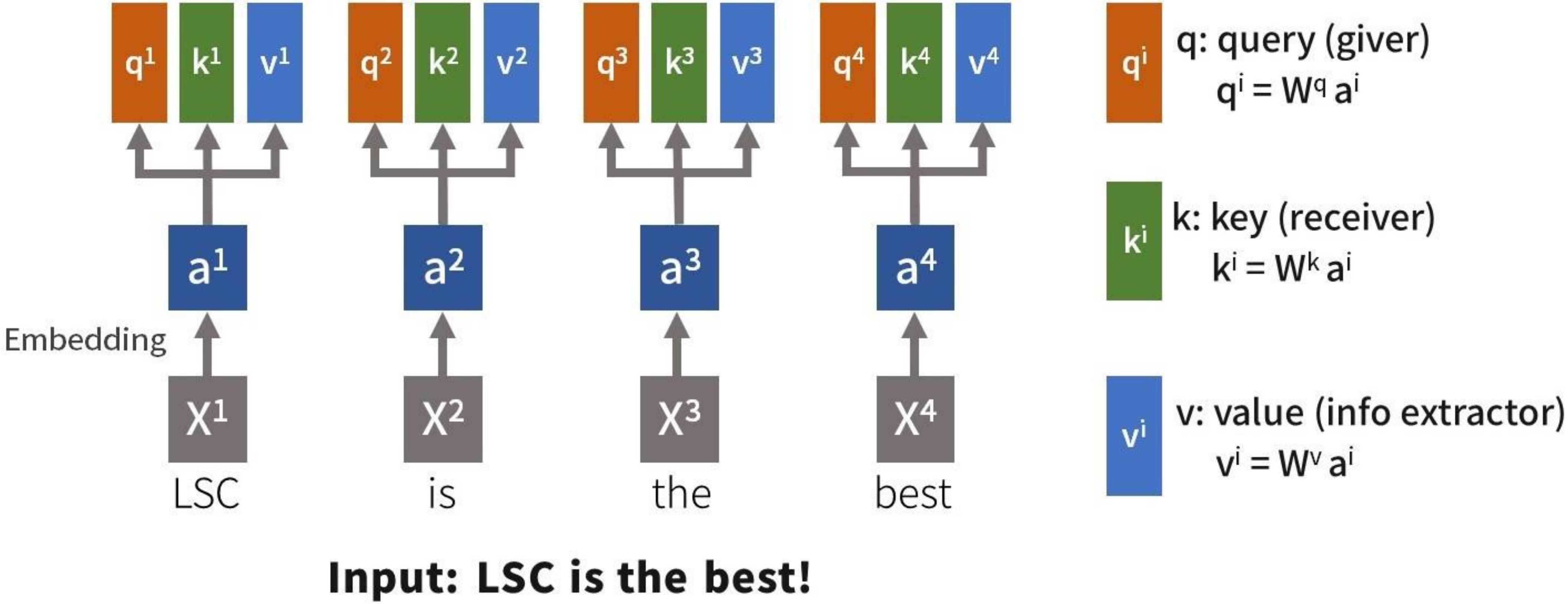
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

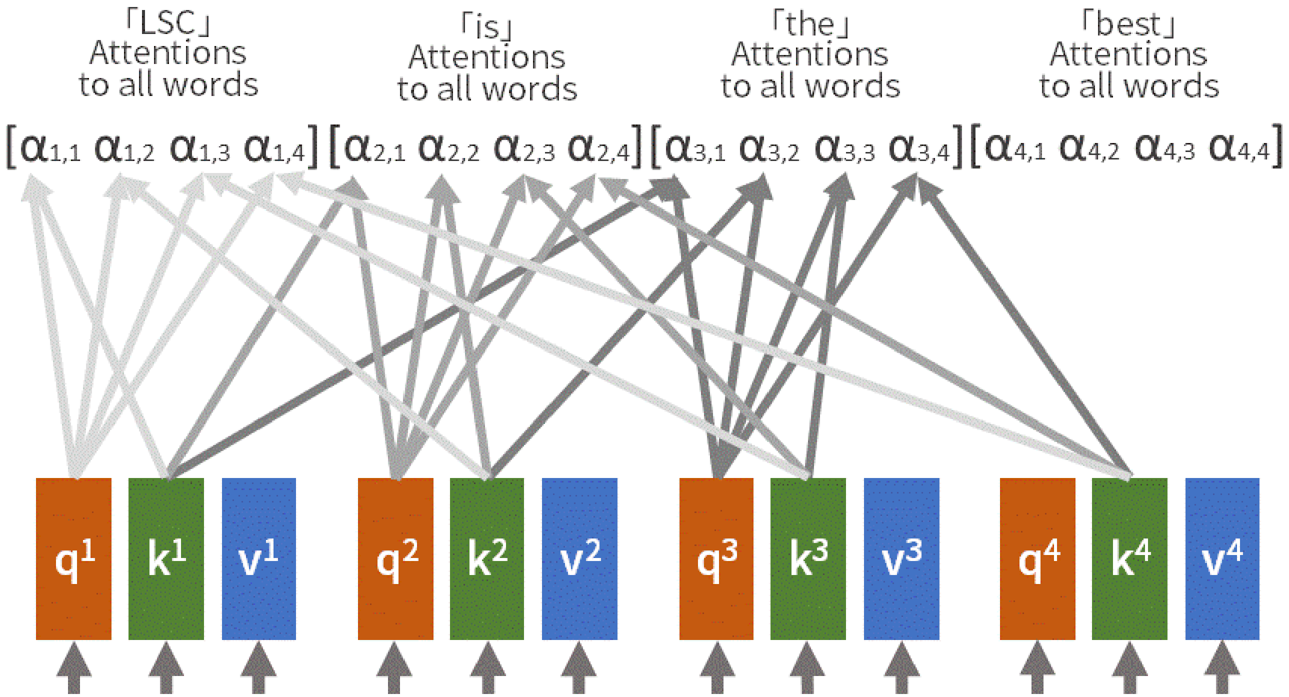
Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer,

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



From https://medium.com/lsc-psd/introduction-of-self-attention-layer-in-transformer-fc7bff63f3bc

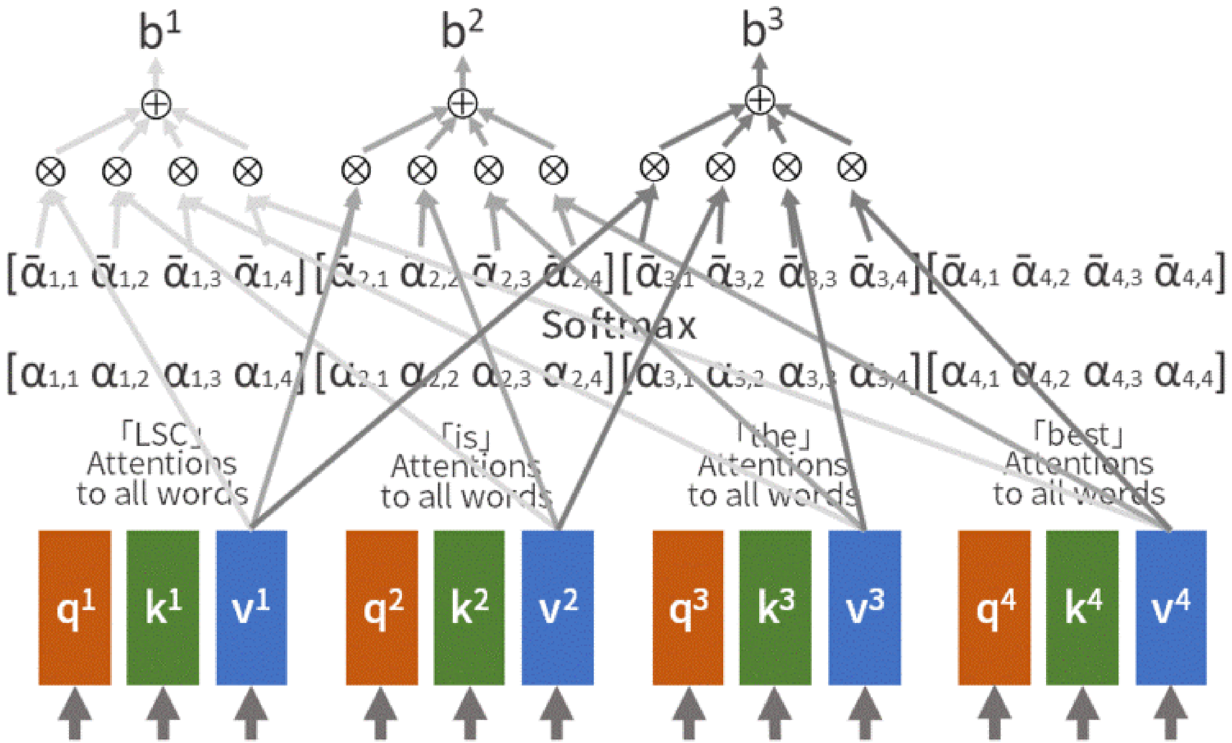


$$\alpha_{i,j} = \frac{q^i \cdot k^j}{\sqrt{d}}$$

d: dimension of q, k

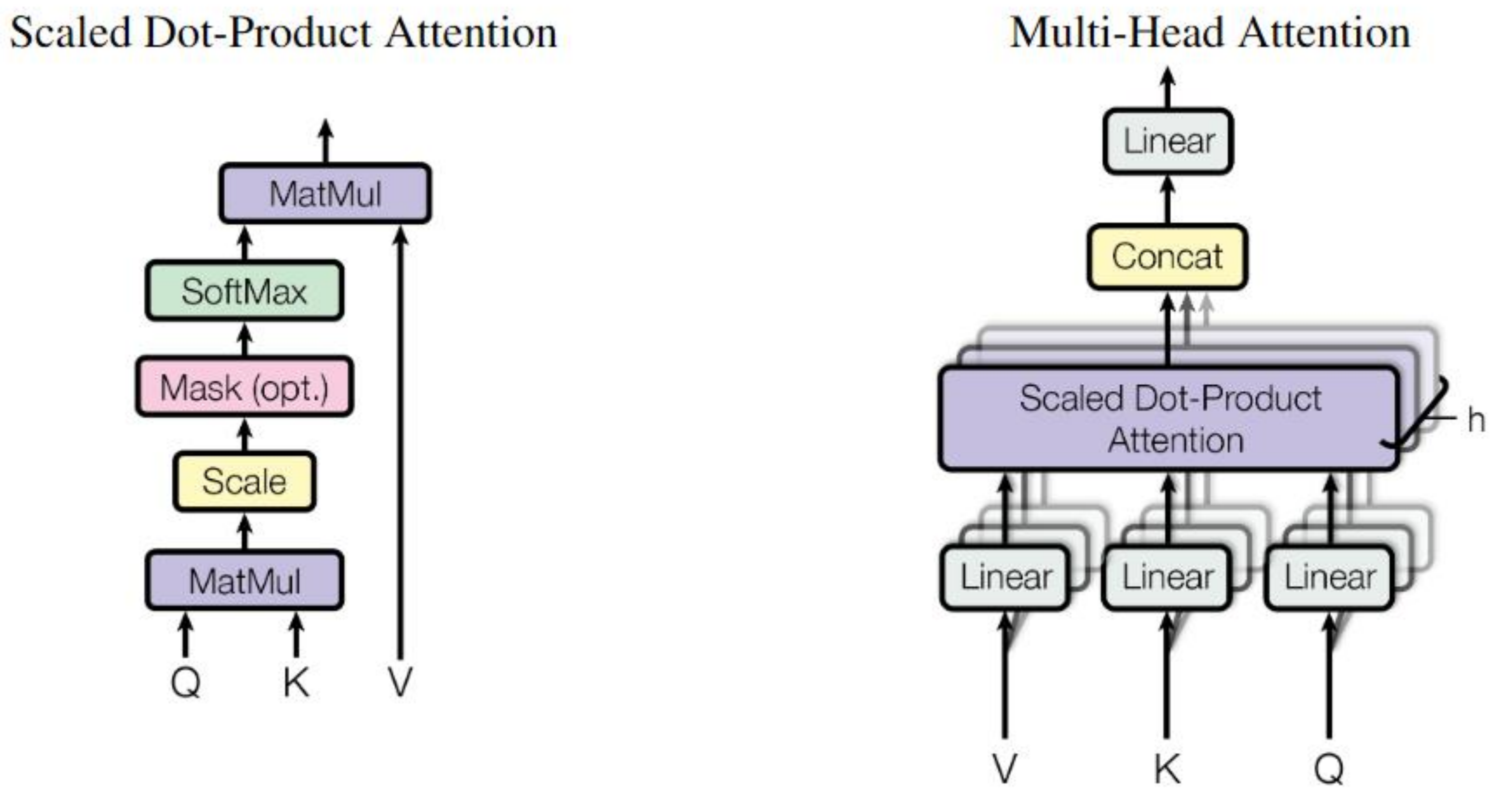
$$\underset{\text{Attention Matrix}}{A} = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \alpha_{1,4} \\ \alpha_{2,1} & \alpha_{2,2} & \alpha_{2,3} & \alpha_{2,4} \\ \alpha_{3,1} & \alpha_{3,2} & \alpha_{3,3} & \alpha_{3,4} \\ \alpha_{4,1} & \alpha_{4,2} & \alpha_{4,3} & \alpha_{4,4} \end{bmatrix}$$

From https://medium.com/lsc-psd/introduction-of-self-attention-layer-in-transformer-fc7bff63f3bc



$$b^i = \sum_j \bar{\alpha}_{i,j} v^j$$

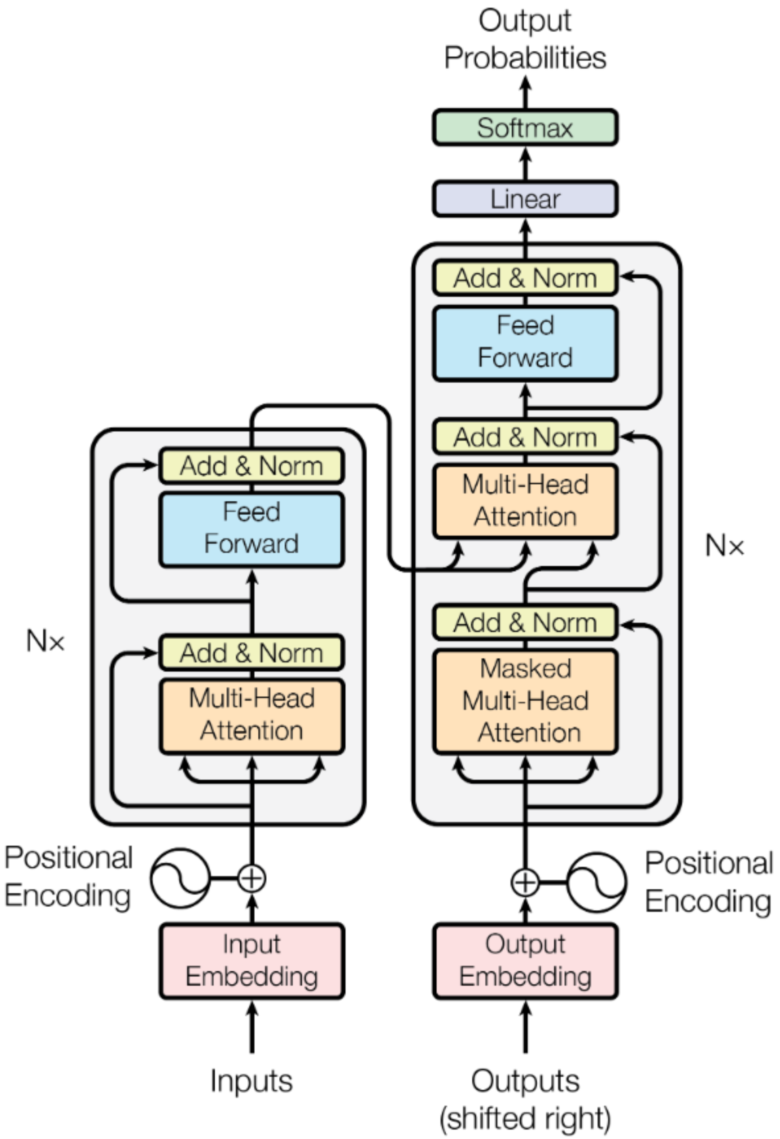
Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$


Scaled Dot-Product Attention
MatMul
SoftMax
Mask (opt.)
Scale
MatMul
Q
K
V
Multi-Head Attention
Linear
Concat
Scaled Dot-Product Attention
h
Linear
Linear
Linear
V
K
Q

Transformer Architecture

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	



Outline

- Context and Receptive Field
- Going Beyond Convolutions in…
    - Text
    - Point Clouds
    - Images

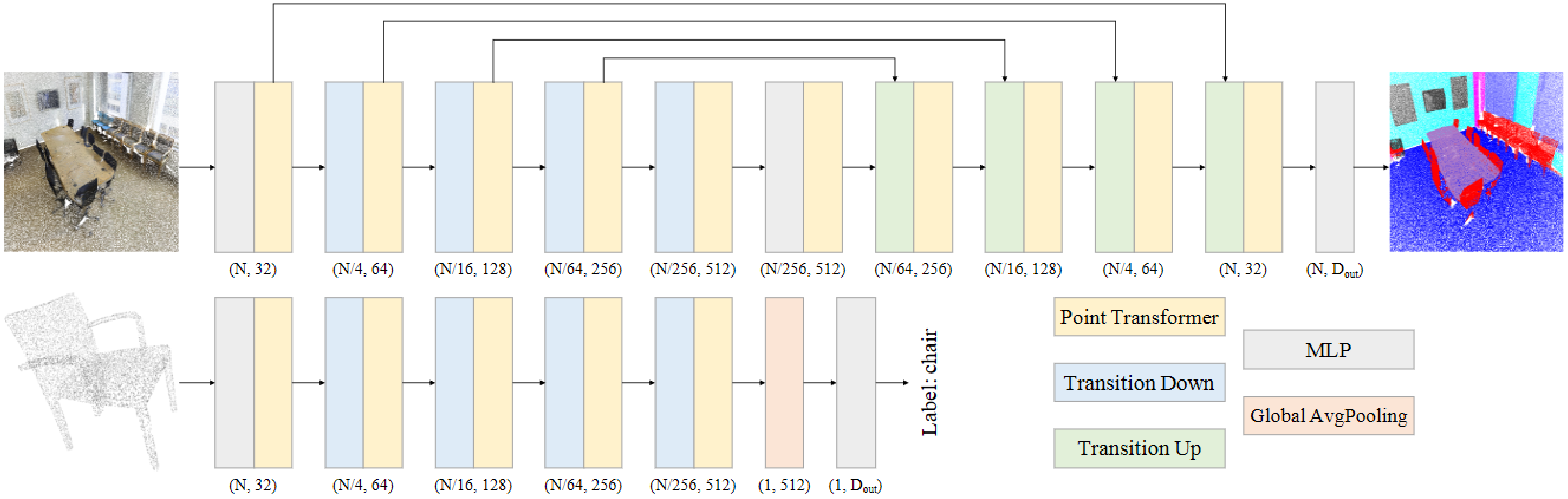
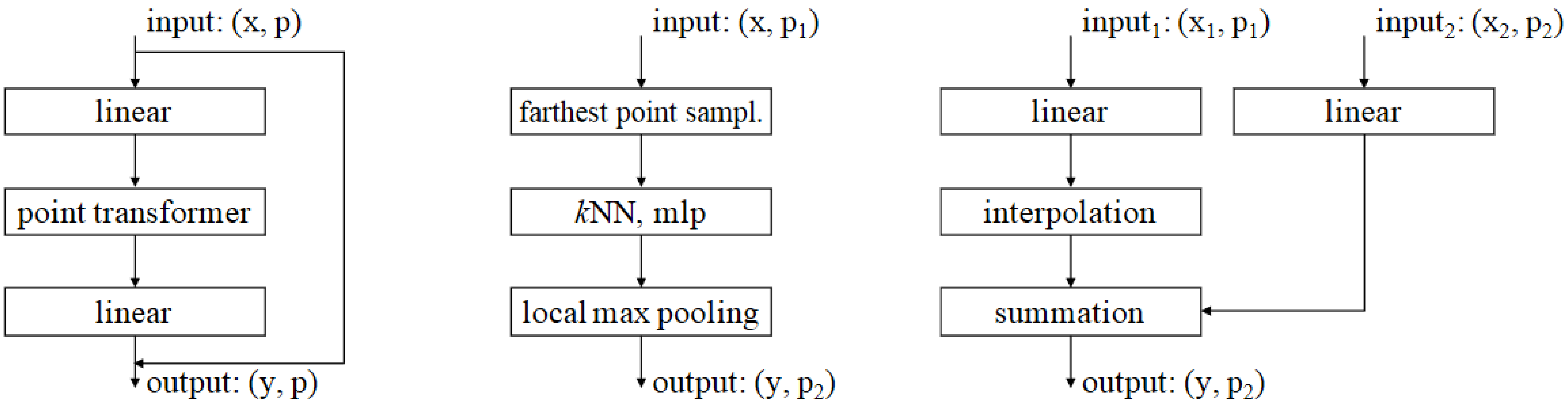
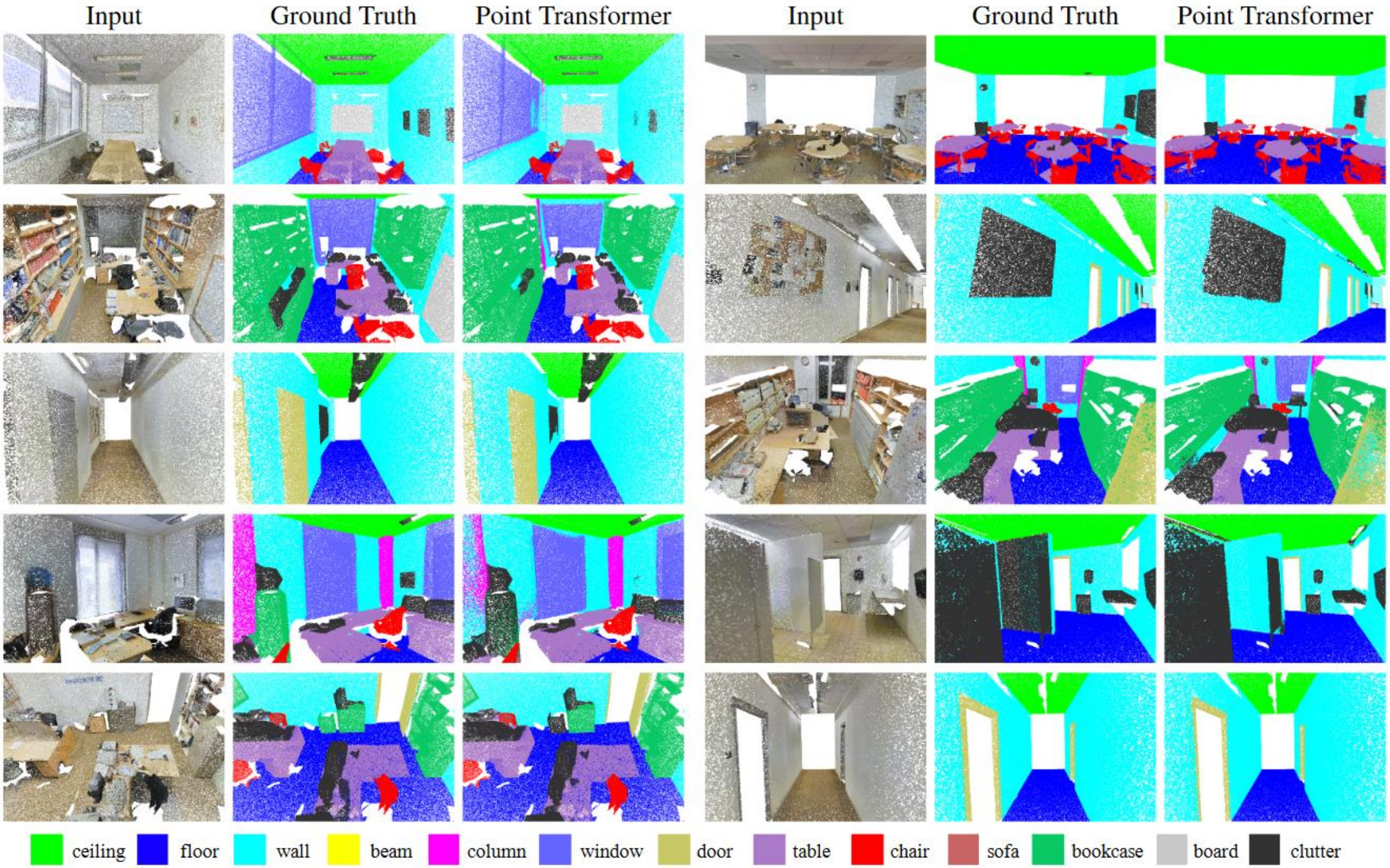


Figure 3. Point transformer networks for semantic segmentation (top) and classification (bottom).



(a) point transformer block (b) transition down (c) transition up

Point Transformer. Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, Vladlen Koltun


Input
Ground Truth
Point Transformer
Input
Ground Truth
Point Transformer
ceiling
floor
wall
beam
column
window
door
table
chair
sofa
bookcase
board
clutter

Method	OA	mAcc	mIoU	ceiling	floor	wall	beam	column	window	door	table	chair	sofa	bookcase	board	clutter
PointNet [22]	–	49.0	41.1	88.8	97.3	69.8	0.1	3.9	46.3	10.8	59.0	52.6	5.9	40.3	26.4	33.2
SegCloud [32]	–	57.4	48.9	90.1	96.1	69.9	0.0	18.4	38.4	23.1	70.4	75.9	40.9	58.4	13.0	41.6
TangentConv [31]	–	62.2	52.6	90.5	97.7	74.0	0.0	20.7	39.0	31.3	77.5	69.4	57.3	38.5	48.8	39.8
PointCNN [18]	85.9	63.9	57.3	92.3	98.2	79.4	0.0	17.6	22.8	62.1	74.4	80.6	31.7	66.7	62.1	56.7
SPGraph [14]	86.4	66.5	58.0	89.4	96.9	78.1	0.0	42.8	48.9	61.6	84.7	75.4	69.8	52.6	2.1	52.2
PCCN [38]	–	67.0	58.3	92.3	96.2	75.9	0.3	6.0	69.5	63.5	66.9	65.6	47.3	68.9	59.1	46.2
PointWeb [50]	87.0	66.6	60.3	92.0	98.5	79.4	0.0	21.1	59.7	34.8	76.3	88.3	46.9	69.3	64.9	52.5
HPEIN [12]	87.2	68.3	61.9	91.5	98.2	81.4	0.0	23.3	65.3	40.0	75.5	87.7	58.5	67.8	65.6	49.4
MinkowskiNet [33]	–	71.7	65.4	91.8	98.7	86.2	0.0	34.1	48.9	62.4	81.6	89.8	47.2	74.9	74.4	58.6
KPConv [33]	–	72.8	67.1	92.8	97.3	82.4	0.0	23.9	58.0	69.0	81.5	91.0	75.4	75.3	66.7	58.9
PointTransformer	**90.8**	**76.5**	**70.4**	94.0	98.5	86.3	0.0	38.0	63.4	74.3	89.1	82.4	74.3	80.2	76.0	59.3

Table 1. Semantic segmentation results on the S3DIS dataset, evaluated on Area 5.

Method	input	mAcc	OA
3DShapeNets [43]	voxel	77.3	84.7
VoxNet [20]	voxel	83.0	85.9
Subvolume [23]	voxel	86.0	89.2
MVCNN [30]	image	–	90.1
PointNet [22]	point	86.2	89.2
PointNet++ [24]	point	–	91.9
SpecGCN [36]	point	–	92.1
PointCNN [18]	point	88.1	92.2
DGCNN [40]	point	90.2	92.2
PointWeb [50]	point	89.4	92.3
SpiderCNN [44]	point	–	92.4
PointConv [42]	point	–	92.5
KPConv [33]	point	–	92.9
InterpCNN [19]	point	–	93.0
PointTransformer	point	**90.6**	**93.7**

Table 3. Shape classification results on the ModelNet40 dataset.

https://paperswithcode.com/sota/3d-point-cloud-classification-on-modelnet40

Outline

- Context and Receptive Field
- Going Beyond Convolutions in…
  - Text
  - Point Clouds
  - Images

Preprint. Under review.

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]

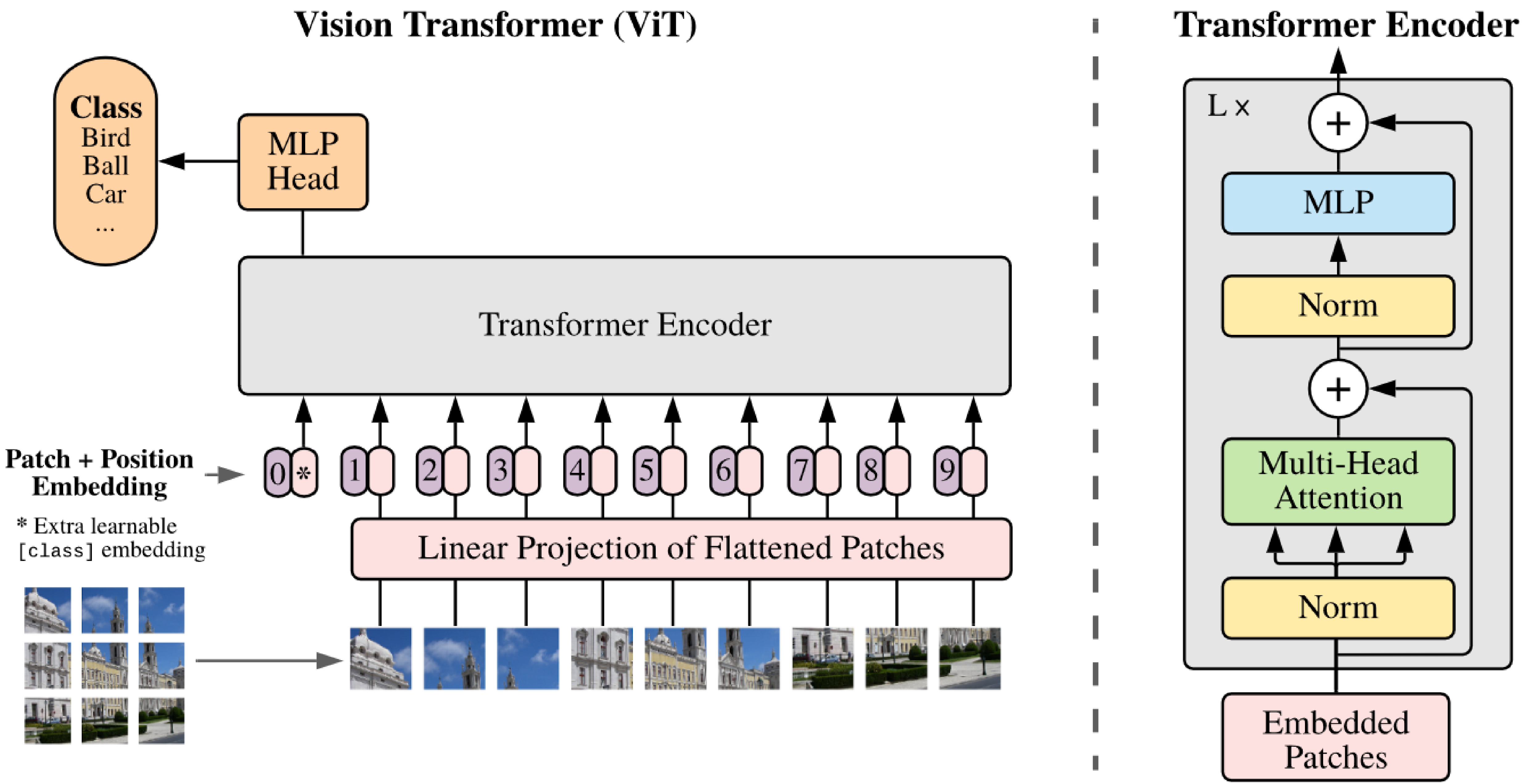
[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

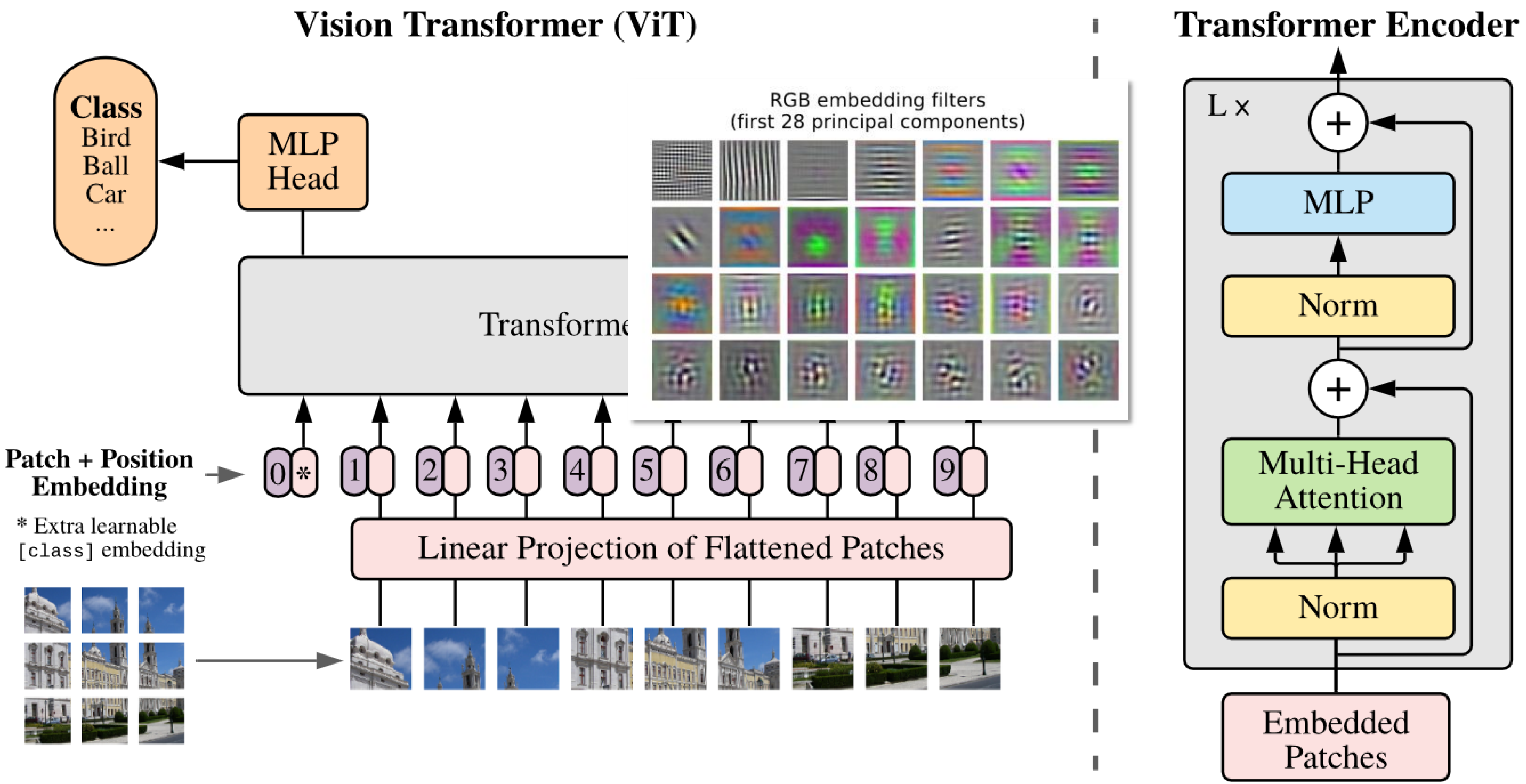
`{adosovitskiy, neilhoulsby}@google.com`

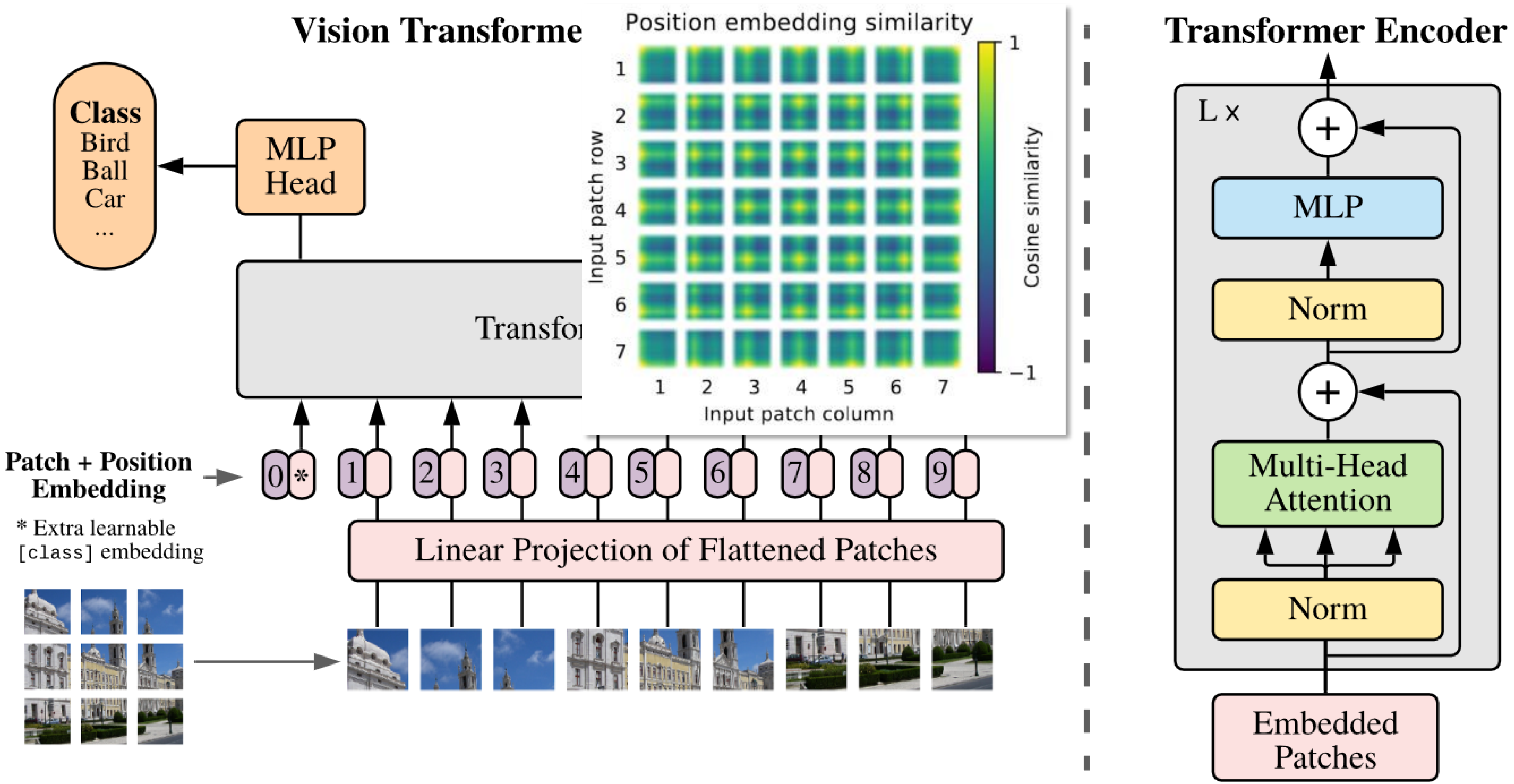
ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]



Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches


Vision Transformer (ViT)
Transformer Encoder
Class
Bird
Ball
Car
...
MLP Head
Transforme
RGB embedding filters
(first 28 principal components)
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches


Vision Transforme
Class
Bird
Ball
Car
...
MLP
Head
Transfor
Patch + Position
Embedding
* Extra learnable
[class] embedding
0 *
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Position embedding similarity
Input patch row
Input patch column
Cosine similarity
1
−1
Transformer Encoder
L ×
MLP
Norm
Multi-Head
Attention
Norm
Embedded
Patches

Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

Table 1: Details of Vision Transformer model variants.

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21K (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** $\pm$ 0.04	87.76 $\pm$ 0.03	85.30 $\pm$ 0.02	87.54 $\pm$ 0.02	88.4/88.5*
ImageNet ReaL	**90.72** $\pm$ 0.05	90.54 $\pm$ 0.03	88.62 $\pm$ 0.05	90.54	90.55
CIFAR-10	**99.50** $\pm$ 0.06	99.42 $\pm$ 0.03	99.15 $\pm$ 0.03	99.37 $\pm$ 0.06	–
CIFAR-100	**94.55** $\pm$ 0.04	93.90 $\pm$ 0.05	93.25 $\pm$ 0.05	93.51 $\pm$ 0.08	–
Oxford-IIIT Pets	**97.56** $\pm$ 0.03	97.32 $\pm$ 0.11	94.67 $\pm$ 0.15	96.62 $\pm$ 0.23	–
Oxford Flowers-102	99.68 $\pm$ 0.02	**99.74** $\pm$ 0.00	99.61 $\pm$ 0.02	99.63 $\pm$ 0.03	–
VTAB (19 tasks)	**77.63** $\pm$ 0.23	76.28 $\pm$ 0.46	72.72 $\pm$ 0.21	76.29 $\pm$ 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

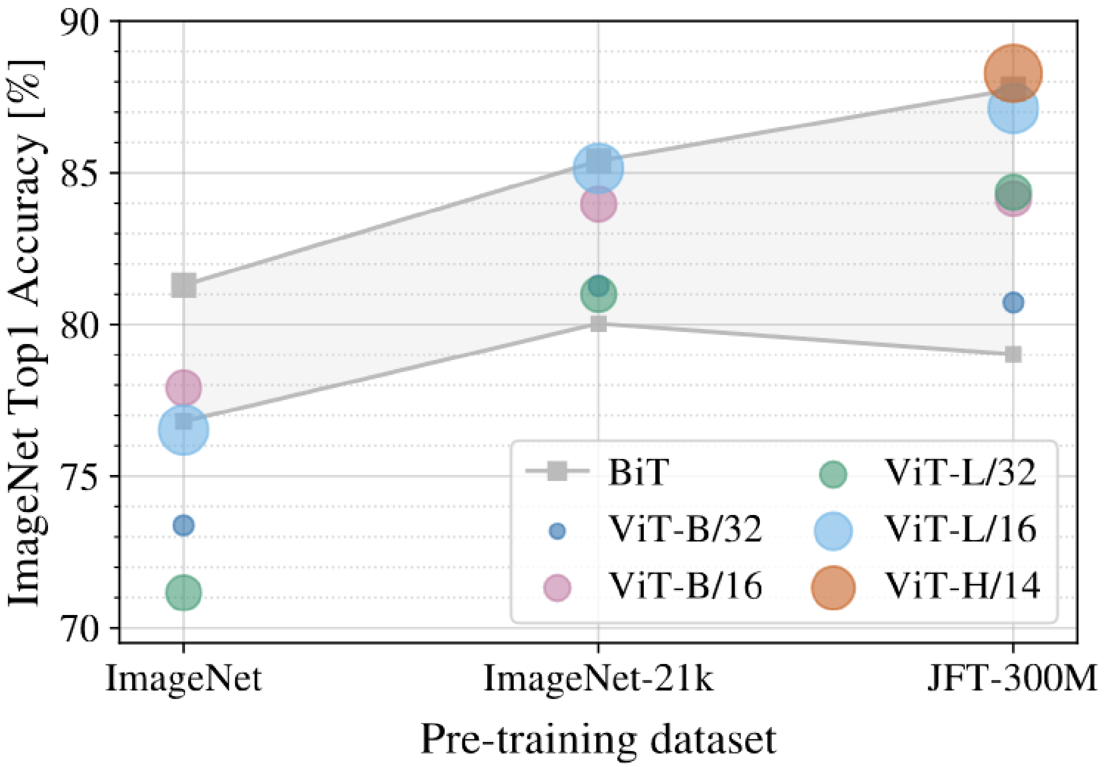


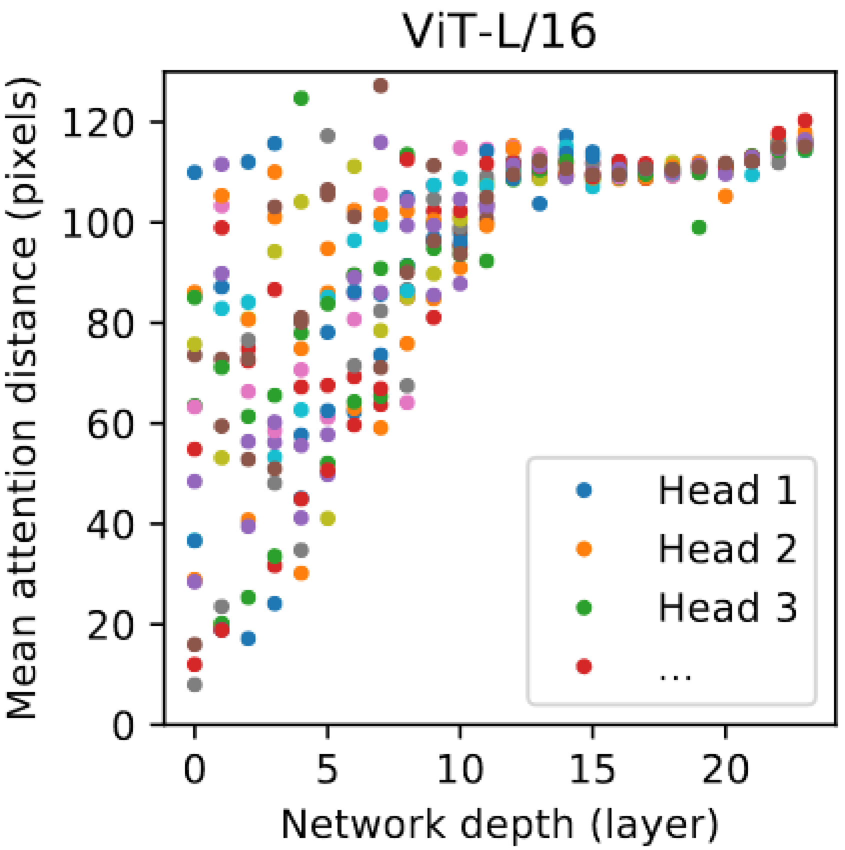
Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

When trained on mid-sized datasets such as ImageNet, such models yield modest accuracies of a few percentage points below ResNets of comparable size. This seemingly discouraging outcome maybe expected: Transformers lack some of the inductive biases inherent to CNNs, such as translation equivariance and locality, and therefore do not generalize well when trained on insufficient amounts of data.

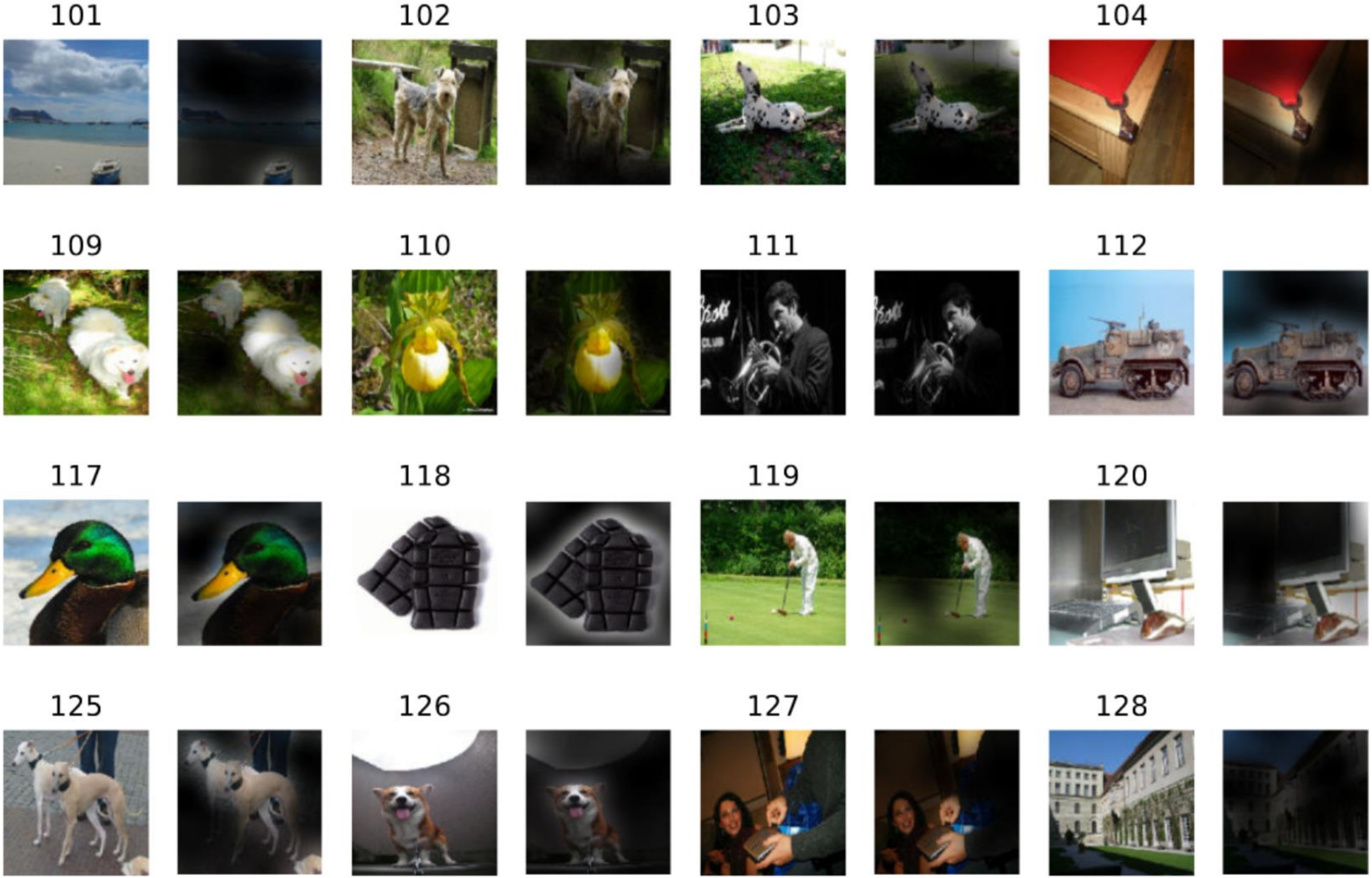
However, the picture changes if the models are trained on larger datasets (14M-300M images). We find that large scale training trumps inductive bias.

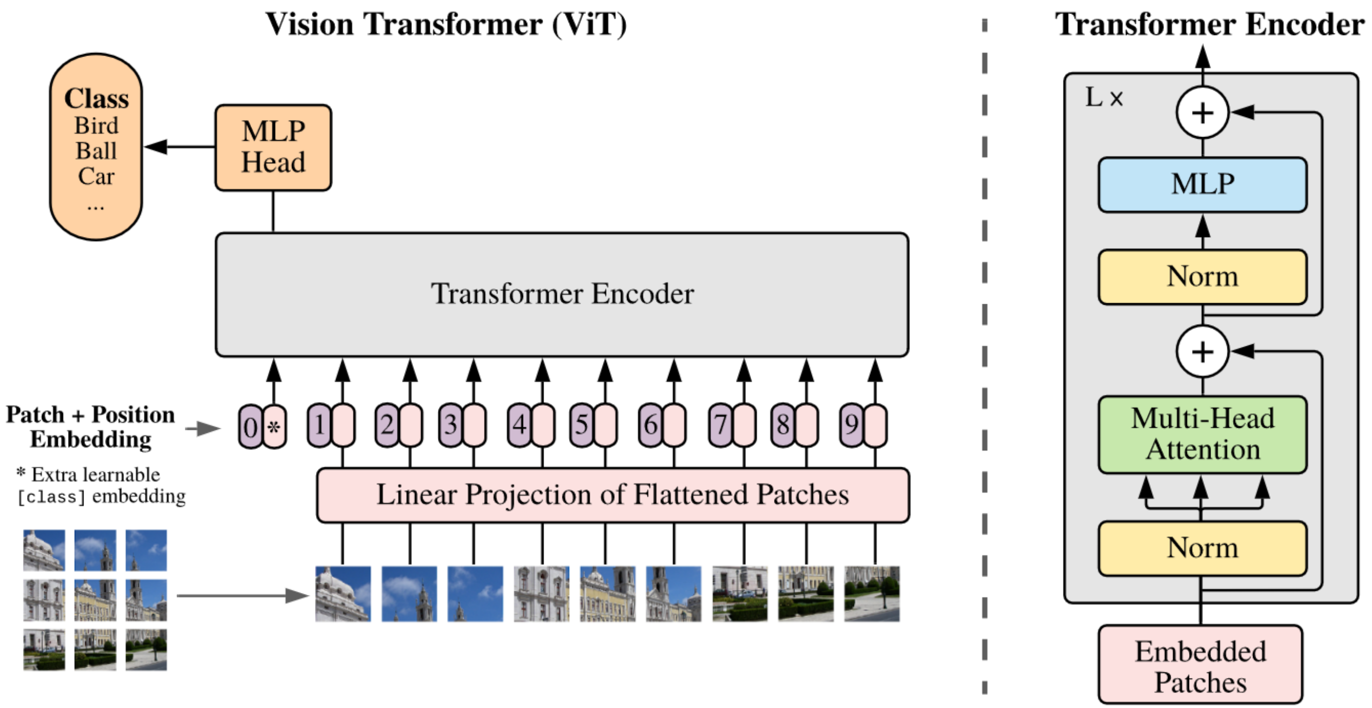
Dosovitskiy et al.

https://paperswithcode.com/sota/image-classification-on-imagenet

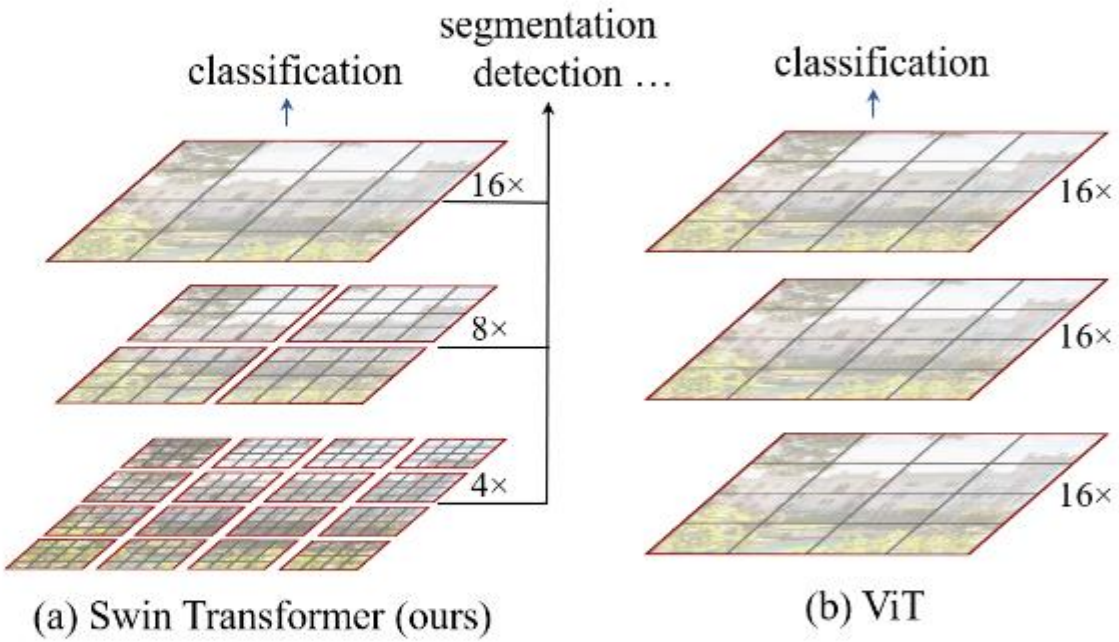


ViT-L/16
Mean attention distance (pixels)
Network depth (layer)
0
20
40
60
80
100
120
0
5
10
15
20
Head 1
Head 2
Head 3
...


101
102
103
104
109
110
111
112
117
118
119
120
125
126
127
128



- This can’t be ideal, right?



Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo

https://paperswithcode.com/sota/instance-segmentation-on-coco

Summary

- "Attention" models outperform recurrent models and convolutional models for sequence processing. They allow long range interactions.
- These models do best with LOTS of training data
- They seem like a good fit for point processing tasks, although maybe there isn't enough data for them to outperform other methods.
- Surprisingly, they seem to outperform convolutional networks for image processing tasks. Again, long range interactions might be more important than we realized.
- Naïve attention mechanisms have quadratic complexity with the number of input tokens, but there are often workarounds for this.