Masked Autoencoders Are Scalable Vision Learners

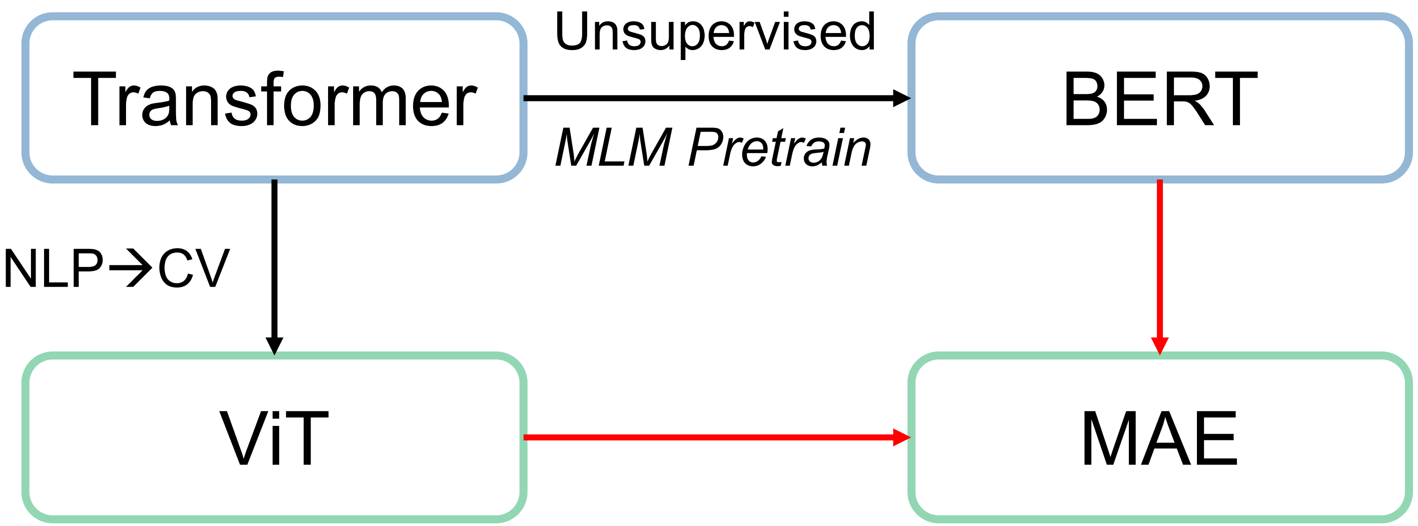
Arxiv 2021

Kaiming He*, Xinlei Chen*, Saining Xie, Yanghao Li, Piotr Doll´ar, Ross Girshick
Facebook AI Research (FAIR)

Outline

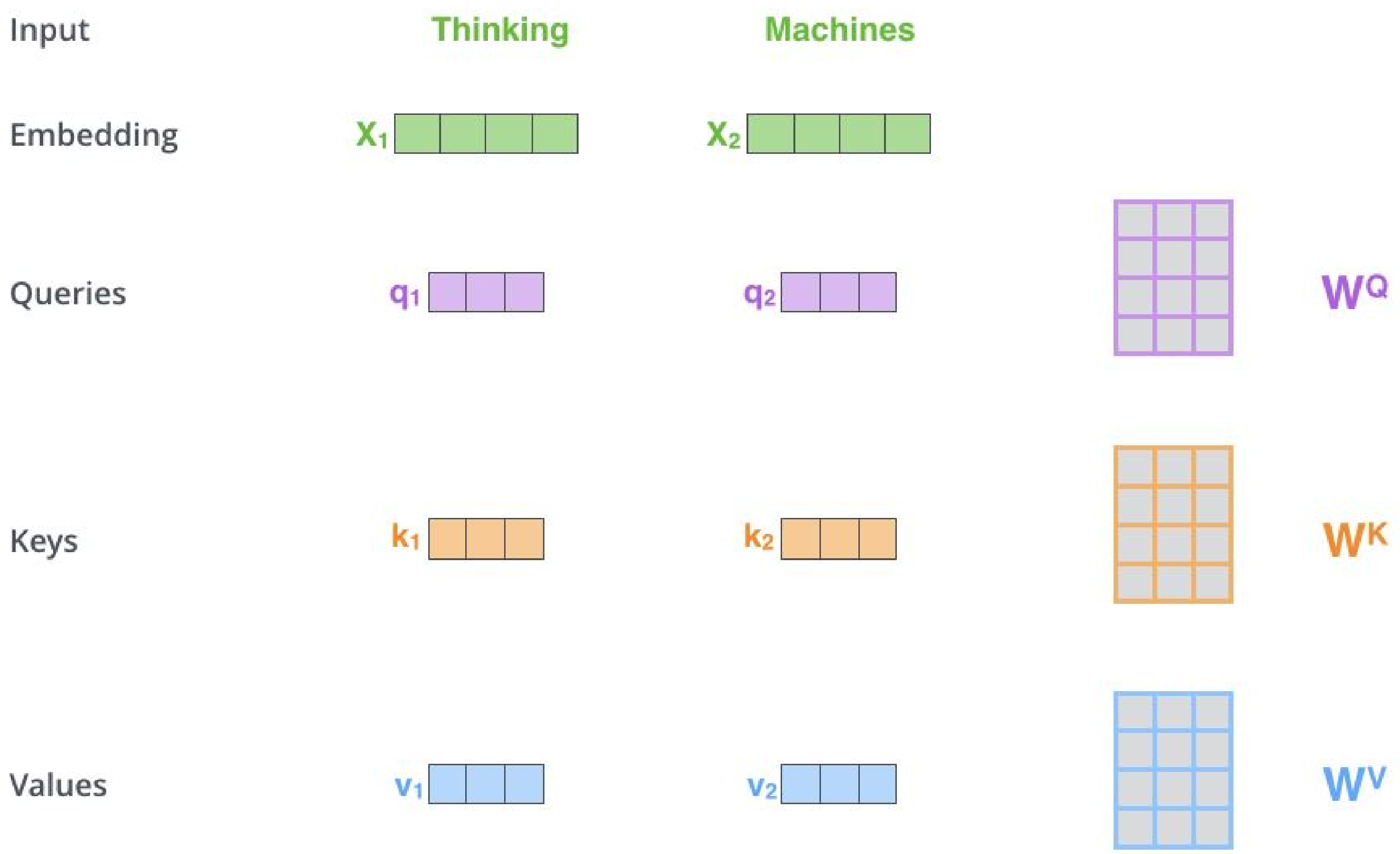
- Background
- Method
- Experiments
- Conclusion

Background



Transformer

- Attention
  - Embedding



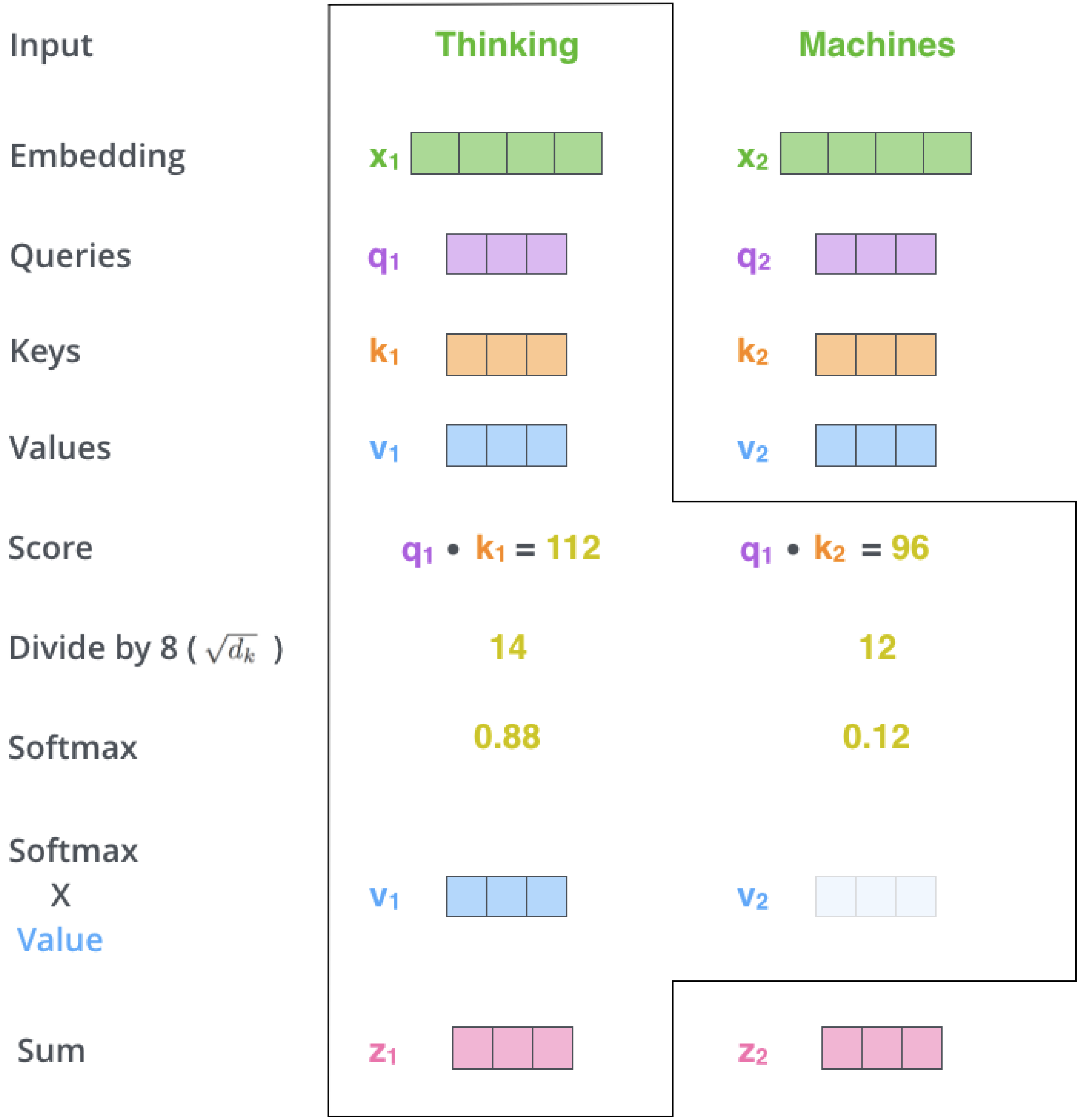
Transformer

- Attention
  - Formulation

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

Transformer

- Attention
  - Pipeline

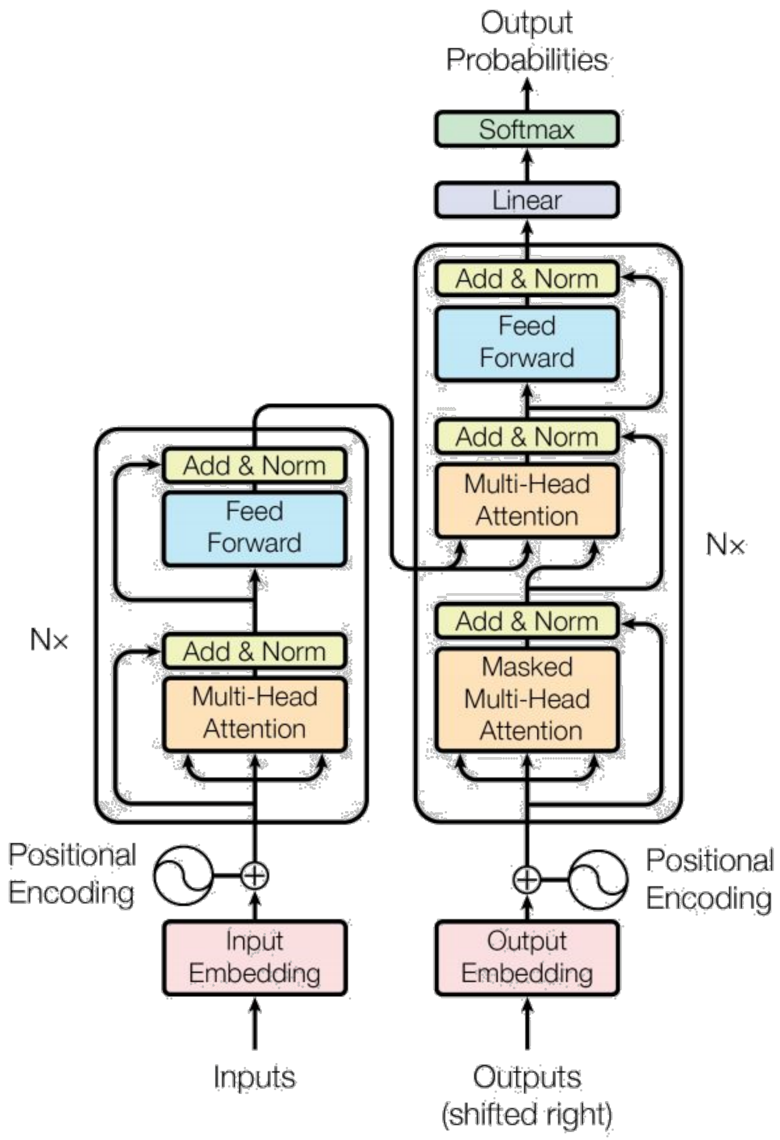


Transformer

- Attention
  - Multi-head Attention
    - Concat results of multiple attention module
    - FC to generate final result
  - Positional Embedding
    - Embed position information (usually sine)
    - Embedding = Word Embedding + Positional Embedding

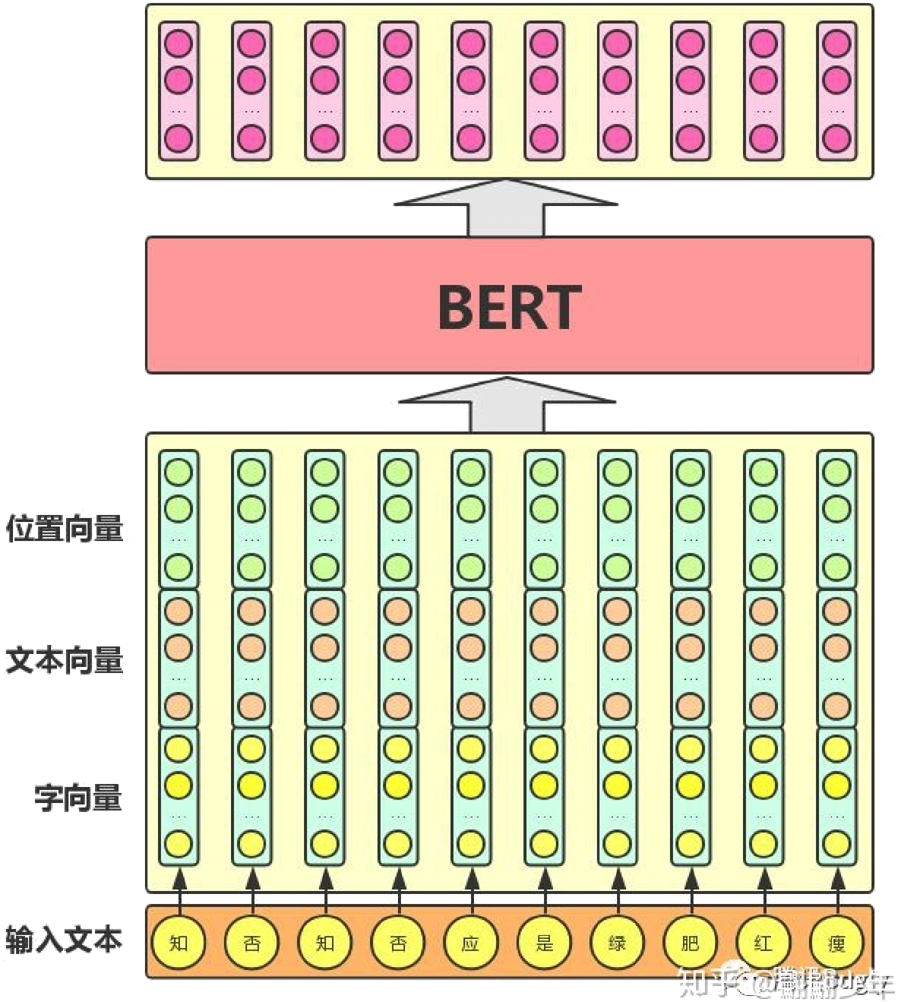
Transformer

- Structure



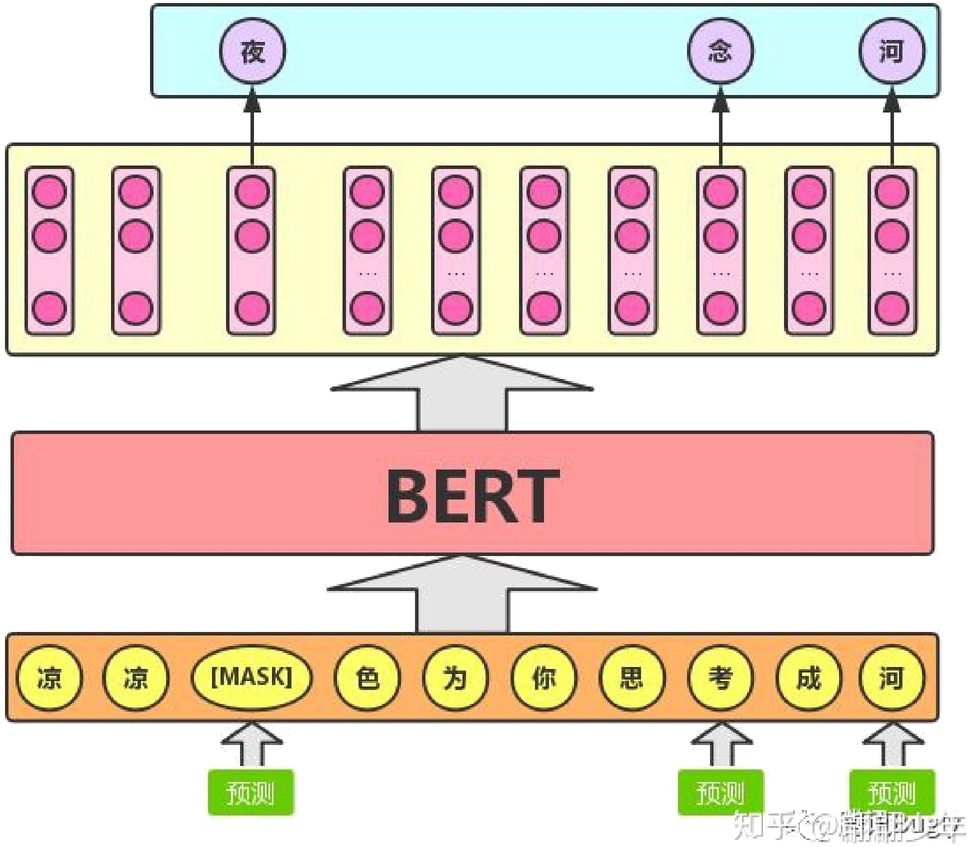
BERT

- Find a better Representation (Embedding)



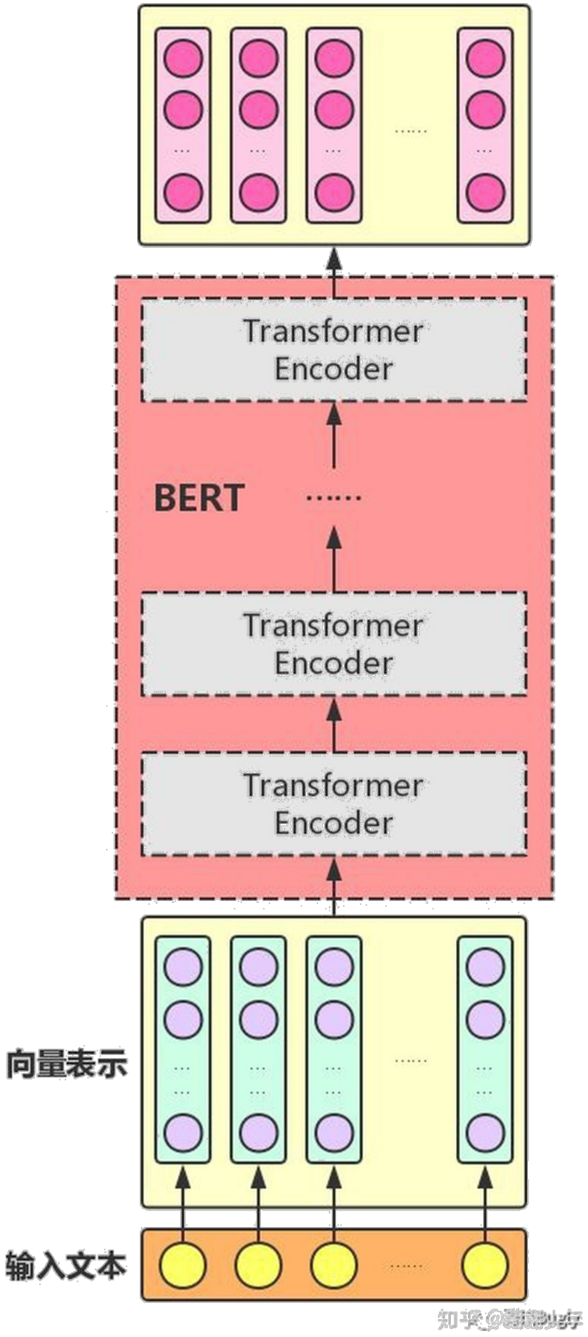
BERT

- Unsupervised Pretraining
  - Task: Masked Language Model



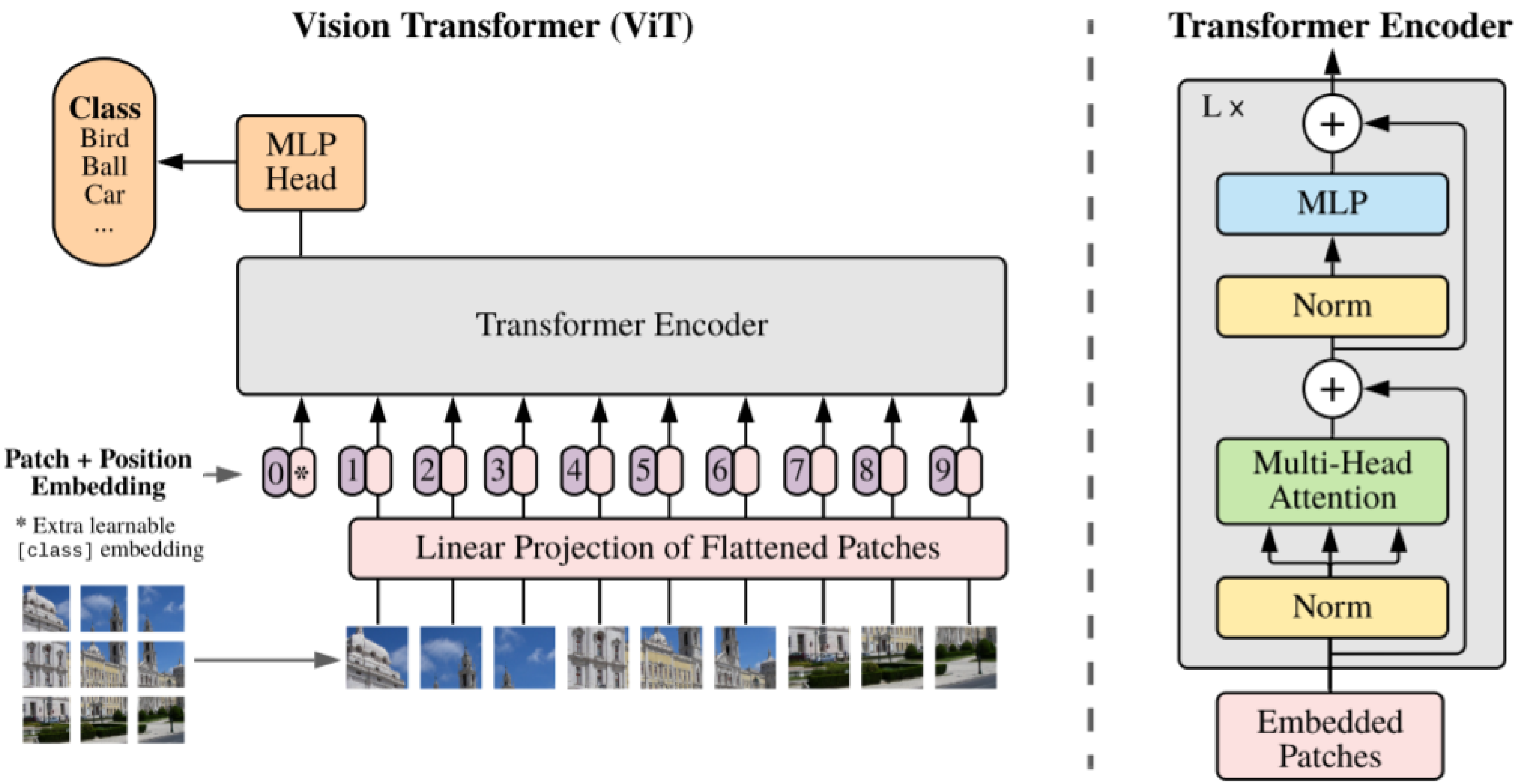
BERT

- Structure
  - Sequential Transformer *Encoder*
    - *That's why Bi-directional*



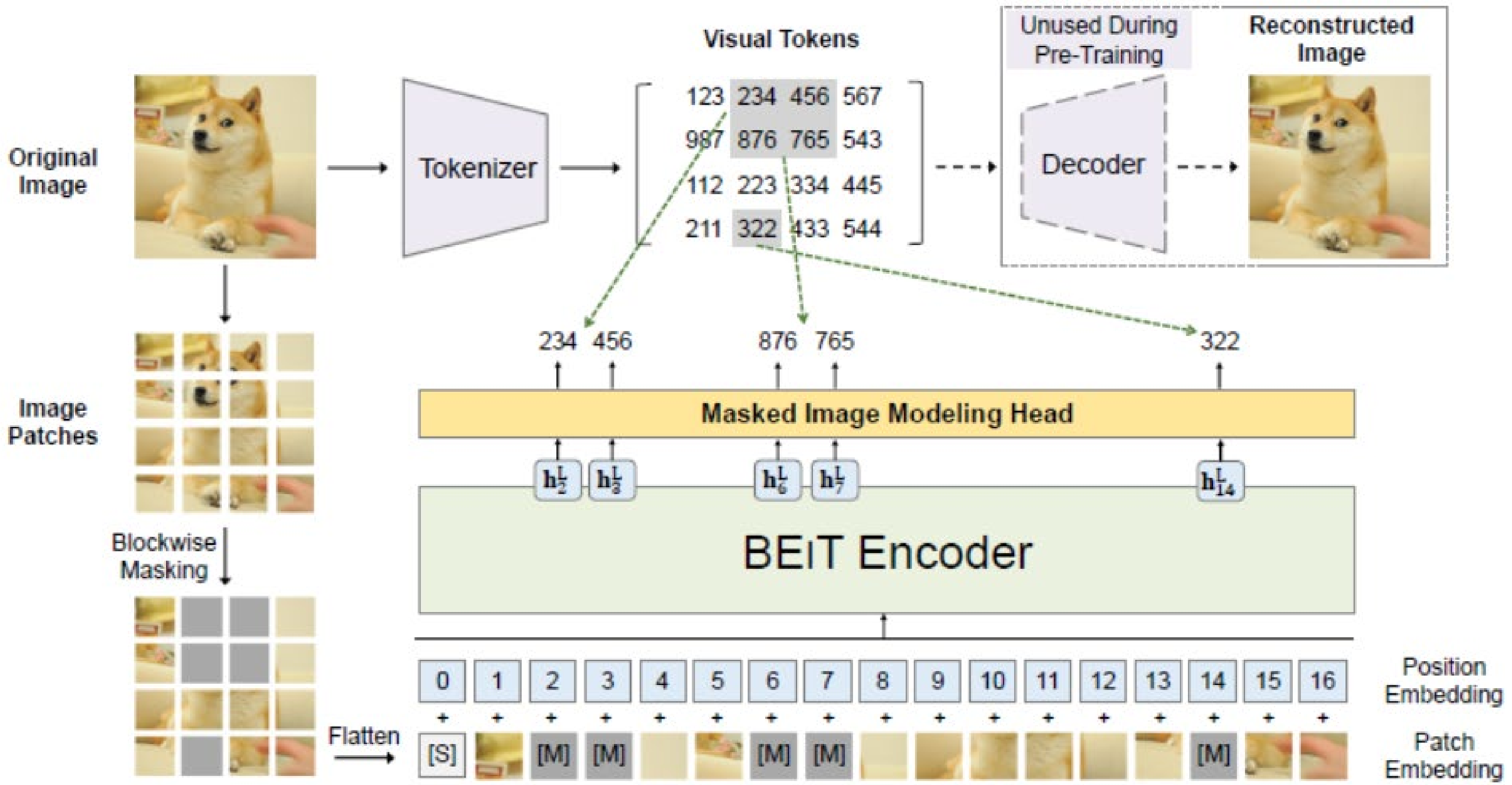
ViT

- Structure



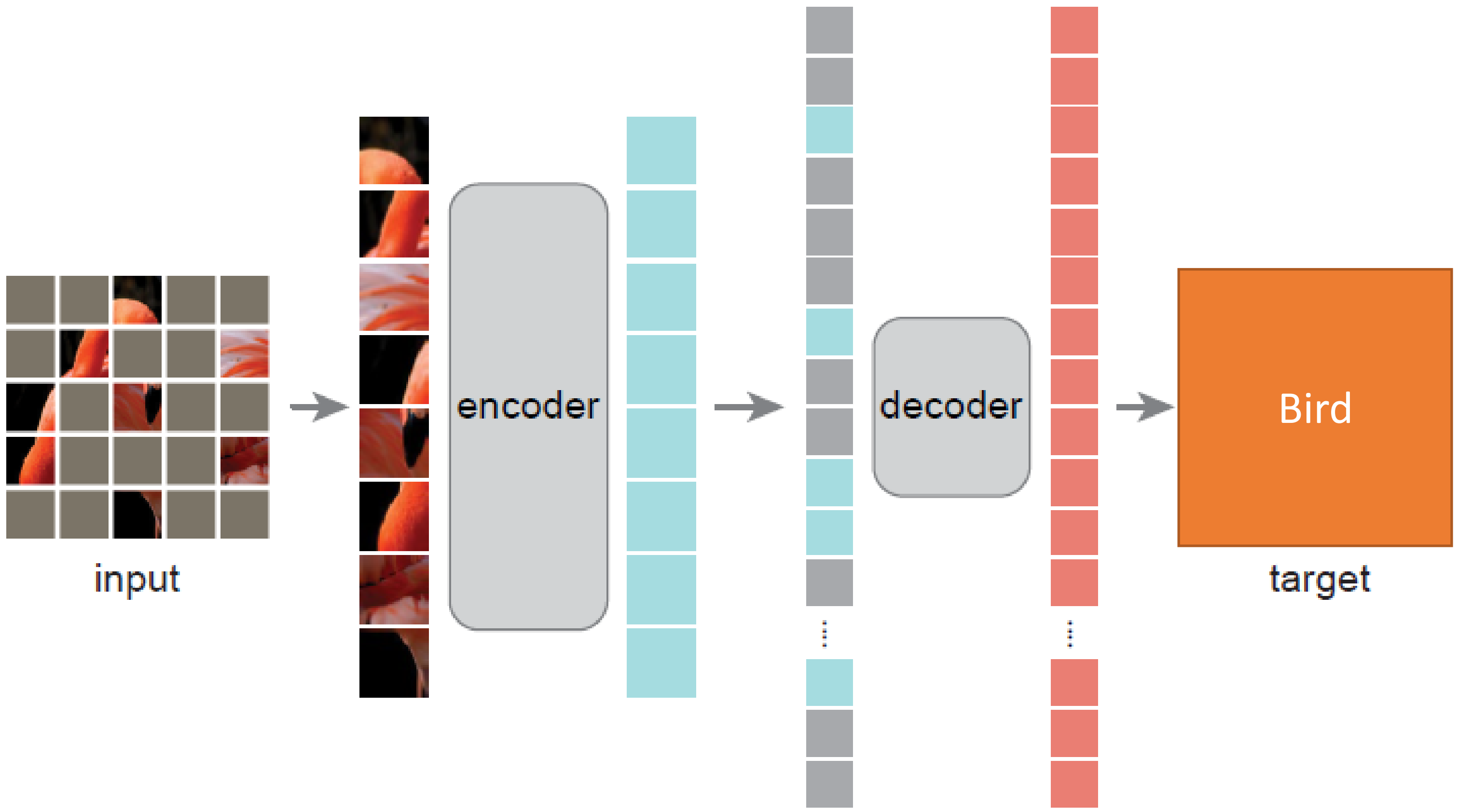
BEiT

- Structure



MAE

- Structure



Question

- *What makes masked autoencoding different between vision and language?*
  - Transform Structure matters
  - Information Density Gap
    - BERT: mask few words
    - MAE: drop a lot of patches (~75%)
  - Decoder Design
    - NLP: reconstruct *words* → semantic
      - Decoder can be trivial (like MLP)
    - CV: reconstruct *pixels* → less semantic
      - Decoder is more important

Details

- Encoder
  - Only takes *unmasked* patches
- Decoder
  - Take *all* patches
  - Light-weight (far smaller than encoder)
  - Only used in pretraining stage
- Reconstruction Target
  - MSE on *masked* patches

Visual Results

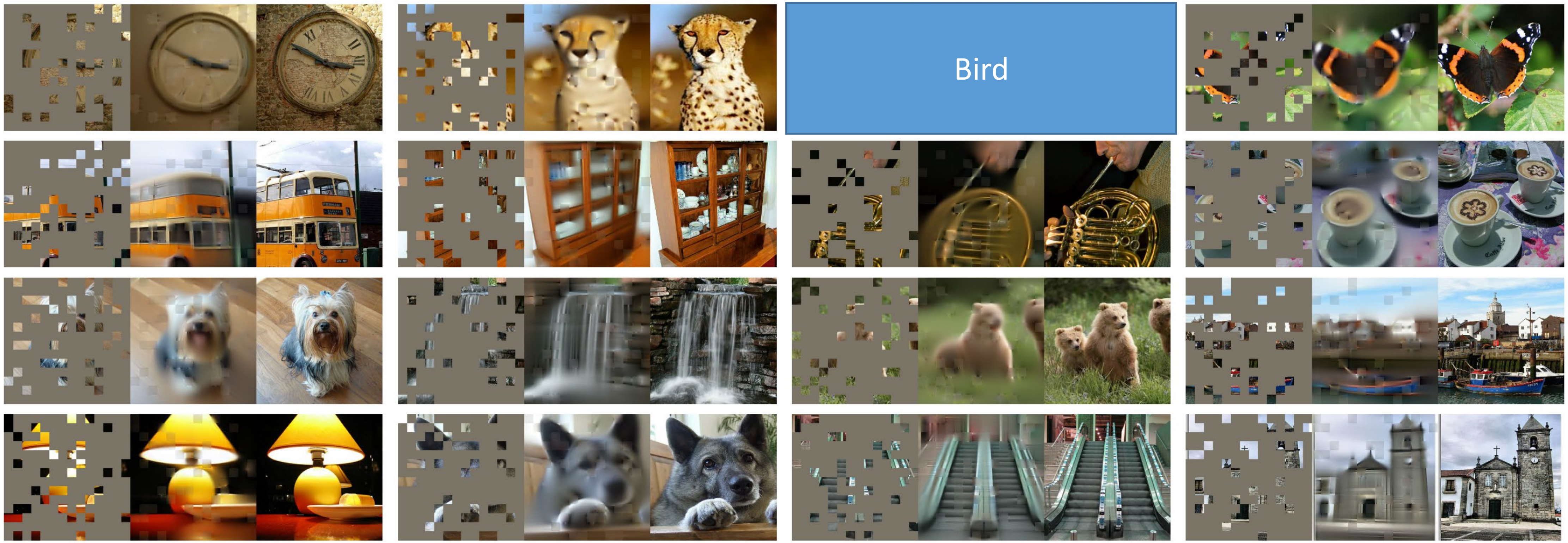


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix. †*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

Quantitive Results

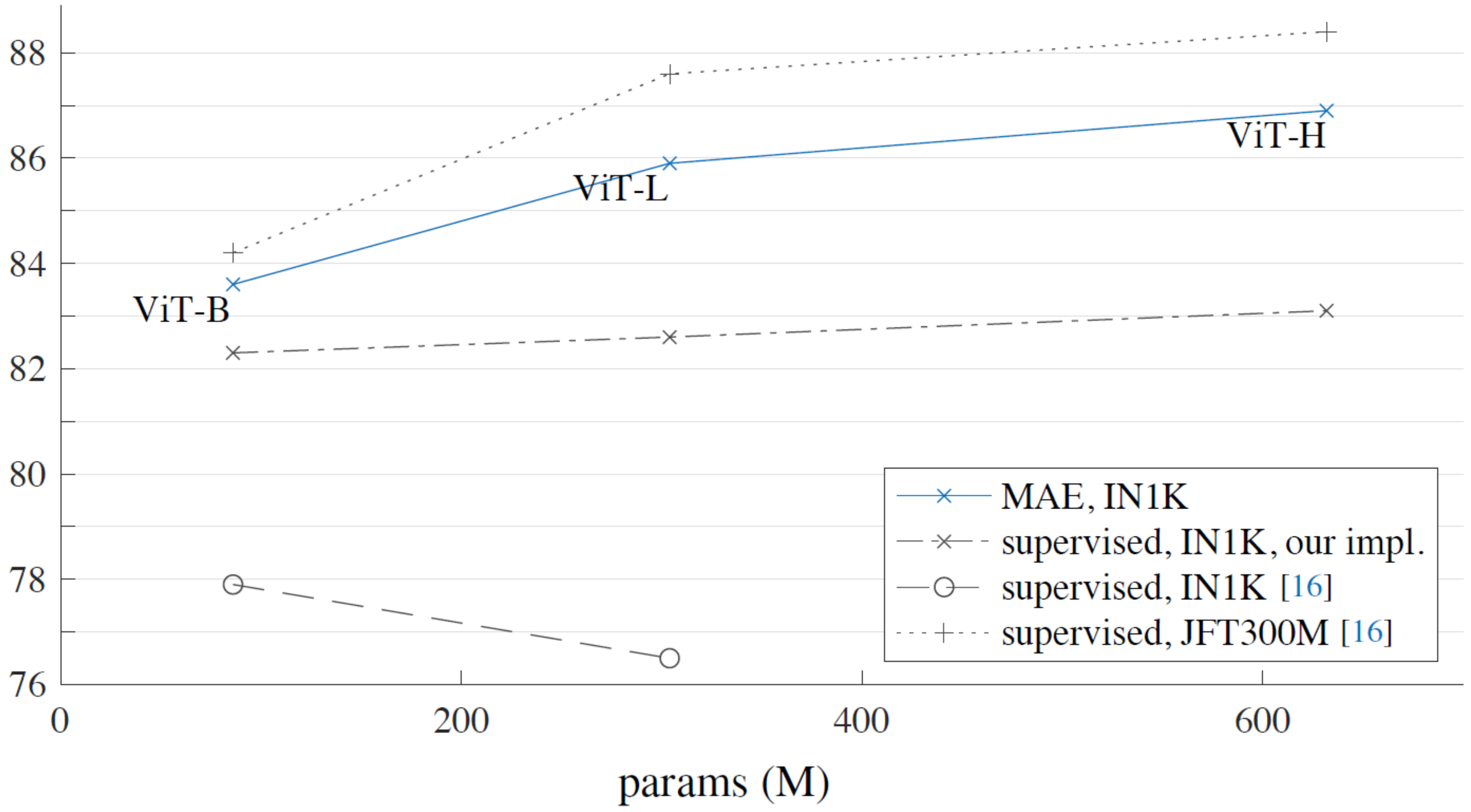
- Compared to Self-supervised

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

Good and Scalable

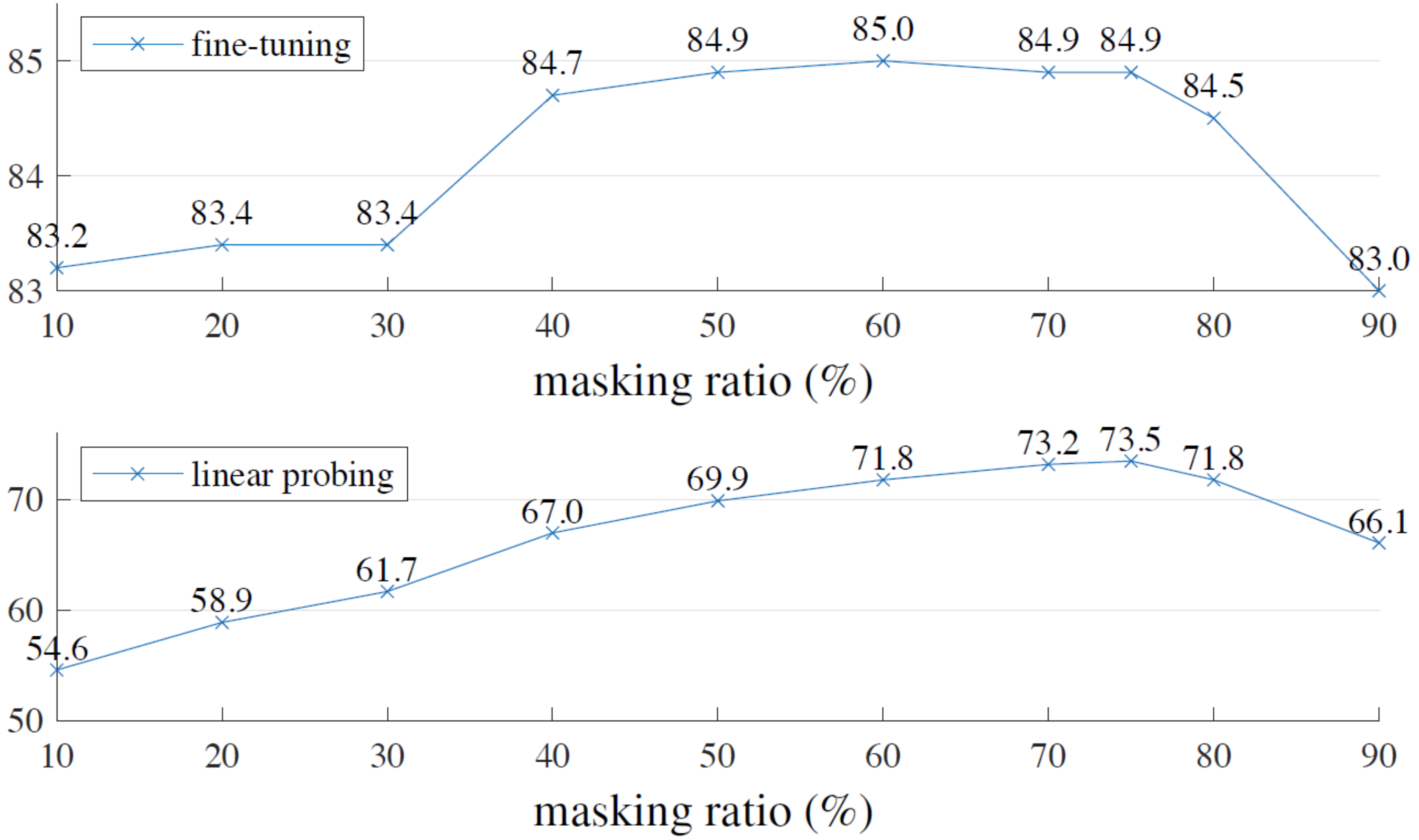
Quantitive Results

- Compared to Supervised



Properties

- Mask Ratio



Ablation Study

blocks	ft	lin
1	84.8	65.5
2	**84.9**	70.0
4	**84.9**	71.9
8	**84.9**	**73.5**
12	84.4	73.3

(a) **Decoder depth**. A deep decoder can improve linear probing accuracy.

dim	ft	lin
128	**84.9**	69.1
256	84.8	71.3
512	**84.9**	**73.5**
768	84.4	73.1
1024	84.3	73.1

(b) **Decoder width**. The decoder can be narrower than the encoder (1024-d).

case	ft	lin	FLOPs
encoder w/ [M]	84.2	59.6	3.3×
encoder w/o [M]	**84.9**	**73.5**	**1**×

(c) **Mask token**. An encoder without mask tokens is more accurate and faster (Table 2).

case	ft	lin
pixel (w/o norm)	84.9	73.5
pixel (w/ norm)	**85.4**	**73.9**
PCA	84.6	72.3
dVAE token	85.3	71.6

(d) **Reconstruction target**. Pixels as reconstruction targets are effective.

case	ft	lin
none	84.0	65.7
crop, fixed size	84.7	73.1
crop, rand size	**84.9**	**73.5**
crop + color jit	84.3	71.9

(e) **Data augmentation**. Our MAE works with minimal or no augmentation.

case	ratio	ft	lin
random	75	**84.9**	**73.5**
block	50	83.9	72.3
block	75	82.8	63.9
grid	75	84.0	66.0

(f) **Mask sampling**. Random sampling works the best. See Figure 6 for visualizations.

Transfer Learning

		AP^{box}		AP^{mask}	
method	pre-train data	ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	**53.3**	44.4	47.1
MAE	IN1K	**50.3**	**53.3**	**44.9**	**47.2**

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.

method	pre-train data	ViT-B	ViT-L
supervised	IN1K w/ labels	47.4	49.9
MoCo v3	IN1K	47.3	49.1
BEiT	IN1K+DALLE	47.1	53.3
MAE	IN1K	**48.1**	**53.6**

Table 5. **ADE20K semantic segmentation** (mIoU) using UperNet. BEiT results are reproduced using the official code. Other entries are based on our implementation. Self-supervised entries use IN1K data *without* labels.

Conclusion

- Simple Masked Autoencoder works
- Rethinking Model or Data
- Effective Training Tricks and Well-organized paper

Thanks

王德昭
wangdz@pku.edu.cn