BEiT: BERT Pre-Training of Image Transformers

A Path to the BERT Moment of CV

Li Dong (董力) on behalf of
Hangbo Bao and Furu Wei
Microsoft Research Asia

2021/11

New AI Paradigm: (Self-supervised) Pre-trained Models

Huge amount of **unlabeled** data

Self-supervised Pre-training

→

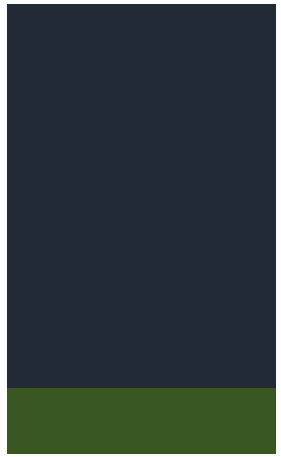
Pre-trained Model

Examples: BERT, GPT-3, Alexander/Turing models

Supervised Fine-tuning

→

Towards few/zero-shot or prompt tuning



Limited (or even no) **labeled** data per task

Downstream AI Tasks:

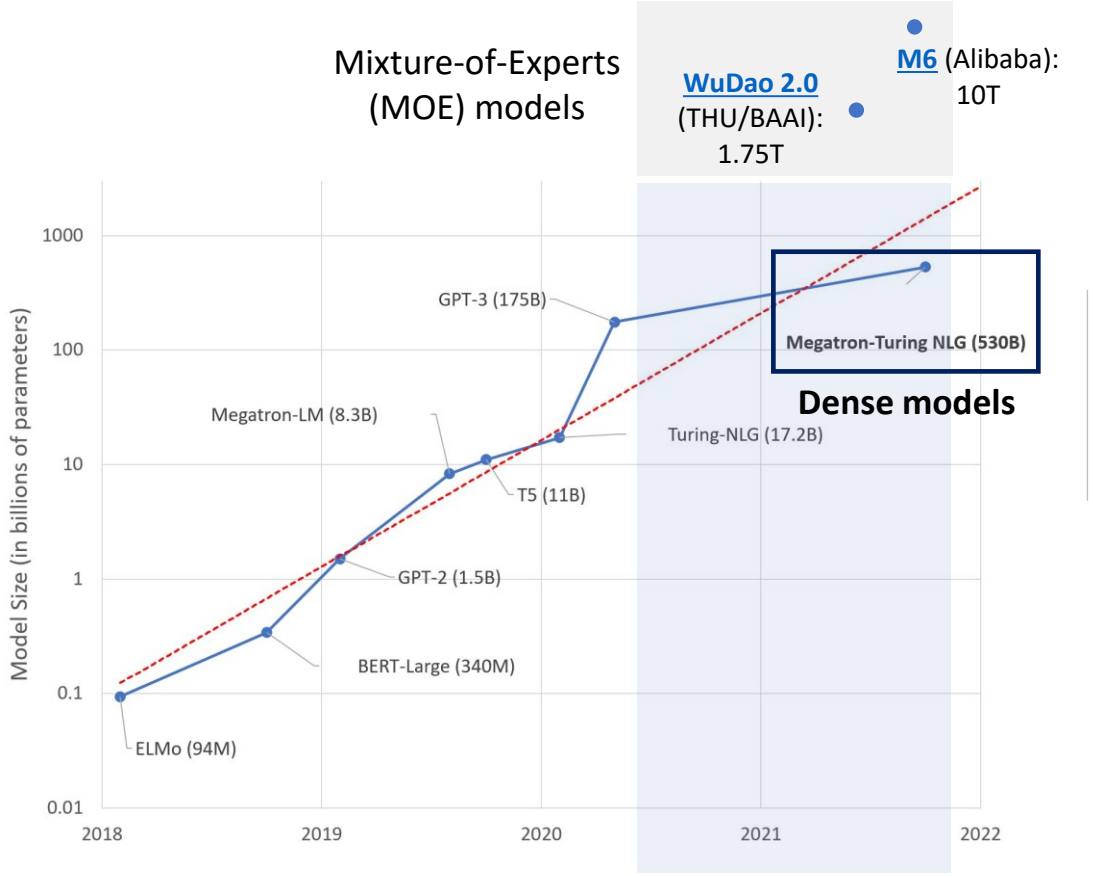
- Language understanding and generation
- Machine translation
- Speech recognition
- Image classification, Object detection
- …

Pre-trained Model Task-specific Model (Layer)

Spanning and converging across different AI areas including NLP, CV, Speech, Multimodal, …

The keywords in the past year (2020 –> 2021)

BIGGER



CONVERGENCE

From **language**, to **vision**, **audio**, and **multimodal** (e.g., vision + language, layout/format + language, etc.)

- **Transformers** becomes the *de facto* backbone networks across AI areas like NLP, CV, Speech, …
- **Self-supervised pre-training tasks** are converging across different modalities
  - **Generative**: language to vision, audio
  - **Contrastive**: vision to language (multilingual)

Key milestone: **convergence of NLP and CV**

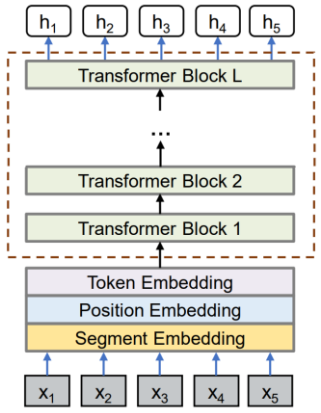
Translate success from language to vision

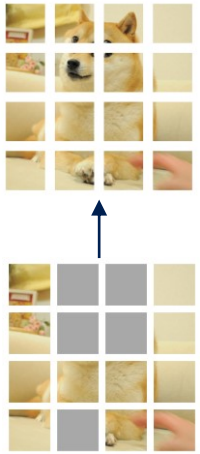
	NLP	CV
Backbone Networks	**Transformer** as the dominant backbone network architecture	**Vision Transformers**
Self-supervised Pre-training Methods/Tasks	**Generative self-supervised pre-training tasks** (e.g., **BERT - Masked Language Modeling**) play the most important role	**BEiT - Masked Image Modeling**

BEiT: BERT Pre-Training of Image Transformers

Data + Model + Task = Pre-Training

Images + Vision Transformer + Masked Image Modeling = BEiT



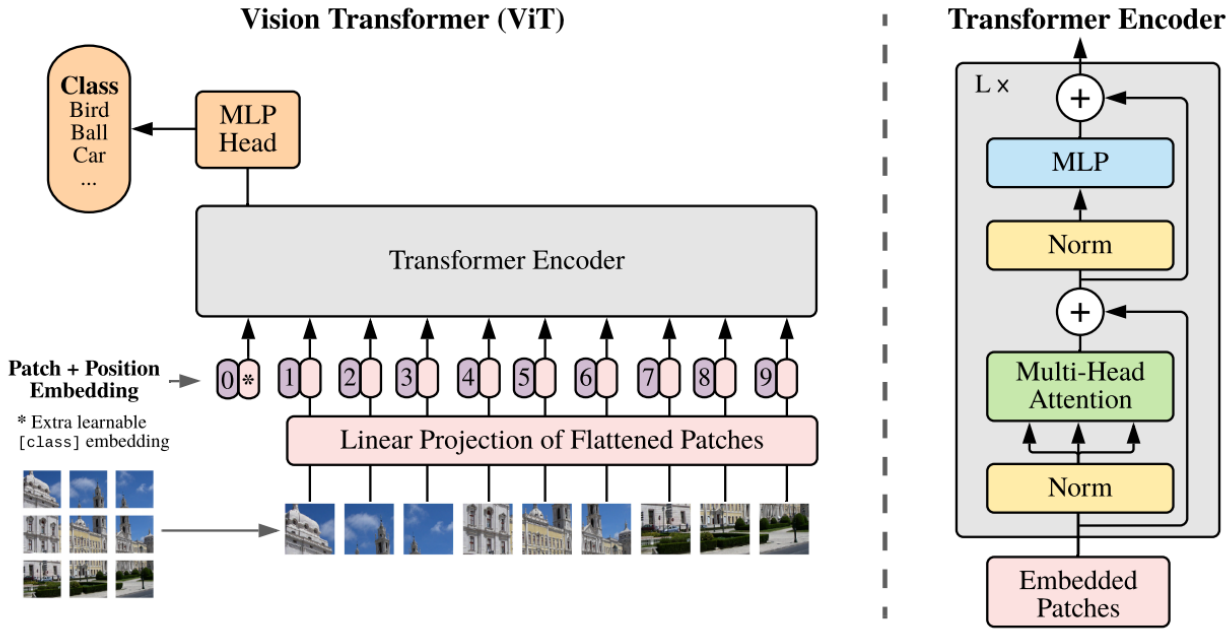


BEiT: BERT Pre-Training of Image Transformers. arXiv 2021.

Vision Transformer (ViT)

Split an image into fixed-size patches, linearly embed each of them, add position embeddings, and **feed the resulting sequence of vectors to a standard Transformer encoder**.

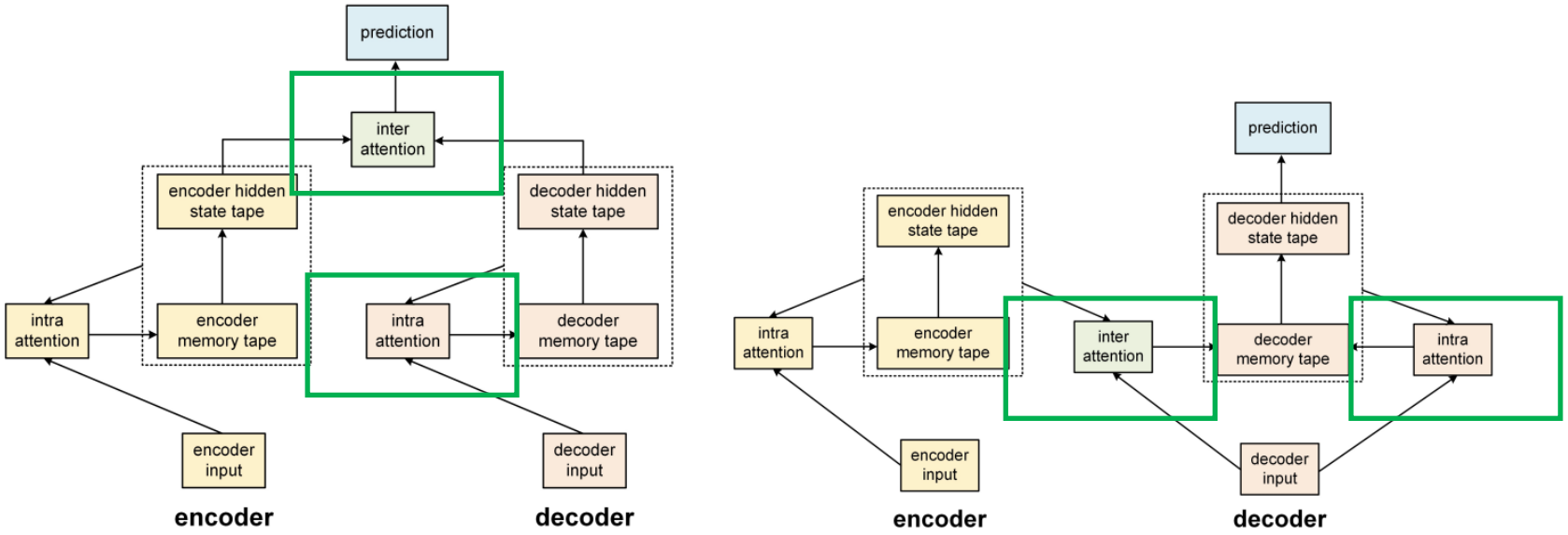
- Two key elements: self-attention, and feed-forward network



Alexey Dosovitskiy et al, (ICLR 2021)

Vision Transformer (ViT)

Self-attention: message passing



(a) Decoder with shallow attention fusion.

(b) Decoder with deep attention fusion.

Proposed in “Long Short-Term Memory-Networks for Machine Reading. Jianpeng Cheng, Li Dong, Mirella Lapata. EMNLP 2016.”

Vision Transformer (ViT)

Feed-forward network: neural knowledge database

Knowledge Neurons in Pretrained Transformers. Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Furu Wei. arXiv 2021.

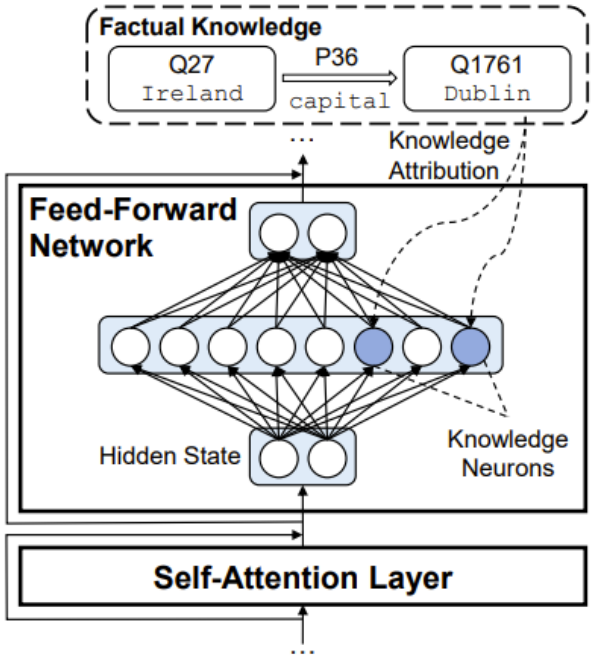


Figure 1: Through knowledge attribution, we identify knowledge neurons that express a relational fact.

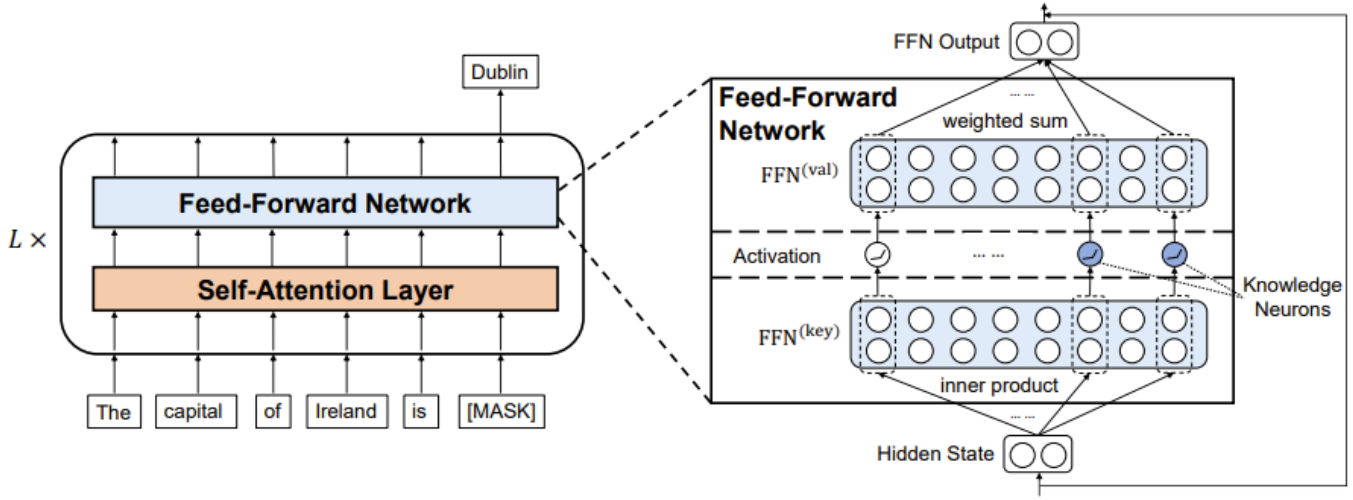
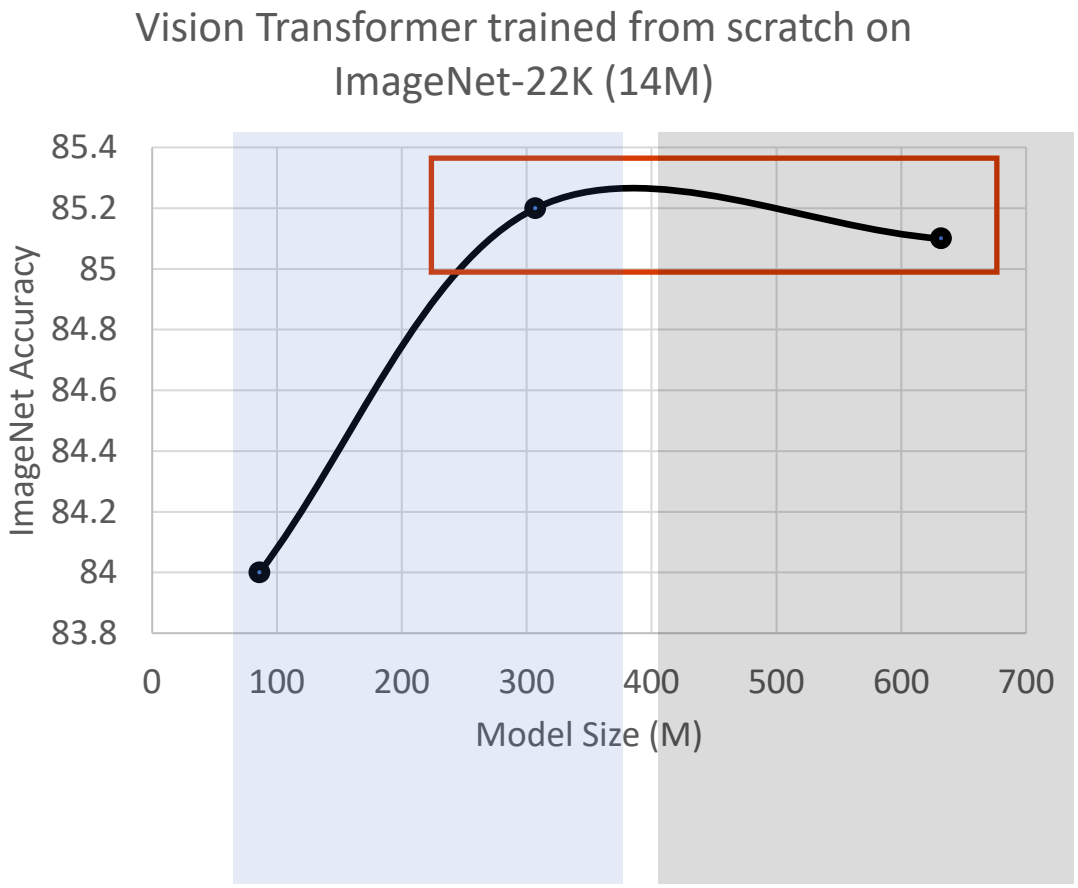


Figure 2: Illustration of how an FFN module in a Transformer block works as a key-value memory. The first linear layer $\mathrm{FFN}^{(\mathrm{key})}$ computes intermediate neurons through inner product. Taking the activation of these neurons as weights, the second linear layer $\mathrm{FFN}^{(\mathrm{val})}$ integrates value vectors through weighted sum. We hypothesize that knowledge neurons in the FFN module are responsible for expressing factual knowledge.

“Annotation hunger” in supervised pre-training in CV

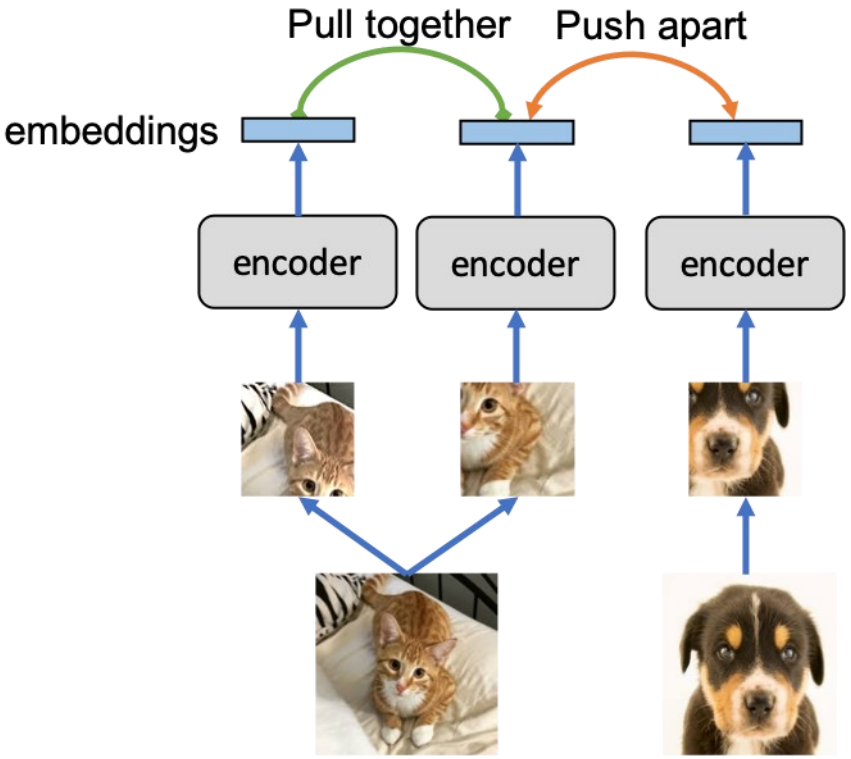


Data > Model: increasing model size brings significant improvements

Data < Model: increasing model size results in performance drop (**overfitting**)

Restricted by the annotation bottleneck even with larger models

Contrastive self-supervised pre-training in CV



Data augmentation

SimCLR (Chen et al., 2020), MoCo (He et al., 2020), DINO (Caron et al., 2021)

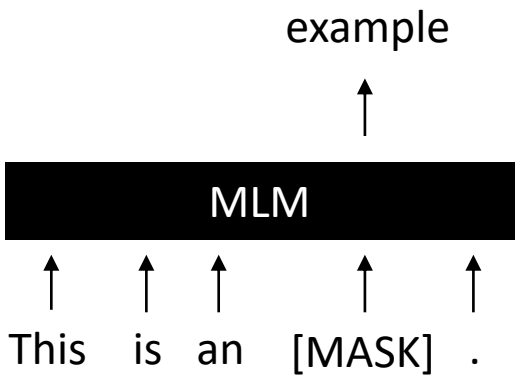
Self-supervised pre-training task:
similar or dissimilar ?

- **Heavily rely on data augmentation**
- **Scaling up issues (memory/computation inefficient)**
  - Multi-pass encoding
  - Large batch size
- **Performance saturation**

Training length. In the following table we report ViT-S/B + MoCo v3 *vs.* training length:

	300-ep	600-ep
ViT-S/16	72.5	73.4
ViT-B/16	76.5	76.7

The new paradigm: **generative** self-supervised pre-training for CV (inspired by language model pre-training)



BERT Masked Language Modeling

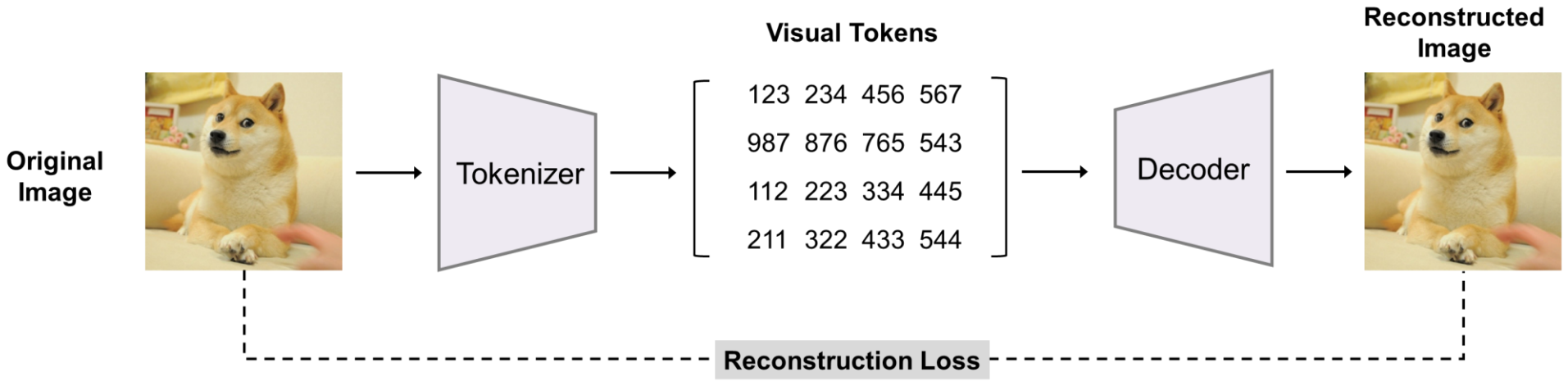
Masked **Image** Modeling ?

BEiT

Visual Tokens

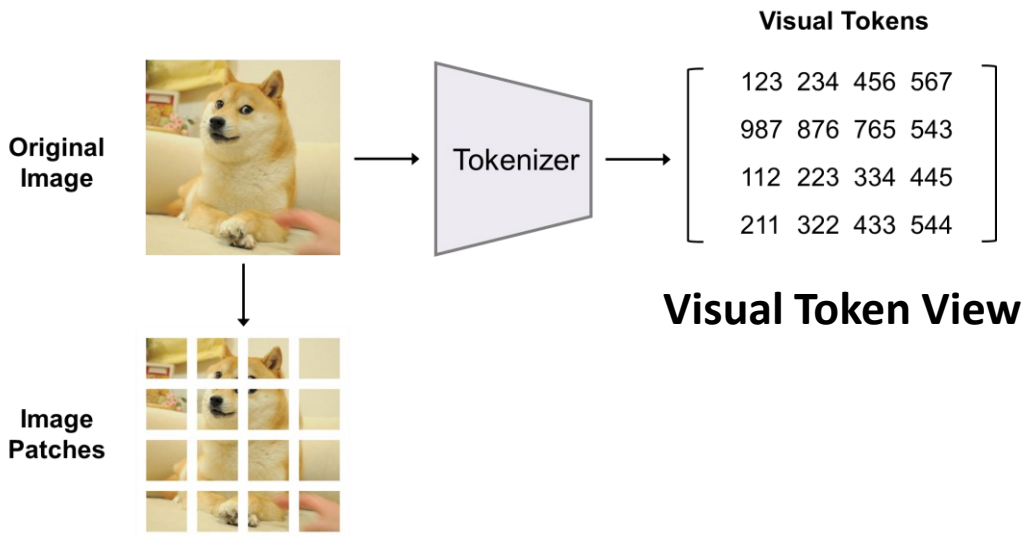
Tokenize images to discrete tokens

- Discrete variational autoencoder (dVAE; Ramesh et al., 2021)
- Learn to reconstruct the original image by conditioning on visual tokens



BEiT Pre-training: Masked Image Modeling

- Image Representations (Two Views)



Visual Token View

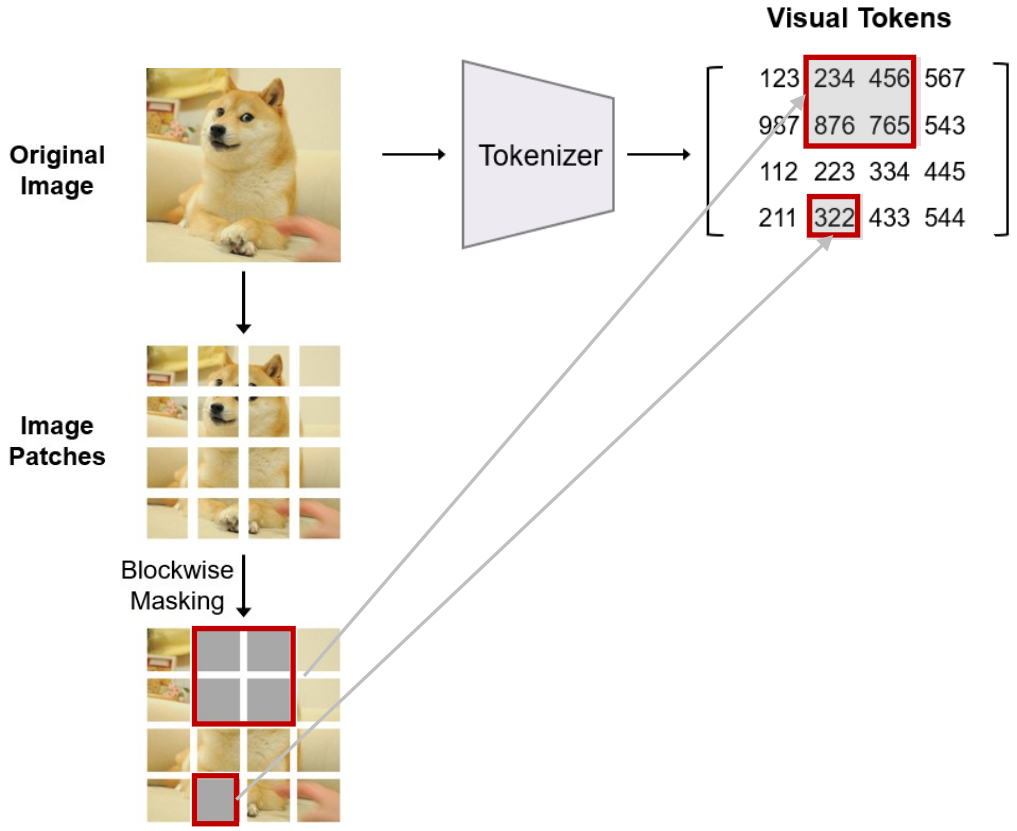
Only used in **pre-training stage** to design pre-training tasks/objectives

Image Patch View

Used in both **pre-training and finetuning** (namely in **downstream tasks** as in existing CV models)

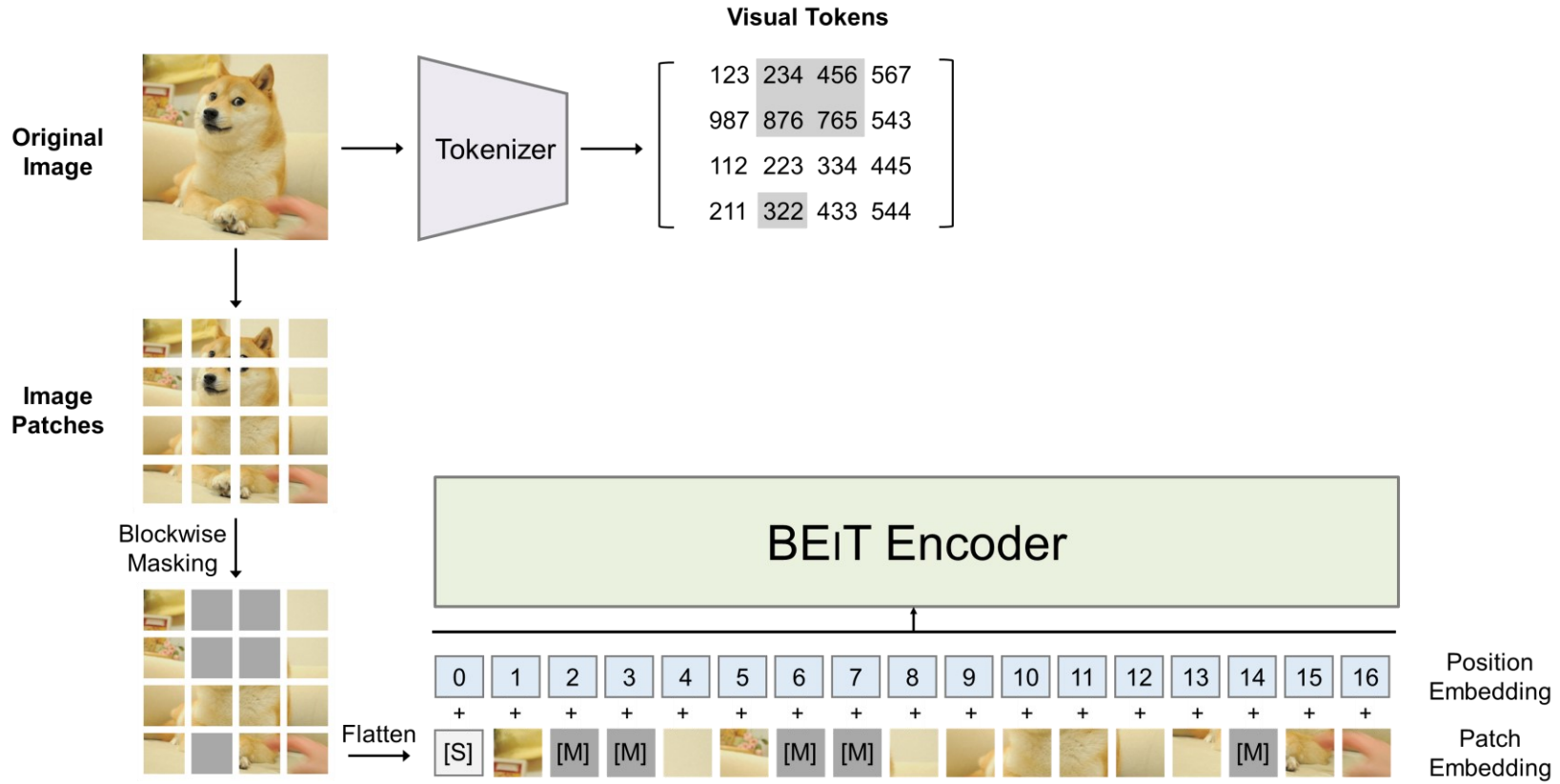
BEiT Pre-training: Masked Image Modeling

- Blockwise Masking
  - A block of image patches is masked each time



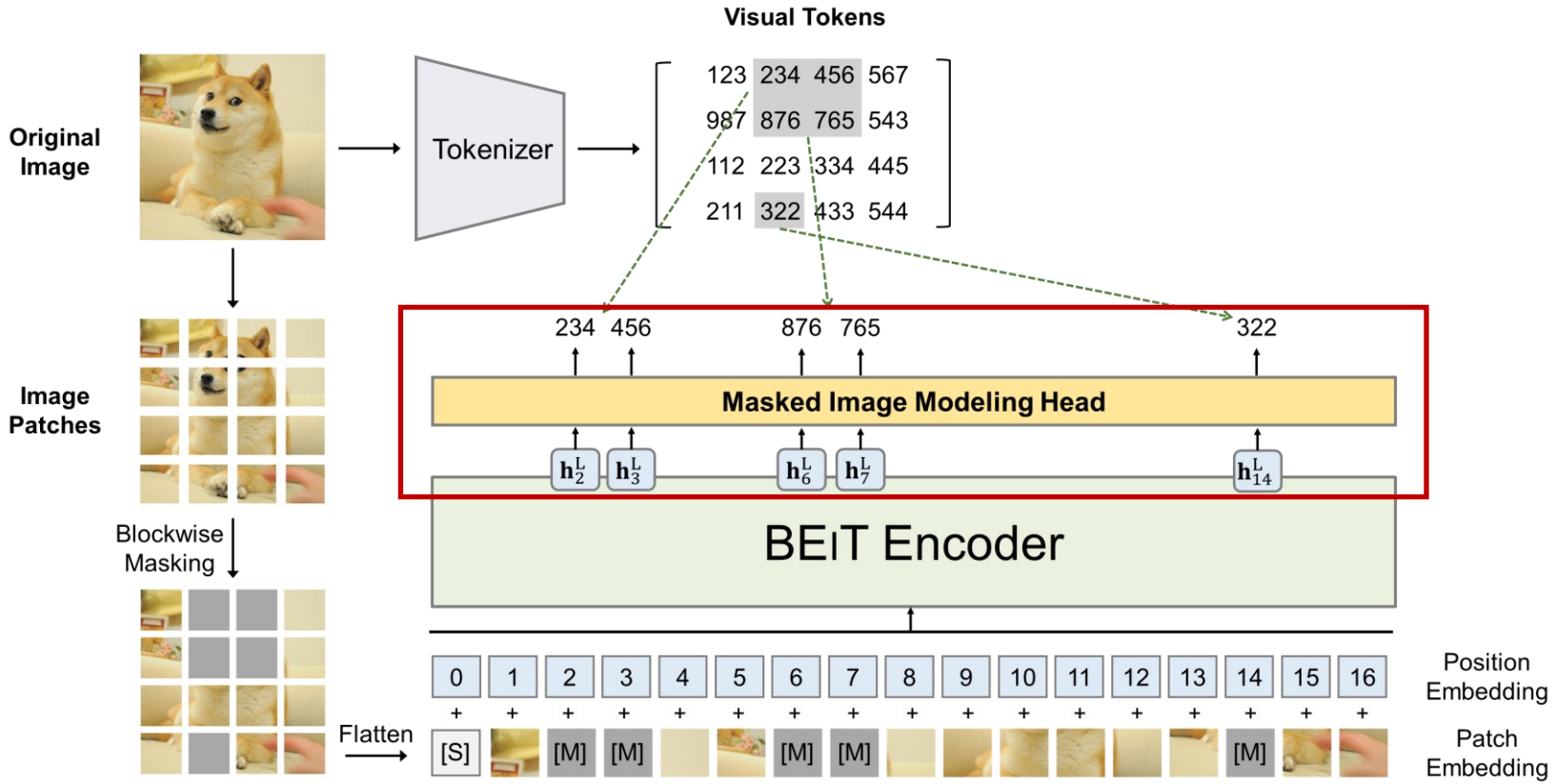
BEiT Pre-training: Masked Image Modeling

- Feed corrupted image patches into Transformer
  - Final hidden vectors are regarded as encoded representations



BEiT Pre-training: Masked Image Modeling

- Recover correct visual tokens given the corrupted image
  - Visual tokens summarize the details to high-level abstractions



From Perspective of Variational Autoencoder (VAE)

- Notations
  - Original image x
  - Corrupted image $\tilde{x}$
  - Visual tokens z
- Consider the evidence lower bound (ELBO) of $p(x|\tilde{x})$

Image Reconstruction

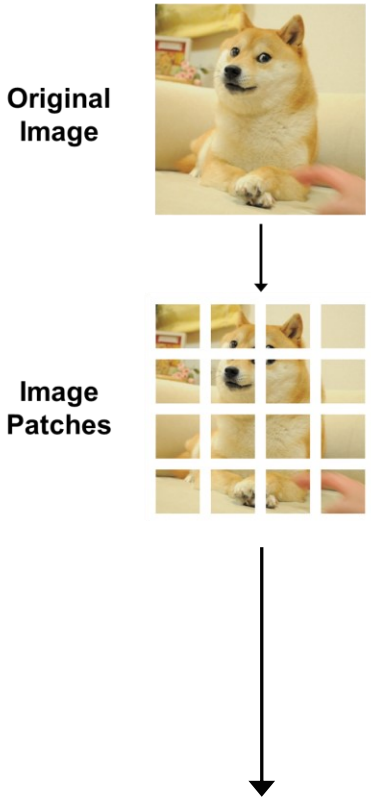
Image to Visual Tokens

Masked Image Modeling

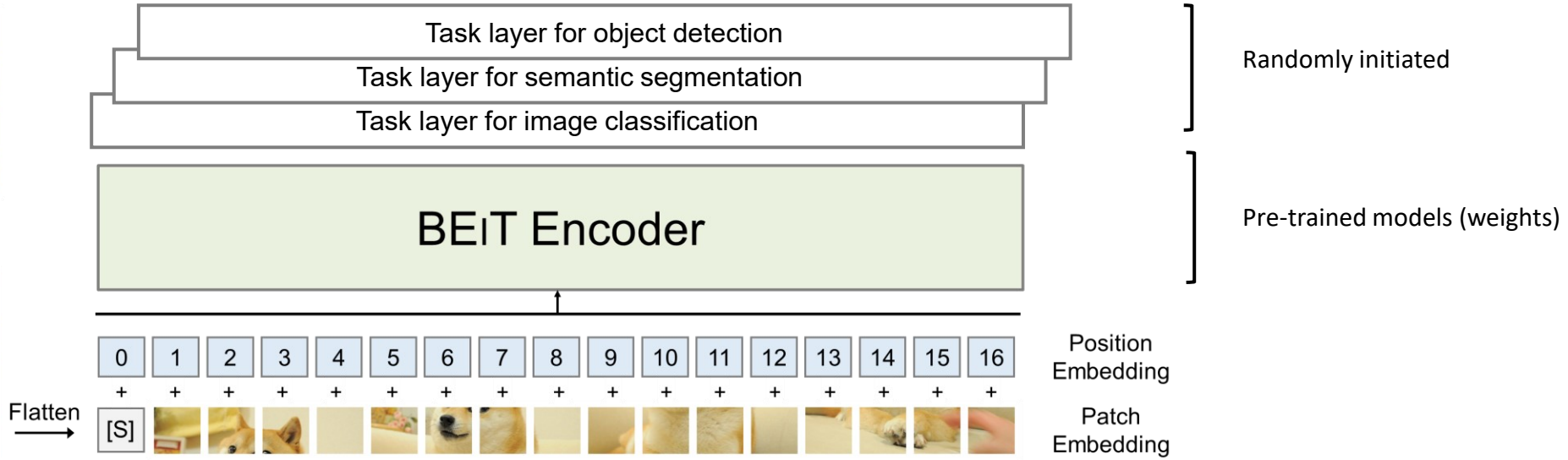
$$\sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \log p(x_i|\tilde{x}_i) \geq \sum_{(x_i,\tilde{x}_i)\in\mathcal{D}} \Big(\underbrace{\mathbb{E}_{z_i\sim q_\phi(\mathbf{z}|x_i)}[\log p_\psi(x_i|z_i)]}_{\text{Visual Token Reconstruction}} - D_{\mathrm{KL}}[q_\phi(\mathbf{z}|x_i), p_\theta(\mathbf{z}|\tilde{x}_i)] \Big)$$

Visual Tokens to Image

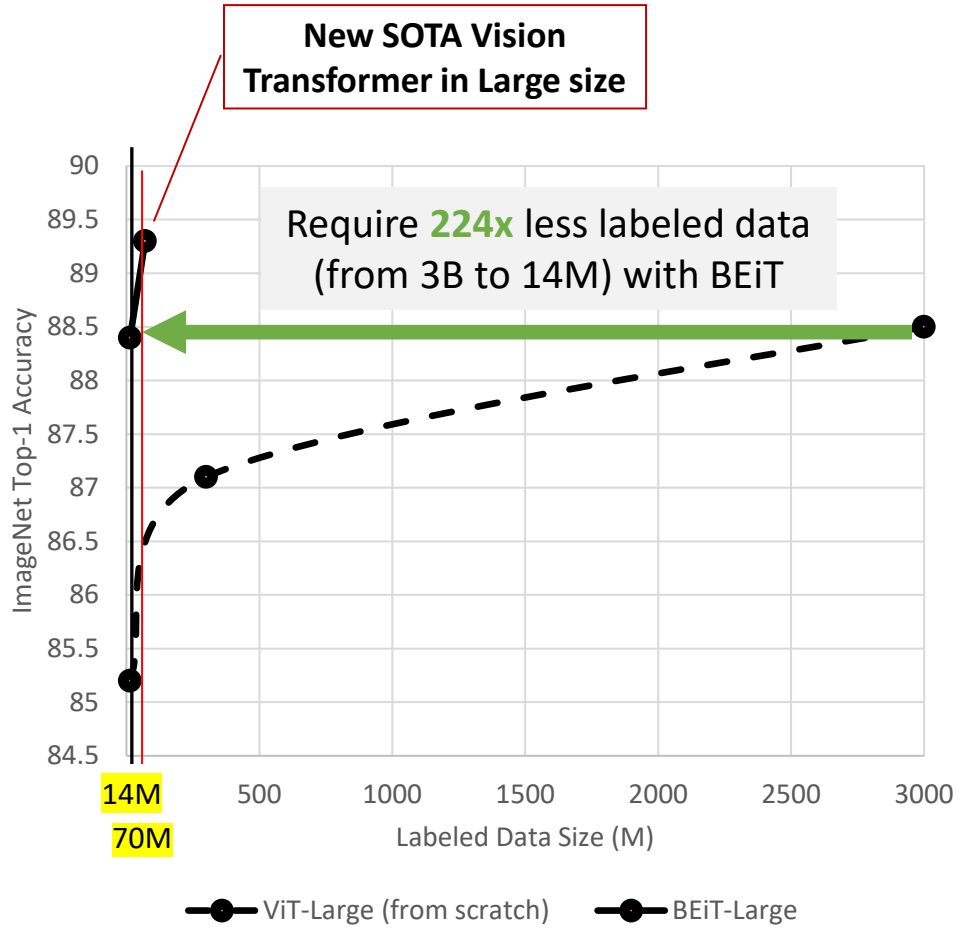
Fine-Tuning BEiT on Downstream Tasks

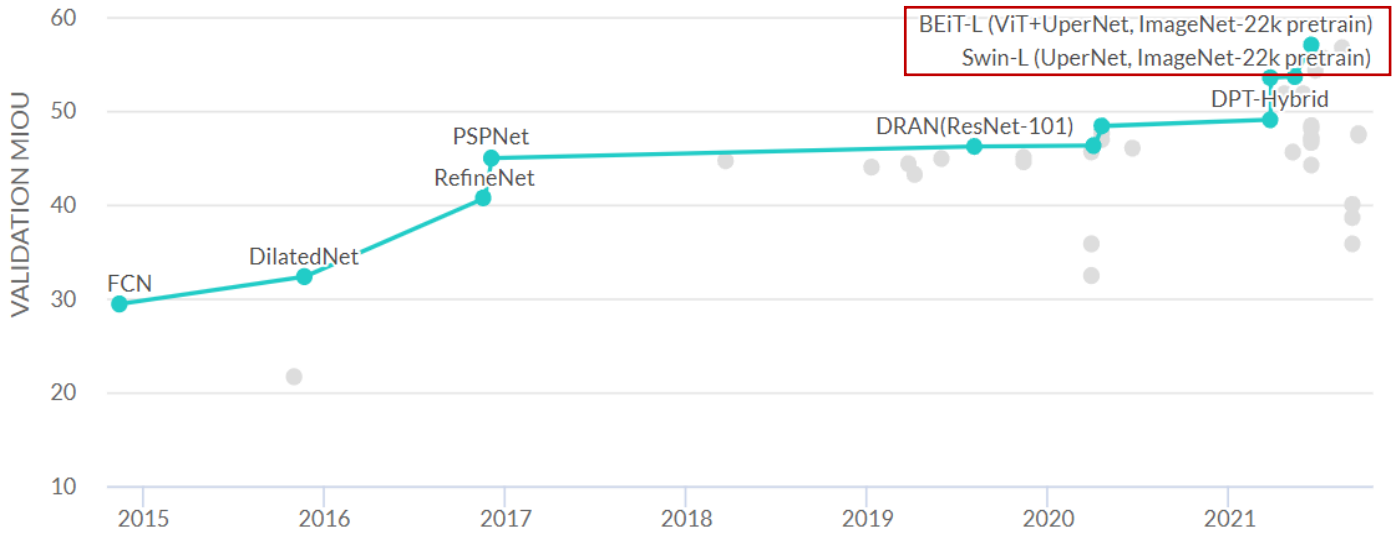


- **Pre-training + Fine-tuning** paradigm
- Append task layers upon pretrained BEiT
  - Image classification: Softmax classifier
  - Semantic segmentation: SETR-PUP (Zheng et al., 2020), or UperNet (Xiao et al., 2019)
  - Other task layers for object detection, etc.



BEiT overcomes annotation hunger for CV

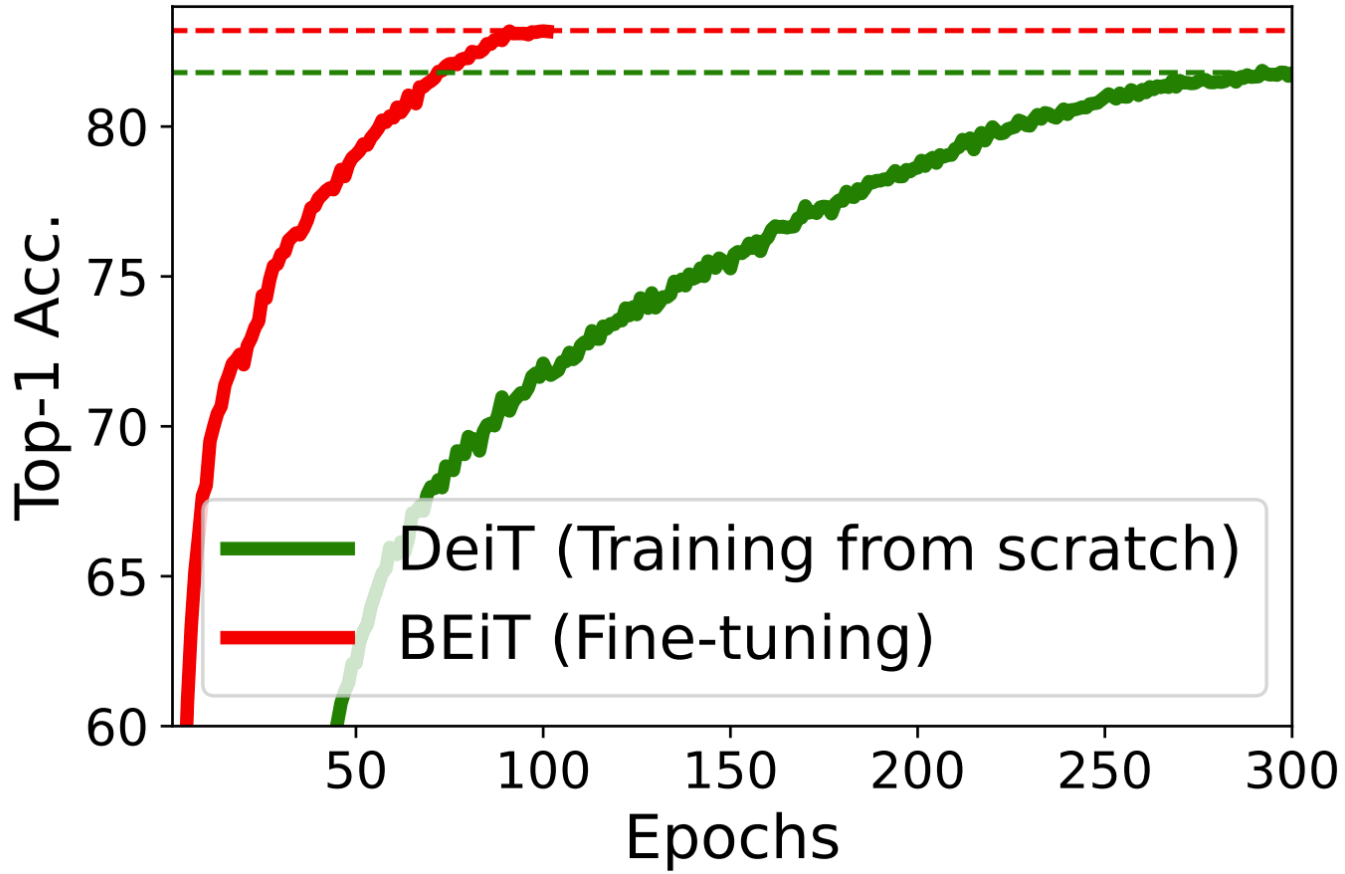




BEiT-L also achieves SOTA results on [semantic segmentation (ADE 20K)]

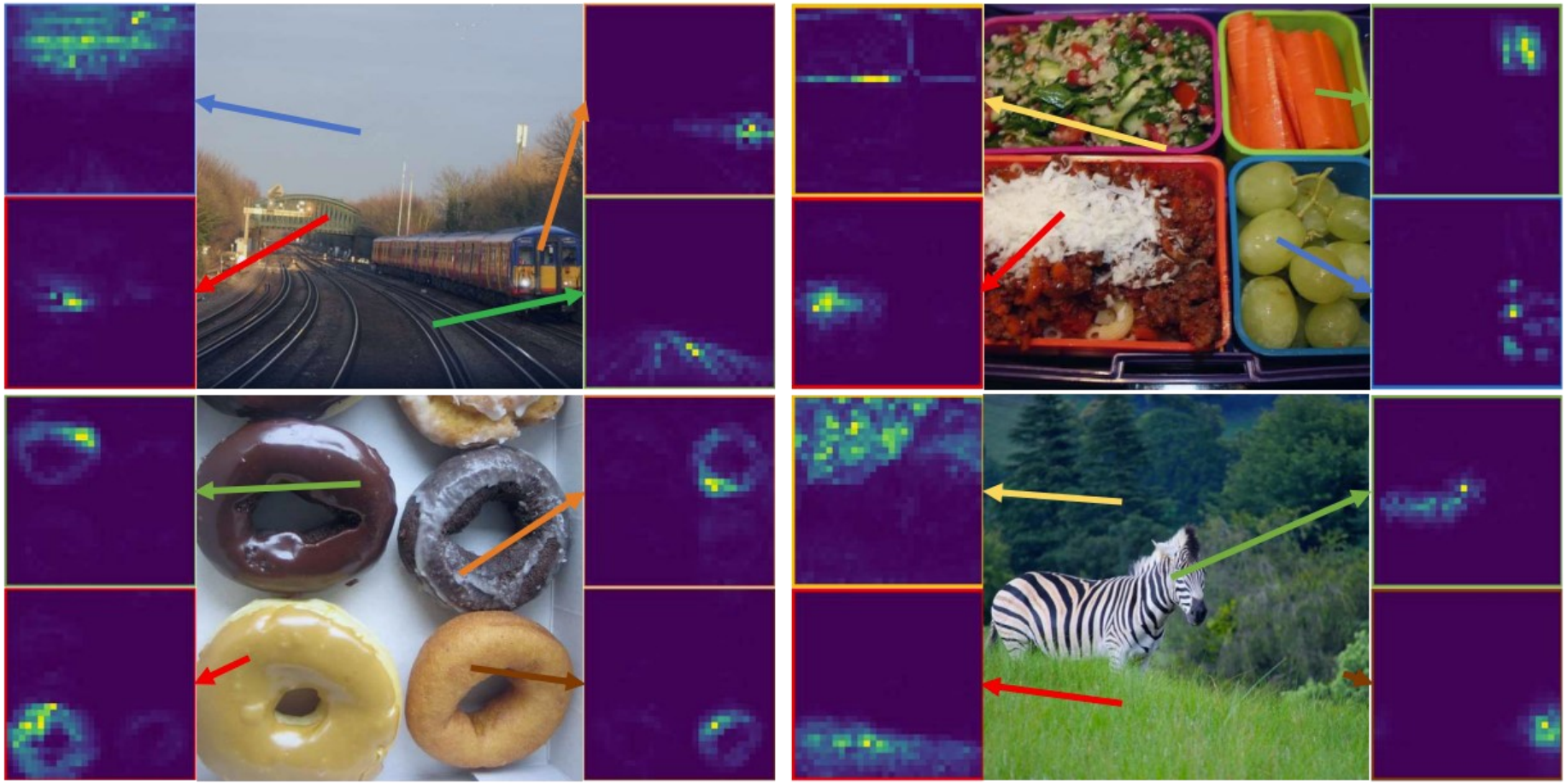
Faster Convergence Speed

BEiT pre-training accelerates fine-tuning convergence



Convergence curves of training DeiT from scratch and fine-tuning BEIT on ImageNet-1K.

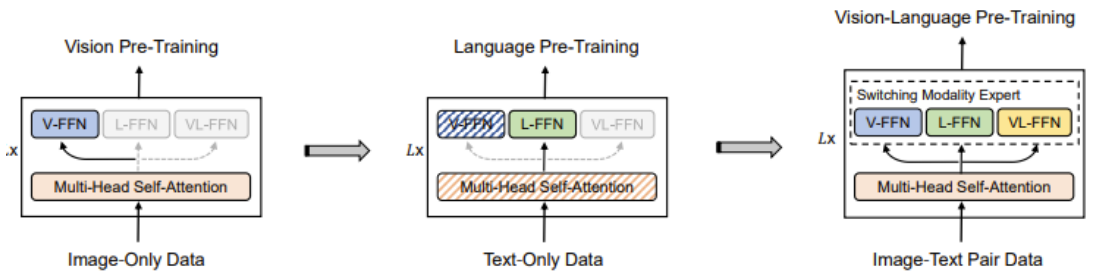
Analysis of Self-Attention Map



Self-attention map for different reference points. The self-attention mechanism in BEiT can separate objects, although self-supervised pre-training does not use manual annotations.

BEiT for Vision-Language Pre-training

- Taking visual question answering (VQA v2.0) as an example
  - VLMo: single large-size model
  - 2nd: ensemble of 48 models
  - 3rd: huge-size model trained on 1.8B image-text pairs



Rank	Participant team	yes/no (↑)	number (↑)	other (↑)	overall (↑)	Last submission at
1	VLMo (Microsoft, Single Large Model)	94.68	67.26	72.87	81.30	7 hours ago
2	Renaissance (AliceMind-MMU)	93.55	72.01	72.67	81.26	3 months ago
3	SimVLM - Google Brain (Single Model)	93.29	66.54	72.23	80.34	3 months ago
4	SFE-NLP (MAP)	92.45	76.57	68.82	79.47	4 months ago
5	UNIMO	93.10	63.06	69.12	78.40	6 months ago
6	ROSITA (ROSITA)	92.66	63.24	69.33	78.34	6 months ago

VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts.
Wenhui Wang, Hangbo Bao#, Li Dong, Furu Wei. arXiv:2111.02358, 2021.

Takeaway Messages

- BEiT translates success from language to vision
  Masked image modeling task
- BEiT achieves strong fine-tuning results on downstream tasks
  Such as image classification, and semantic segmentation
- BEiT is scaling-up-friendly
  Critical for large-scale self-supervised pre-training
- BEiT overcomes performance saturation
  Longer training length consistently improves end-task performance
- BEiT is an important step towards multimodal pre-training
  Greatly reduce the reliance on image-text pairs

Microsoft

Code and pretrained models are available at https://aka.ms/beit.

We are hiring at all levels (including FTE researchers and interns)!
Let's work together to achieve BERT moment for CV.
lidong1@microsoft.com