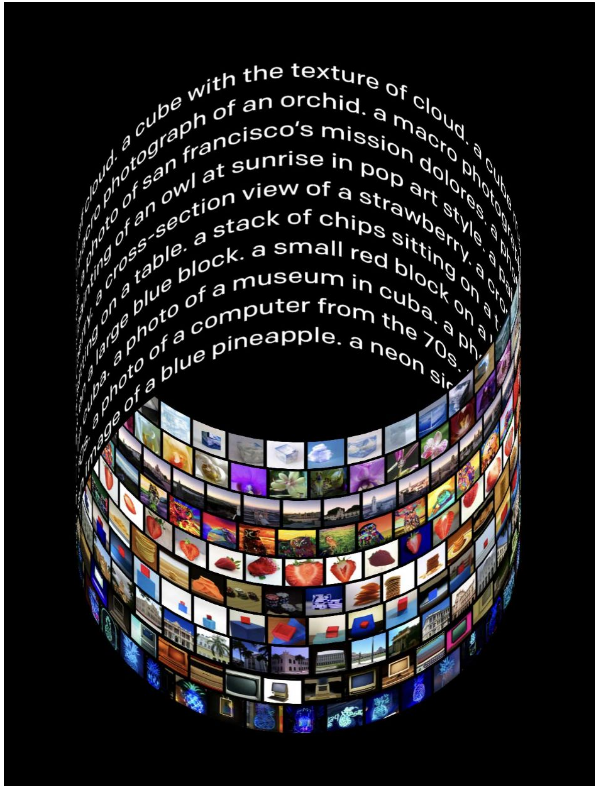
Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks

Kriti Ghosh   Yuchen Zhang

Content

- Background
- Related work & Motivation
- Methods
- Evaluation & Experiments
- Summary



Background

Convergence of language, vision, and multimodal pretraining

Background

- A big convergence of language, vision, and multimodal pretraining is emerging
- By performing large-scale pretraining on massive data, we can easily transfer the models to various downstream tasks
- Try to pretrain a general-purpose foundation model that handles multiple modalities

Related work & Motivation

Transformer
Masked data modeling
Scaling up

Advance the convergence trend for vision-language pretraining

- The success of Transformers on translated from language to vision and multimodal problems.
- Pretraining task based on masked data modeling has applied to various modalities.
- Scaling up the model size and data size improves the generalization quality of foundation models.

Success of Transformer on translated from language to vision

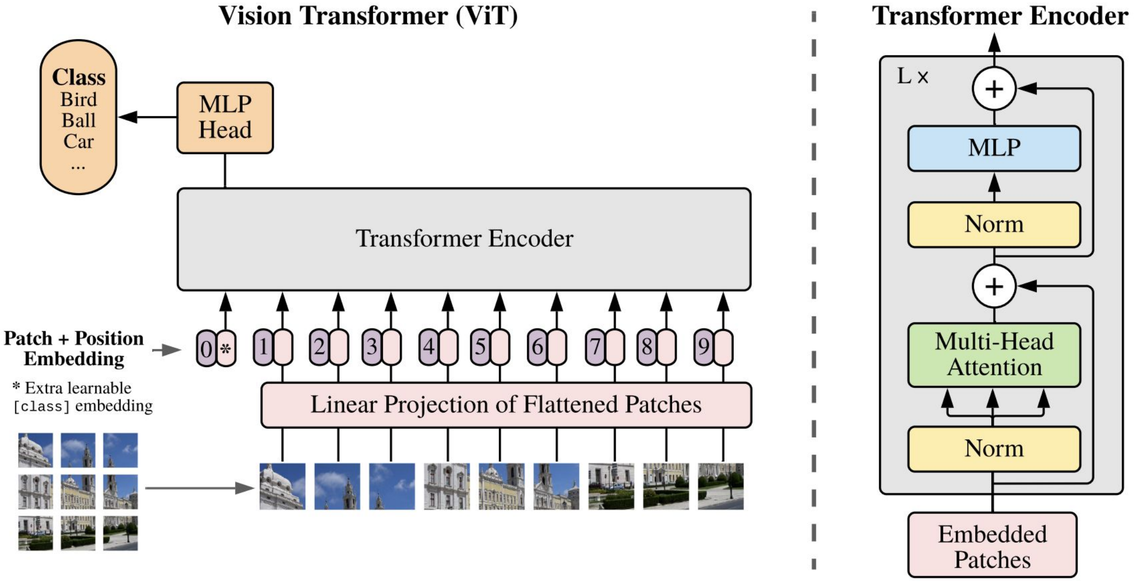


Figure: ViT model overview (VSP+17)

src

Transformer on multimodal problems

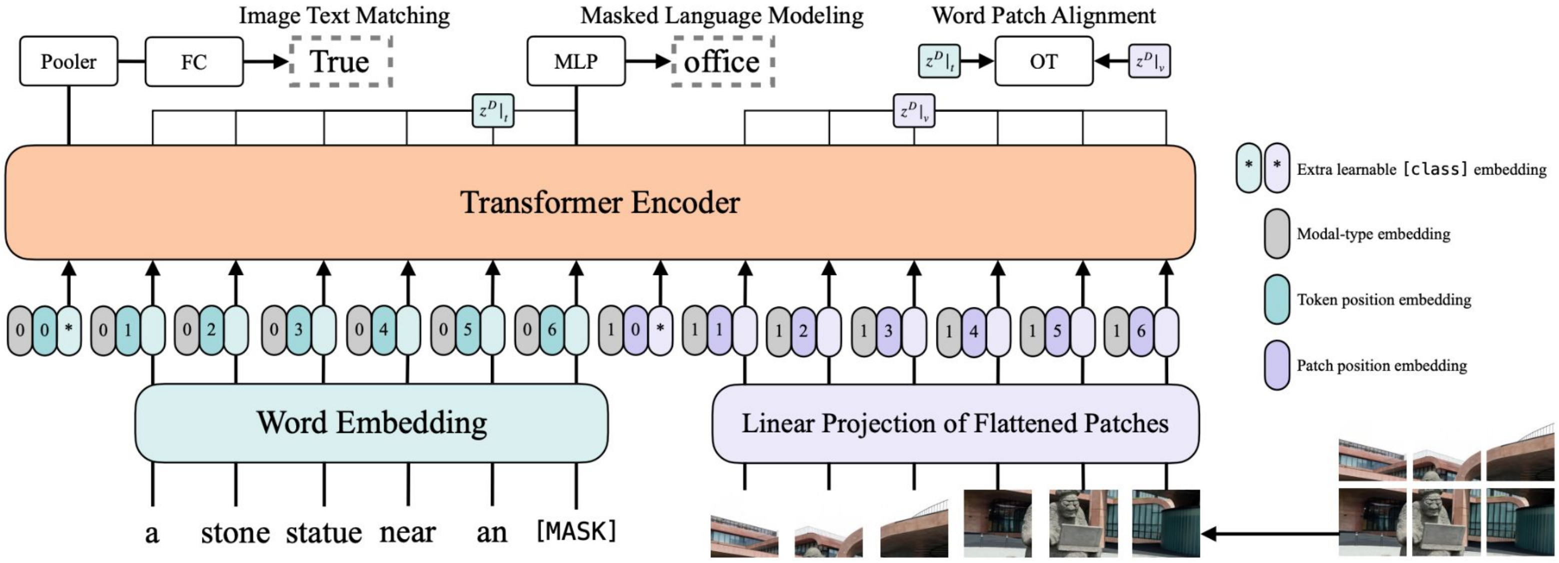


Figure: Vision-and-language transformer (ViLT) model overview (KSK21)

src

Success of Transformer

- The unification of network architectures enables us to handle multiple modalities
- There are various ways to apply Transformer due to the natures of downstream tasks
  - Dual-encoder architecture → efficient retrieval
  - Encoder-decoder networks → generation tasks
  - Fusion-encoder architecture → image-text encoding
- However,
  - Have to **manually convert the end-task formats** according to the specific architecture
  - The parameters are usually **not effectively shared** across modalities

Multiway Transformer for general-purpose modeling

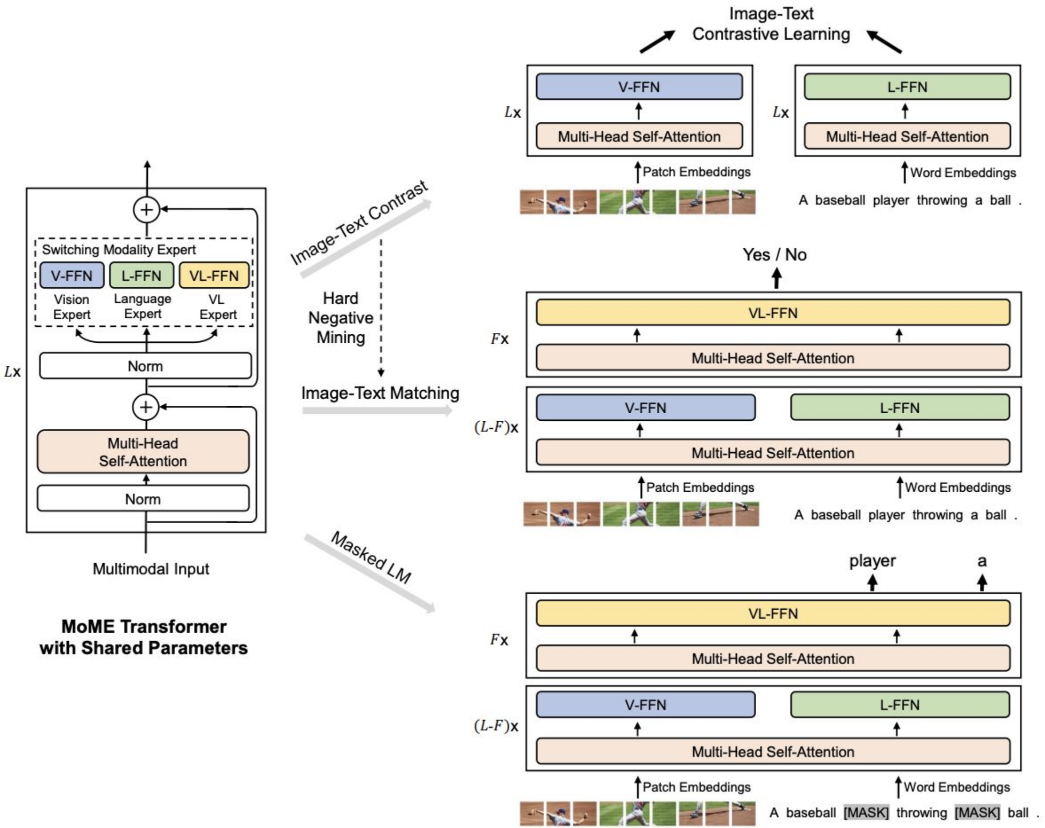


Figure: Overview of VLMo pretraining (WBDW21)

src

Pretraining task based on masked data modeling

Text

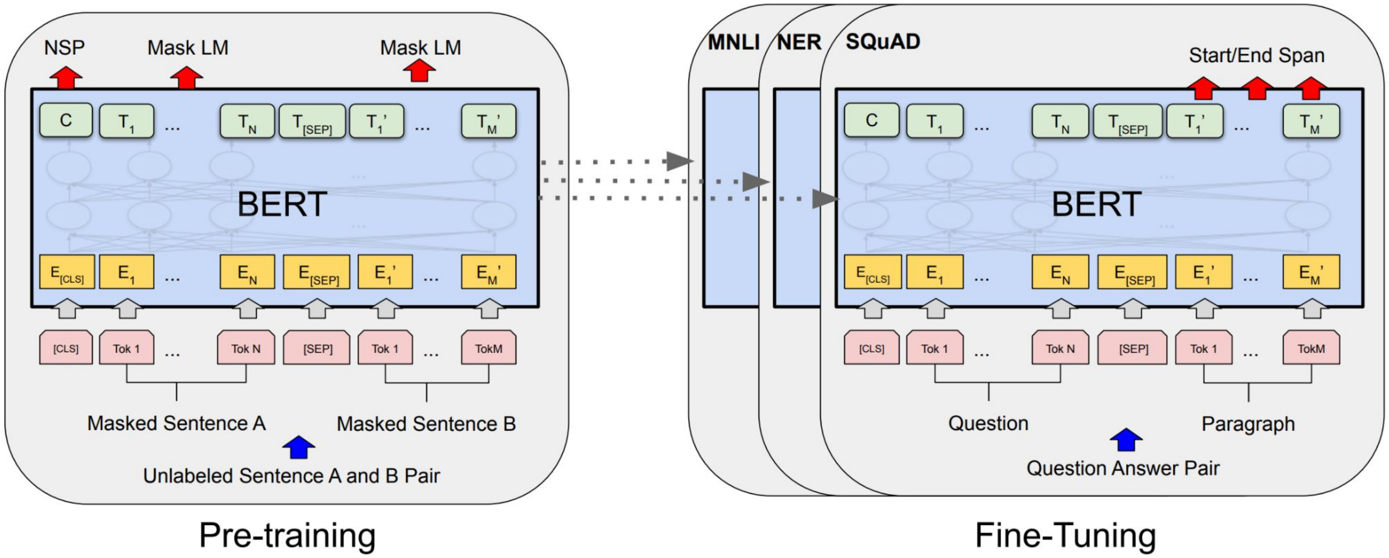


Figure: Overall pre-training and fine-tuning procedures for BERT (DCLT19)

src

Image

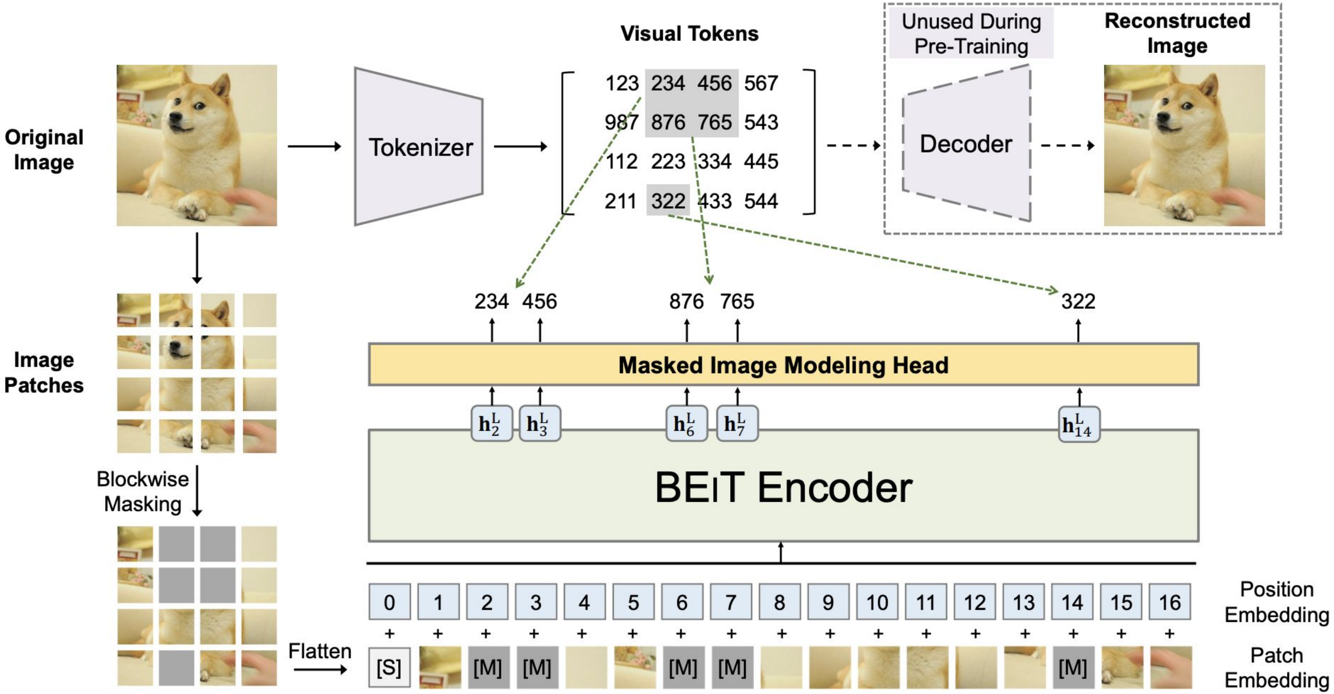


Figure: Overview of BEiT pretraining (BDPW22)

src

Motivation for masked data modeling

Drawbacks:

- Current vision-language foundation models usually multitask other pretraining objectives
- Scaling-up unfriendly and inefficient

In contrast,

- Only use one pretraining task, i.e., mask-then-predict, to train a general-purpose multimodal foundation model
- Treat the image as a foreign language (i.e., *Imglish*), then handle texts and images in the same manner

Scaling up the model and data size

Scaling up the model size and data size universally improves the generalization quality of foundation models

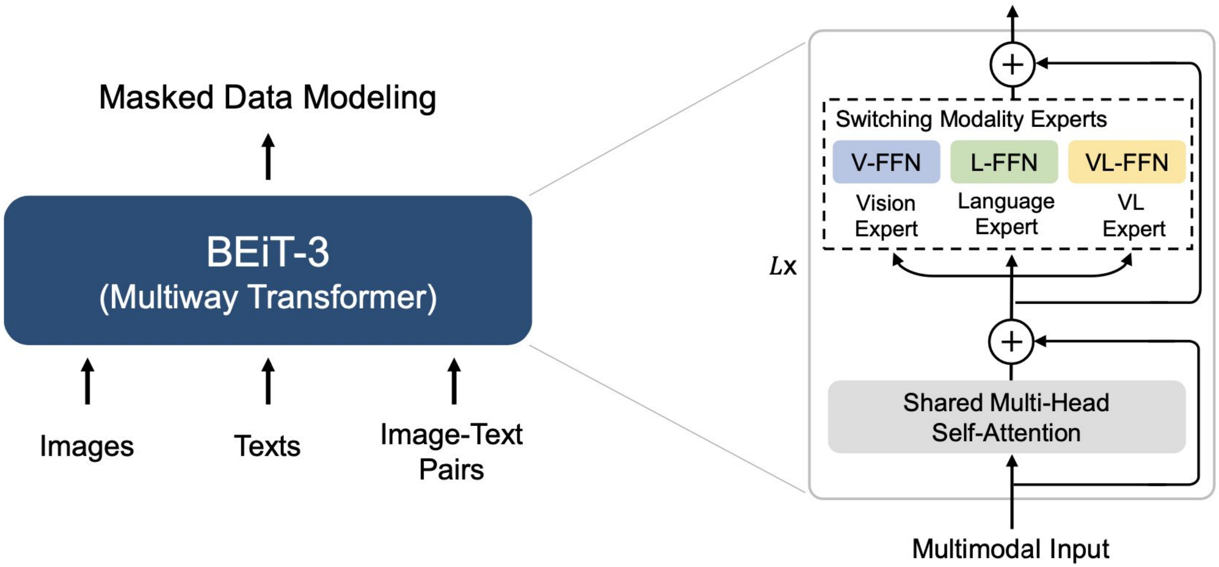
- Follow the philosophy and scale up the model size to billions of parameters
- Scale up the pretraining data size only using publicly accessible resources
- Directly reuse the pipeline developed for large-scale language model pretraining because of treating images as a foreign language

Methods

Backbone Network
Pre-training Task
Scaling Up : Pre-training BEiT

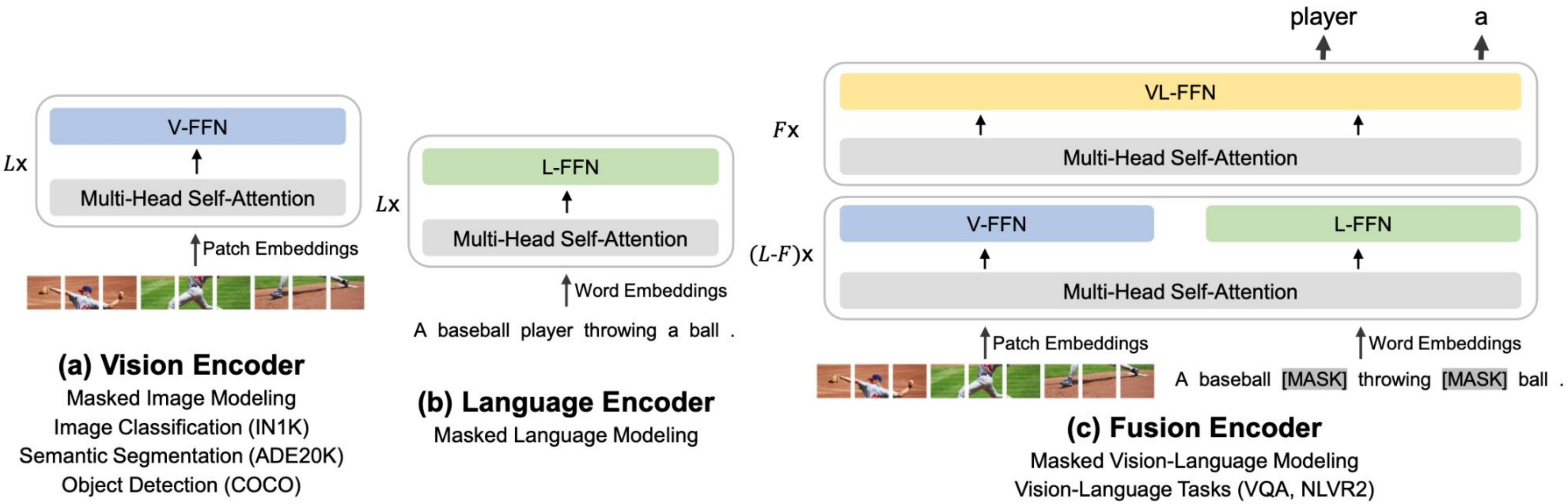
Backbone Network: Multiway Transformers

- Shared self-attention module
- Pool of Feed Forward Networks



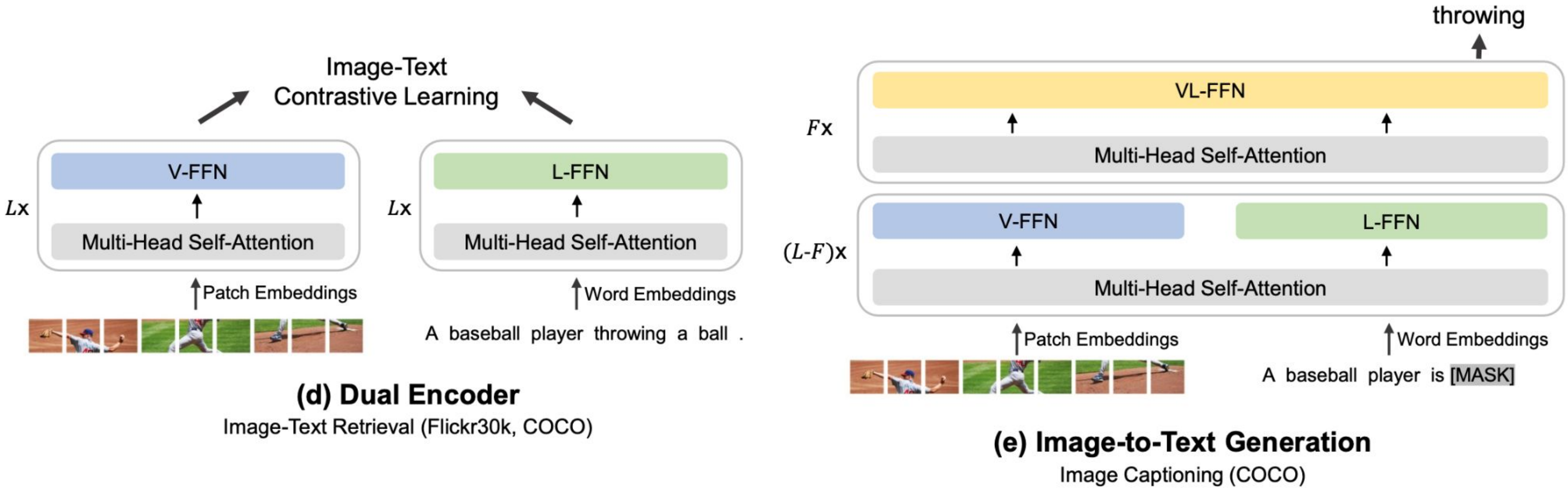
src

How Does a Multilayer Transformer Help?



src

How Does a Multilayer Transformer Help?



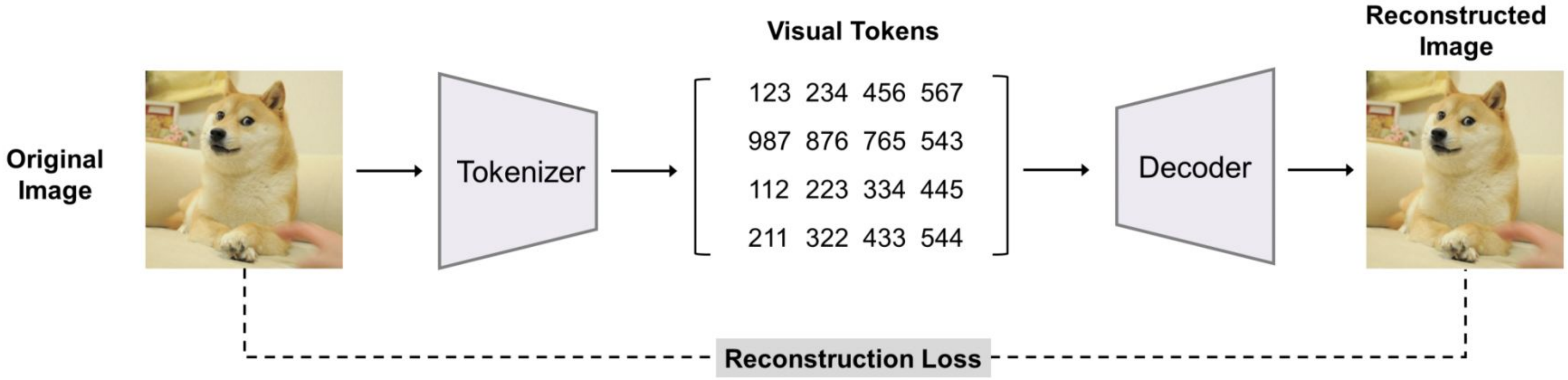
src

Pretraining

- Pre-trained for 1M steps
- Each batch contains 6144 samples with 2048 images, 2048 texts and 2048 image-text pairs
- Patch size used: 14x14
- Resolution 224X224
- SentencePiece Tokenizer is used for text
- For images, tokenizer from previous BEiT paper is used

src

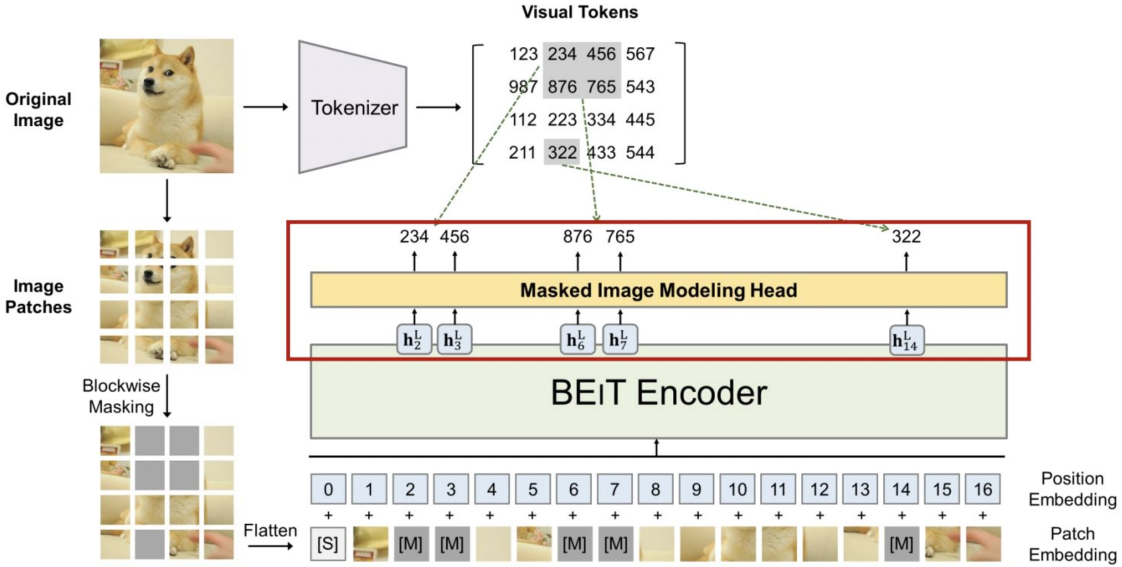
Tokenization in BEiT-3



src

Masked Data Modelling

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



src

Single Pre-training Task based on Block-wise Masking

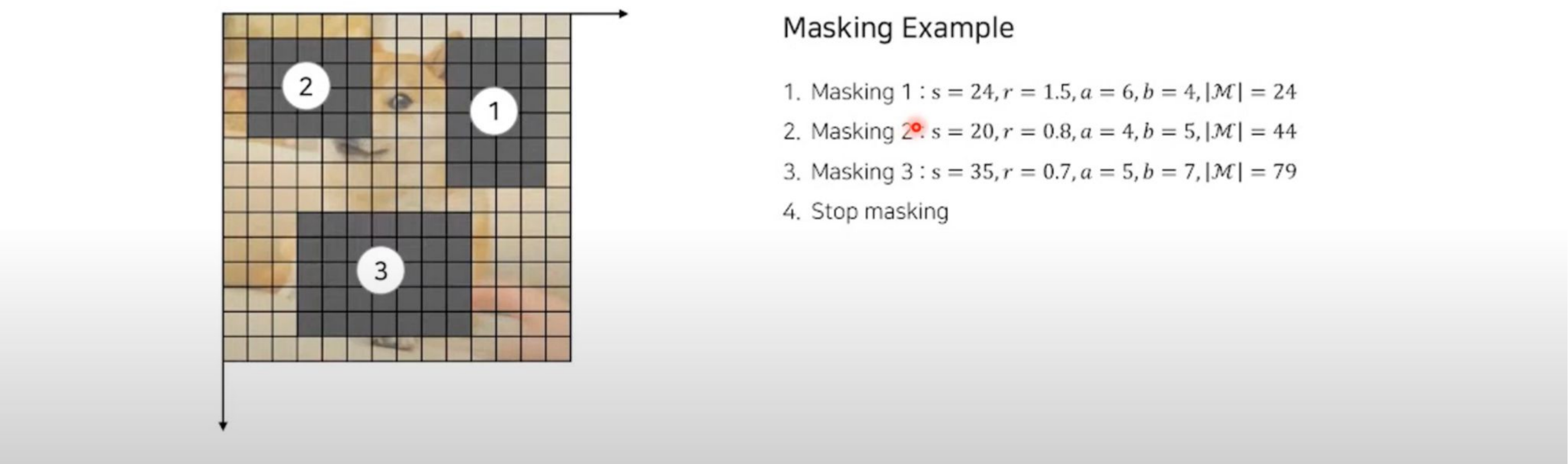
Algorithm 1 Blockwise Masking

Input: $N(= h \times w)$ image patches
Output: Masked positions $\mathcal{M}$
$\mathcal{M} \leftarrow \{\}$
repeat
  $s \leftarrow \mathsf{Rand}(16, 0.4N - |\mathcal{M}|)$ ▷ *Block size*
  $r \leftarrow \mathsf{Rand}(0.3, \frac{1}{0.3})$ ▷ *Aspect ratio of block*
  $a \leftarrow \sqrt{s \cdot r}; b \leftarrow \sqrt{s/r}$
  $t \leftarrow \mathsf{Rand}(0, h - a)\ ; l \leftarrow \mathsf{Rand}(0, w - b)$
  $\mathcal{M} \leftarrow \mathcal{M} \bigcup \{(i, j) : i \in [t, t + a), j \in [l, l + b)\}$
until $|\mathcal{M}| > 0.4N$ ▷ *Masking ratio is* 40%
return $\mathcal{M}$

src

Blockwise Masking(BEIT) Example

- Illustrated below is an example run of blockwise masking algorithm.



src

Previous models vs BEIT-3

• Multiple pre-training tasks	• single pre-training task
• large batch size	• small batch size
• Convert end task format according to specific architectures	• single architecture for various downstream tasks
• Parameters not shared across modalities	• cross-modality fusion
• Private data	• Public data

src

What role does the BEiT-3 pre-training phase play?

- Scale-up friendly
- Eliminate engineering challenges

src

Scaling Up: BEIT-3 Pre-Training

Scale up both model size and size of parameters

Model	#Layers	Hidden Size	MLP Size	#Parameters				
				V-FFN	L-FFN	VL-FFN	Shared Attention	Total
BEiT-3	40	1408	6144	692M	692M	52M	317M	1.9B

Table 2: Model configuration of BEiT-3. *The architecture layout follows ViT-giant* [scaling:vit].

Data	Source	Size
Image-Text Pair	CC12M, CC3M, SBU, COCO, VG	21M pairs
Image	ImageNet-21K	14M images
Text	English Wikipedia, BookCorpus, OpenWebText, CC-News, Stories	160GB documents

Table 3: Pretraining data of BEiT-3. *All the data are academically accessible.*

src

Evaluation & Experiments

Vision–Language Tasks
Vision Tasks

Vision-Language Downstream Tasks:

- VQA
- Visual Reasoning
- Image Captioning
- Image-Text Retrieval [src]

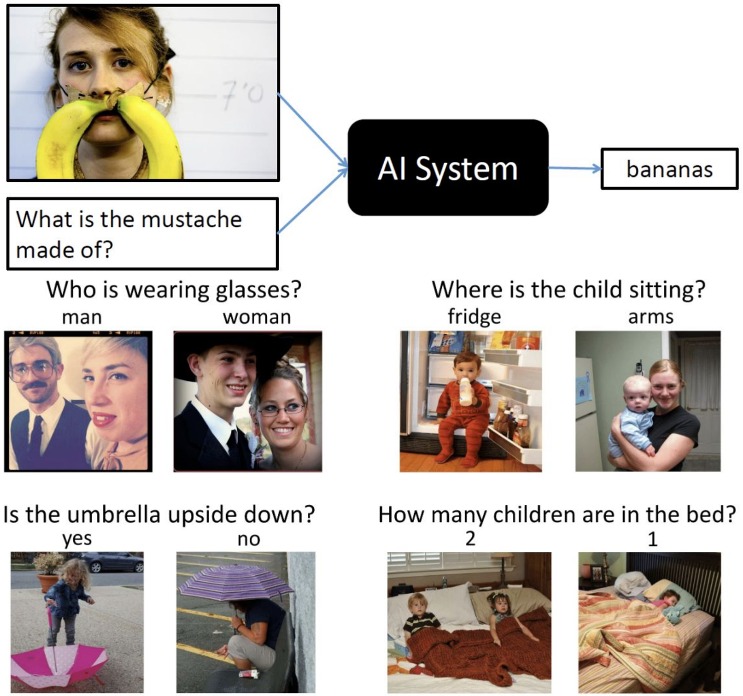
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

Image Captioning

Image from visualqa.org



[src]

Vision-Language Downstream tasks Contd...

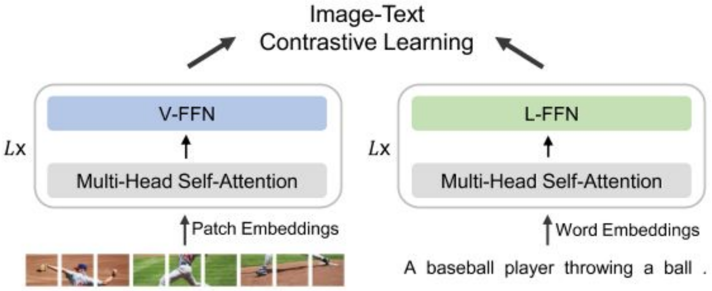
Model	VQAv2		NLVR2		COCO Captioning			
	test-dev	test-std	dev	test-P	B@4	M	C	S
Oscar [oscar]	73.61	73.82	79.12	80.37	37.4	30.7	127.8	23.5
VinVL [vinvl]	76.52	76.60	82.67	83.98	38.5	30.4	130.8	23.4
ALBEF [albef]	75.84	76.04	82.55	83.14	-	-	-	-
BLIP [blip]	78.25	78.32	82.15	82.24	40.4	-	136.7	-
SimVLM [simvlm]	80.03	80.34	84.53	85.15	40.6	33.7	143.3	**25.4**
Florence [florence]	80.16	80.36	-	-	-	-	-	-
OFA [ofa]	82.00	82.00	-	-	43.9	31.8	145.3	24.8
Flamingo [flamingo]	82.00	82.10	-	-	-	-	138.1	-
CoCa [coca]	82.30	82.30	86.10	87.00	40.9	**33.9**	143.6	24.7
BEiT-3	**84.19**	**84.03**	**91.51**	**92.58**	**44.1**	32.4	**147.6**	**25.4**

Table 4: Results of visual question answering, visual reasoning, and image captioning tasks. We report vqa-score on VQAv2 test-dev and test-standard splits, accuracy for NLVR2 development set and public test set (test-P). For COCO image captioning, we report BLEU@4 (B@4), METEOR (M), CIDEr (C), and SPICE (S) on the Karpathy test split. For simplicity, we report captioning results without using CIDEr optimization.

src

BEIT-3 at work (example) : Image-Text Retrieval

- Measure similarity between image and texts : I2T, T2I
- Directly finetune BEiT-3 on COCO and Flickr30K: no image-text contrastive objective during pre-training



(d) Dual Encoder
Image-Text Retrieval (Flickr30k, COCO)

Cosine similarity between both representations

A baseball player throwing a ball

Vision-Language Downstream tasks Contd...

src

Model	MSCOCO (5K test set)						Flickr30K (1K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Fusion-encoder models												
UNITER [uniter]	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA [villa]	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
Oscar [oscar]	73.5	92.2	96.0	57.5	82.8	89.8	-	-	-	-	-	-
VinVL [vinvl]	75.4	92.9	96.2	58.8	83.5	90.3	-	-	-	-	-	-
Dual encoder + Fusion encoder reranking												
ALBEF [albef]	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	**100.0**	85.6	97.5	98.9
BLIP [blip]	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0
Dual-encoder models												
ALIGN [align]	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	**100.0**	84.9	97.4	98.6
FILIP [filip]	78.9	94.4	97.4	61.2	84.3	90.6	96.6	**100.0**	**100.0**	87.1	97.7	99.1
Florence [florence]	81.8	95.2	-	63.2	85.7	-	97.2	99.9	-	87.9	98.1	-
BEiT-3	**84.8**	**96.5**	**98.3**	**67.2**	**87.7**	**92.8**	**98.0**	**100.0**	**100.0**	**90.3**	**98.7**	**99.5**

Model	Flickr30K (1K test set)					
	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
FLAVA [flava]	67.7	94.0	-	65.2	89.4	-
CLIP [clip]	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [align]	88.6	98.7	99.7	75.7	93.8	96.8
FILIP [filip]	89.8	99.2	99.8	75.0	93.4	96.3
Florence [florence]	90.9	99.1	-	76.7	93.6	-
Flamingo [flamingo]	89.3	98.8	99.7	79.5	95.3	**97.9**
CoCa [coca]	92.5	99.5	99.9	80.4	**95.7**	97.7
BEiT-3	**94.9**	**99.9**	**100.0**	**81.5**	95.6	97.8

Table 6: Zero-shot image-to-text retrieval and text-to-image retrieval on Flickr30K.

Vision Downstream Tasks

src

Object Detection & Instance Segmentation

Model	Extra OD Data	Maximum Image Size	COCO test-dev	
			AP^{box}	AP^{mask}
ViT-Adapter [vit-adapter]	-	1600	60.1	52.1
DyHead [dyhead]	ImageNet-Pseudo Labels	2000	60.6	-
Soft Teacher [soft_teacher]	Object365	-	61.3	53.0
GLIP [glip]	FourODs	-	61.5	-
GLIPv2 [glipv2]	FourODs	-	62.4	-
Florence [florence]	FLOD-9M	2500	62.4	-
SwinV2-G [swinv2]	Object365	1536	63.1	54.4
Mask DINO [mask_dino]	Object365	1280	-	54.7
DINO [dino-od]	Object365	2000	63.3	-
BEiT-3	Object365	1280	**63.7**	**54.8**

Semantic Segmentation

Model	Crop Size	ADE20K	
		mIoU	+MS
HorNet [HorNet]	640^2	57.5	57.9
SeMask [jain2021semask]	640^2	57.0	58.3
SwinV2-G [swinv2]	896^2	59.3	59.9
ViT-Adapter [vit-adapter]	896^2	59.4	60.5
Mask DINO [mask_dino]	-	59.5	60.8
FD-SwinV2-G [fd-swin]	896^2	-	61.4
BEiT-3	896^2	**62.0**	**62.8**

Vision Downstream Tasks Contd...

Image Classification

Model	Extra Data	Image Size	ImageNet
*With extra **private** image-tag data*			
SwinV2-G [swinv2]	IN-22K-ext-70M	640^2	90.2
ViT-G [scaling:vit]	JFT-3B	518^2	90.5
CoAtNet-7 [coatnet]	JFT-3B	512^2	90.9
Model Soups [modelsoups]	JFT-3B	500^2	91.0
CoCa [coca]	JFT-3B	576^2	91.0
*With only **public** image-tag data*			
BEiT [beit]	IN-21K	512^2	88.6
CoAtNet-4 [coatnet]	IN-21K	512^2	88.6
MaxViT [maxvit]	IN-21K	512^2	88.7
MViTv2 [mvitv2]	IN-21K	512^2	88.8
FD-CLIP [fd-swin]	IN-21K	336^2	89.0
BEiT-3	IN-21K	336^2	**89.6**

Table 9: Top-1 accuracy on ImageNet-1K.

src

Overview of BEIT results:

Category	Task	Dataset	Metric	Previous SOTA	BEiT-3
Vision	Semantic Segmentation	ADE20K	mIoU	61.4 (FD-SwinV2)	**62.8 (+1.4)**
	Object Detection	COCO	AP	63.3 (DINO)	**63.7 (+0.4)**
	Instance Segmentation	COCO	AP	54.7 (Mask DINO)	**54.8 (+0.1)**
	Image Classification	ImageNet†	Top-1 acc.	89.0 (FD-CLIP)	**89.6 (+0.6)**
Vision-Language	Visual Reasoning	NLVR2	Acc.	87.0 (CoCa)	**92.6 (+5.6)**
	Visual QA	VQAv2	VQA acc.	82.3 (CoCa)	**84.0 (+1.7)**
	Image Captioning	COCO‡	CIDEr	145.3 (OFA)	**147.6 (+2.3)**
	Finetuned Retrieval	COCO	R@1	72.5 (Florence)	**76.0 (+3.5)**
		Flickr30K		92.6 (Florence)	**94.2 (+1.6)**
	Zero-shot Retrieval	Flickr30K	R@1	86.5 (CoCa)	**88.2 (+1.7)**

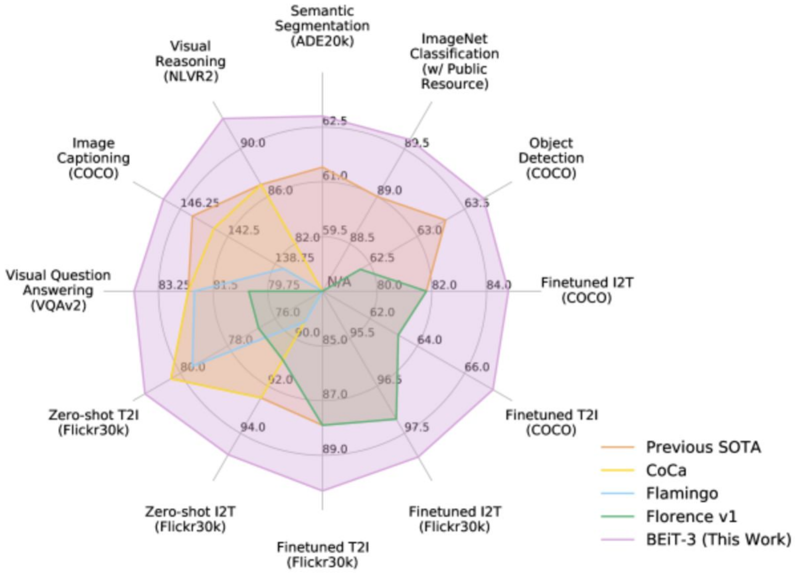


Table 1: Overview of BEiT-3 *results on various vision and vision-language benchmarks. We compare with previous state-of-the-art models, including FD-SwinV2* [fd-swin], *DINO* [dino-od], *Mask DINO* [dino-od], *FD-CLIP* [fd-swin], *CoCa* [coca], *OFA* [ofa], *Florence* [florence]. *We report the average of top-1 image-to-text and text-to-image results for retrieval tasks. "†" indicates ImageNet results only using publicly accessible resources. "‡" indicates image captioning results without CIDEr optimization.*

src

Pros

- One unified architecture shared for various downstream tasks
- Scaling Up friendly
- Overcoming GPU memory cost challenges
- Same pipeline developed for large language model pretraining used for images as they are treated as a foreign language
- Publicly accessible data is used
- State of the art performance over both vision and vision-language tasks

Cons

- Large amount data required to train
- Large memory requirement for storing the tokens
- May not be unified in true sense

src

Future Directions

- Extend the model across more modalities: audio, video etc.
- Pretraining Multilingual BEiT
- Enable in-context learning capabilities
- Further exploration of alignment of modalities
- Use single codebook/vocab for both image and text

src

Summary

- It's a general purpose multimodal foundation model that achieves SOTA transfer performance on both vision and vision-language tasks.
- Introduces multiway transformer for general purpose modelling
- Uses masked language modelling used on images(Imglish), texts(English) and image-text pairs (parallel sentences)
- Pretrained using single task, mask-then-predict
- Scale Up friendly model that outperforms both previous foundation models and specialized models on vision and vision-language tasks.

src

References

1. Wang, Wenhui, et al. "Image as a foreign language: Beit pretraining for all vision and vision-language tasks." arXiv preprint arXiv:2208.10442 (2022).
2. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).
3. Kim, Wonjae, Bokyung Son, and Ildoo Kim. "Vilt: Vision-and-language transformer without convolution or region supervision." International Conference on Machine Learning. PMLR, 2021.
4. Bao, Hangbo, et al. "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts." arXiv preprint arXiv:2111.02358 (2021).
5. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
6. Bao, Hangbo, et al. "Beit: Bert pre-training of image transformers." arXiv preprint arXiv:2106.08254 (2021).
7. Bao, Hangbo, Wenhui Wang, Li Dong, and Furu Wei. "Vl-beit: Generative vision-language pretraining." arXiv preprint arXiv:2206.01127 (2022).
8. Peng, Zhiliang, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. "Beit v2: Masked image modeling with vector-quantized visual tokenizers." arXiv preprint arXiv:2208.06366 (2022).

Thank You!

Q&A