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

Authors: *Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, Furu Wei, Microsoft Corporation*

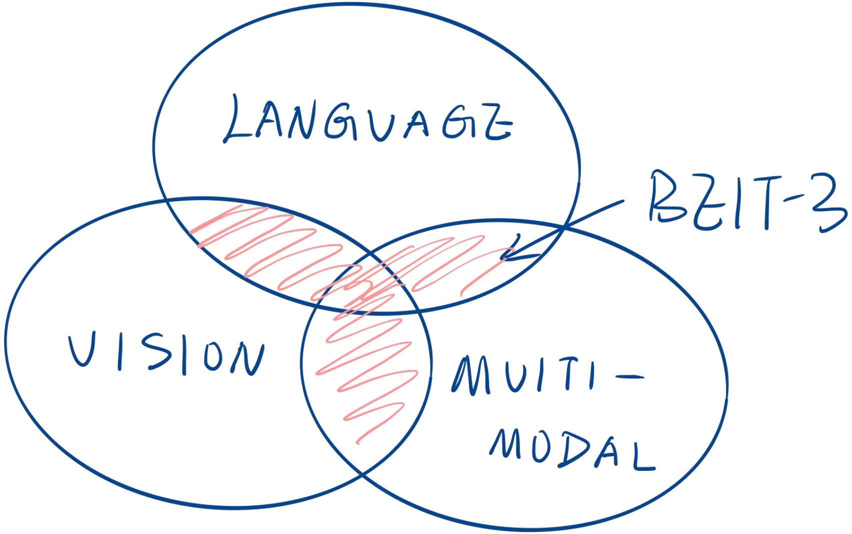
Presenter: *Tianyi Wang*

Overview

- Introduction of BEIT-3
- Artitecture of BEIT-3
- Pretraing task and setup
- Experiment
- Conclusion
- Q & A

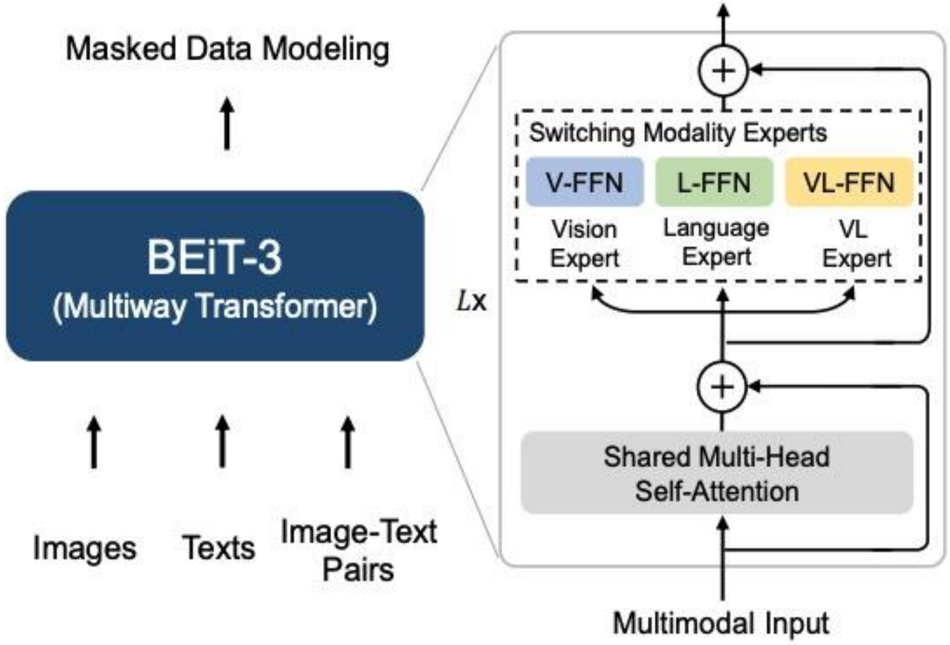
Why BEIT-3?

- Convergence of Modalities:
    - ***Trend***: Increasing integration of language, vision, and multimodal pretraining.
    - ***Objective***: Build a versatile foundation model capable of handling multiple modalities.



Why BEIT-3?

- Transformers in Vision and Multimodal Problems:
  - ***Unified Architectures***: Adoption of Transformer models for various modalities.
  - ***Tailored Solutions***: Providing seamless and effective solutions for diverse downstream tasks.

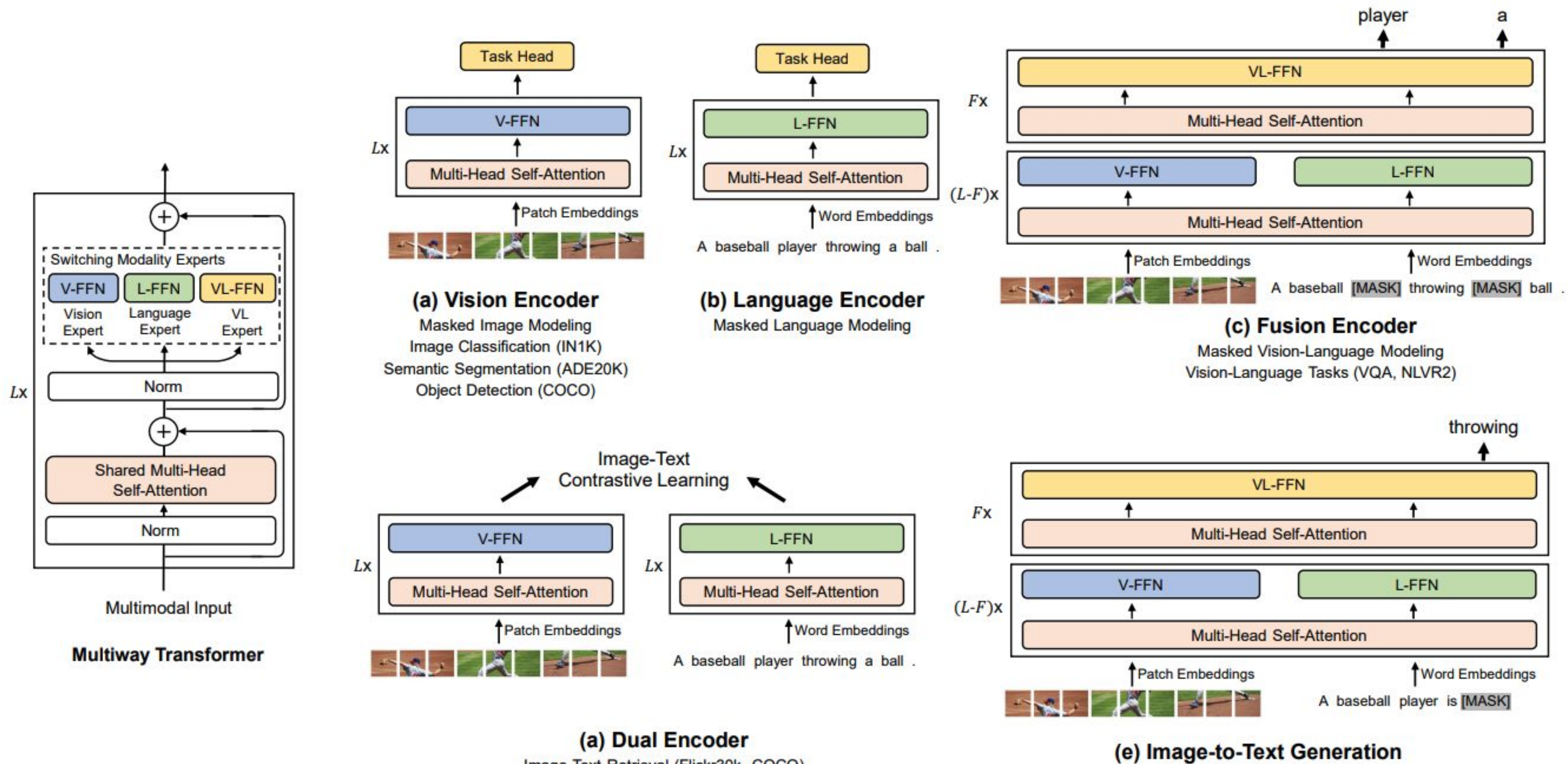


Why BEIT-3?

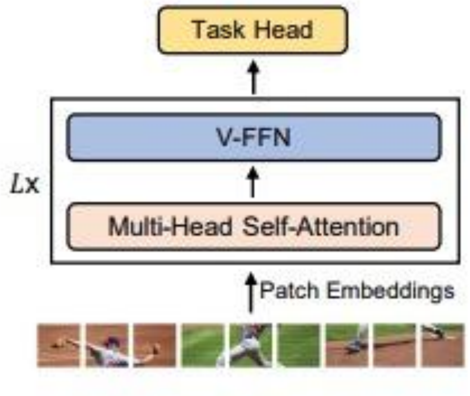
- Masked Data Modeling and Scaling:
  - ***Simplification***: Adopting a mask-then-predict approach for pretraining tasks.
  - ***Scaling***: Focusing on enlarging the model size and dataset to boost generalization and performance.

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)

Understanding BEIT-3: A Multimodal Foundation Model



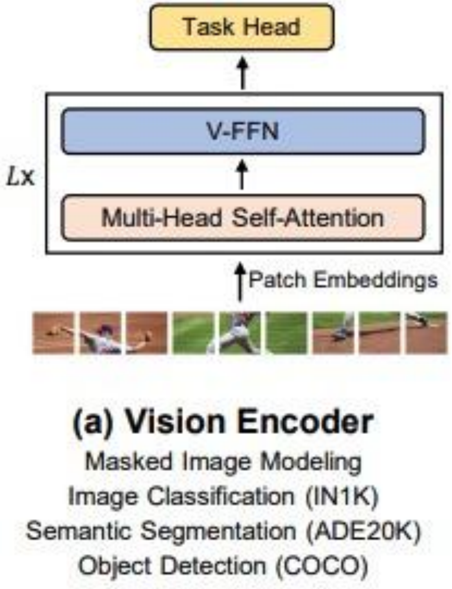
Understanding BEIT-3: A Multimodal Foundation Model



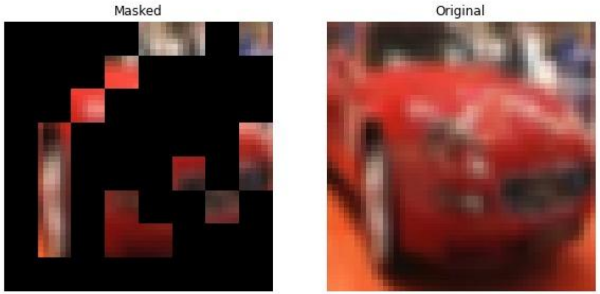
(a) Vision Encoder

Masked Image Modeling
Image Classification (IN1K)
Semantic Segmentation (ADE20K)
Object Detection (COCO)

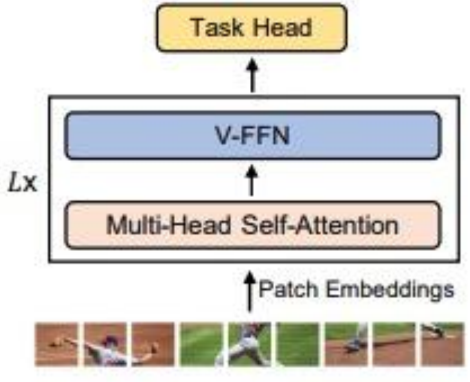
Understanding BEIT-3: A Multimodal Foundation Model



Masked Image Modeling



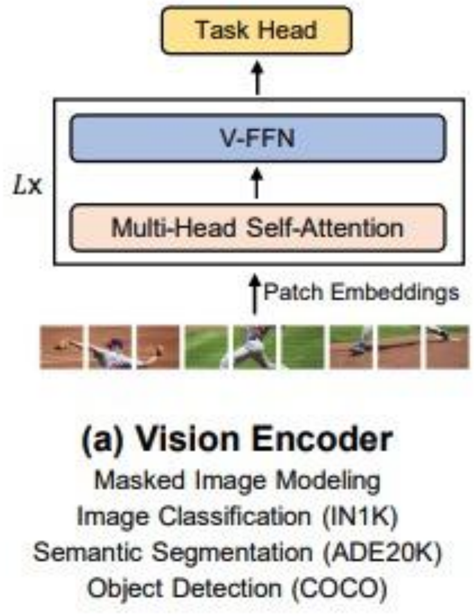
Understanding BEIT-3: A Multimodal Foundation Model



(a) Vision Encoder
Masked Image Modeling
Image Classification (IN1K)
Semantic Segmentation (ADE20K)
Object Detection (COCO)

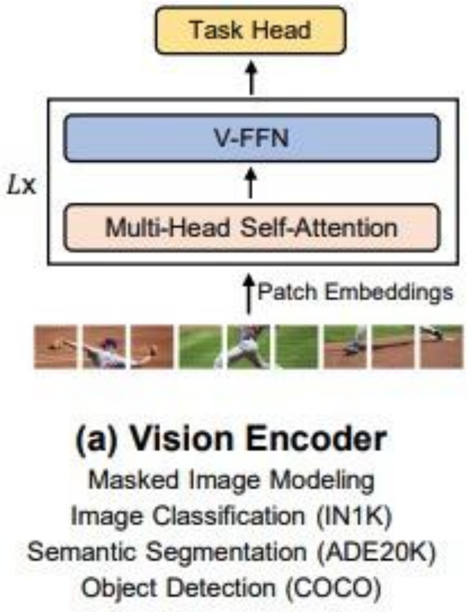
Image Classification

Understanding BEIT-3: A Multimodal Foundation Model

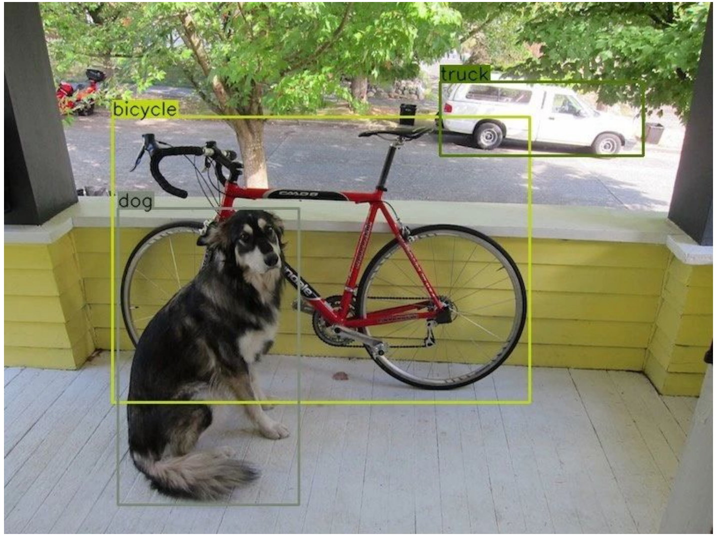


Semantic Segmentation

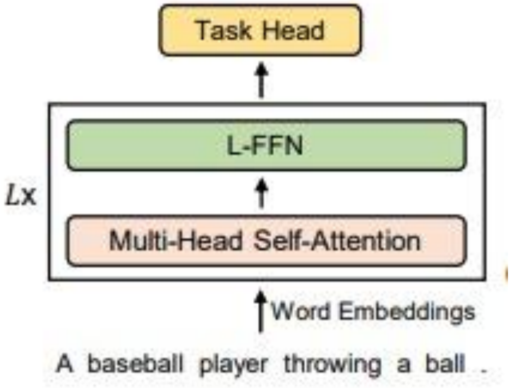
Understanding BEIT-3: A Multimodal Foundation Model



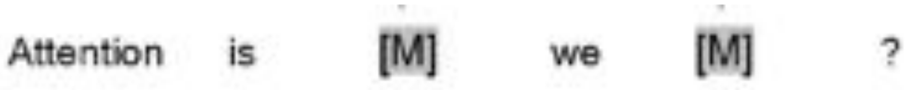
Object Detection



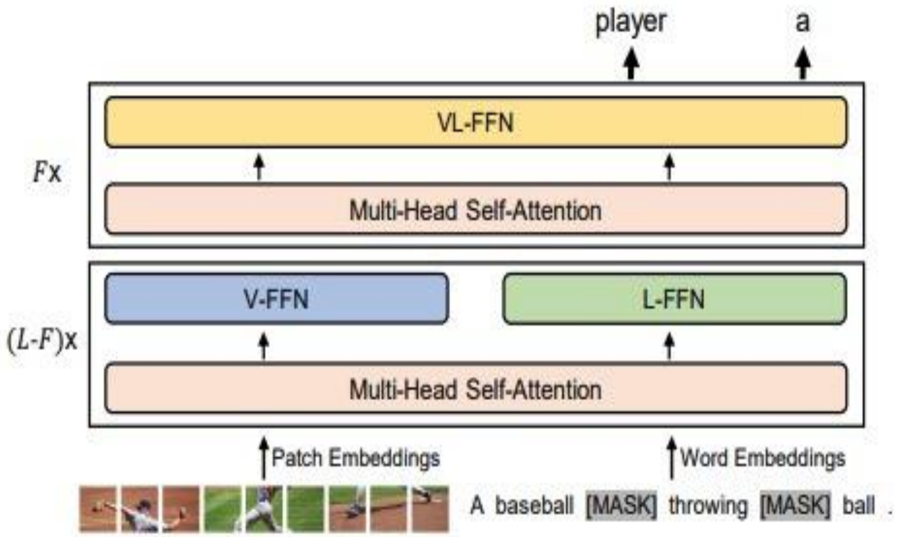
Understanding BEIT-3: A Multimodal Foundation Model



(b) Language Encoder
Masked Language Modeling



Understanding BEIT-3: A Multimodal Foundation Model

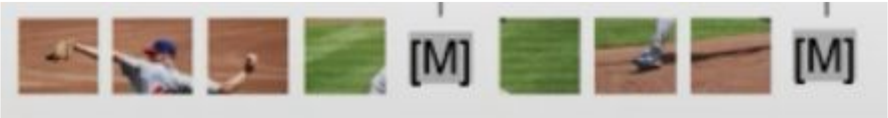


(c) Fusion Encoder

Masked Vision-Language Modeling

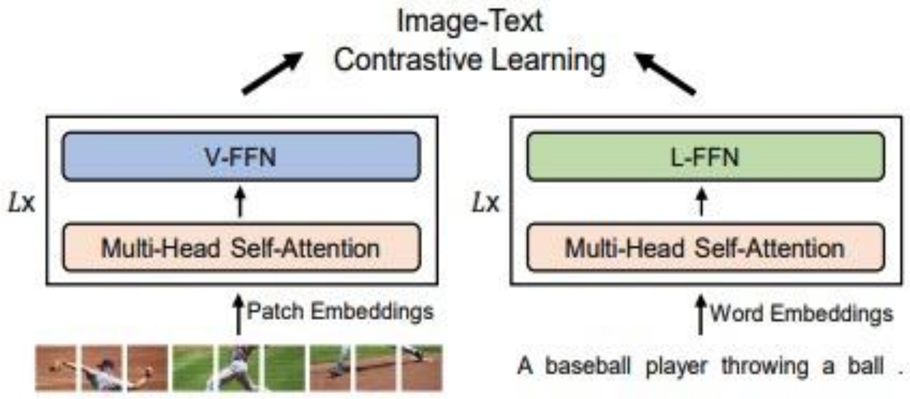
Vision-Language Tasks (VQA, NLVR2)

A baseball [MASK] throwing [MASK] ball .



A baseball player throwing a ball .

Understanding BEIT-3: A Multimodal Foundation Model

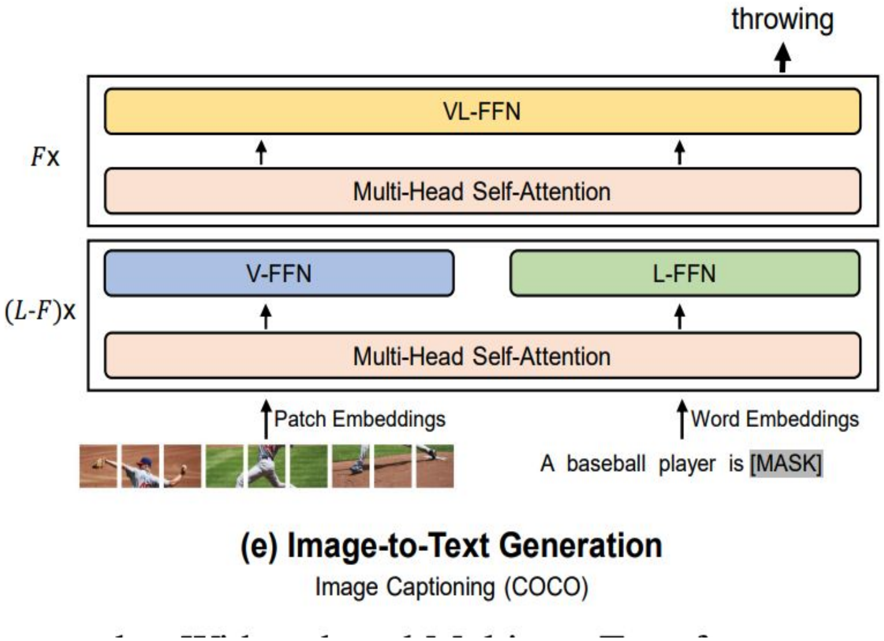


(a) Dual Encoder

Image-Text Retrieval (Flickr30k, COCO)

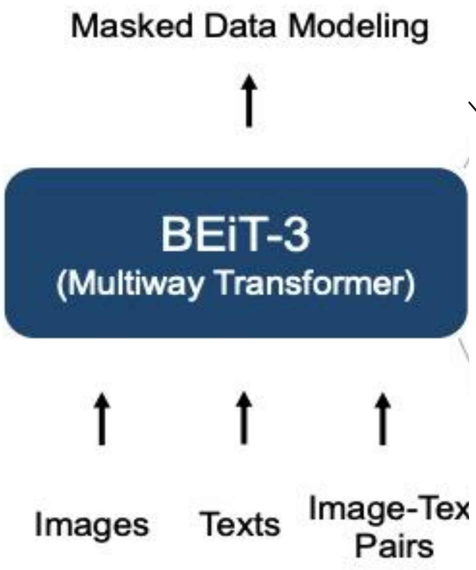
A baseball player throwing a ball .

Understanding BEIT-3: A Multimodal Foundation Model



A baseball player is [MASK]

Pretraining tasks



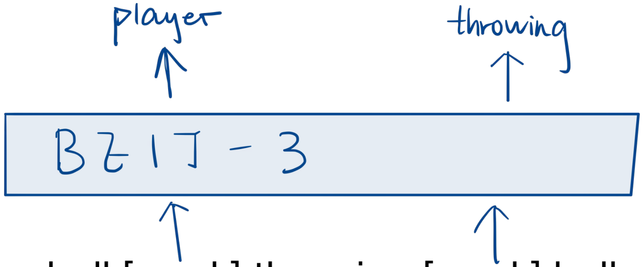
Masked Language Modeling

Masked Image Modeling

Masked Vision - Language Modeling

Pretraining tasks

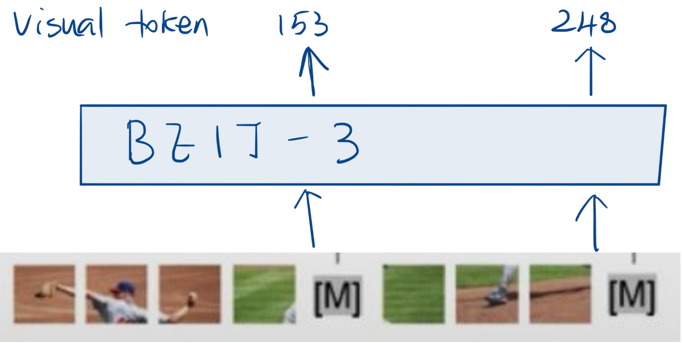
Masked Language Modeling



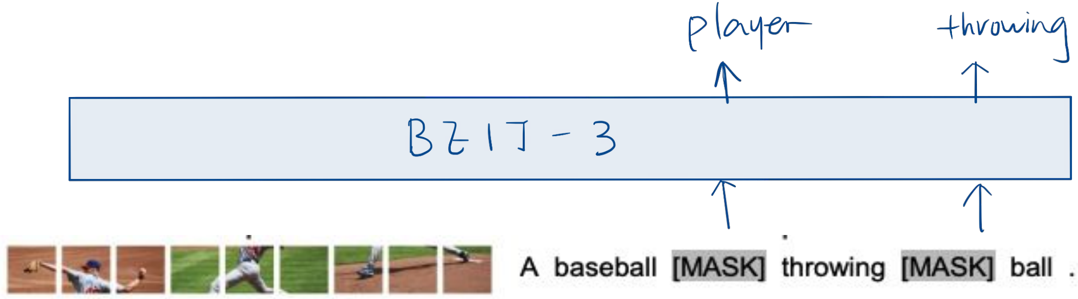
A baseball [mask] throwing [mask] ball.

Pretraining tasks

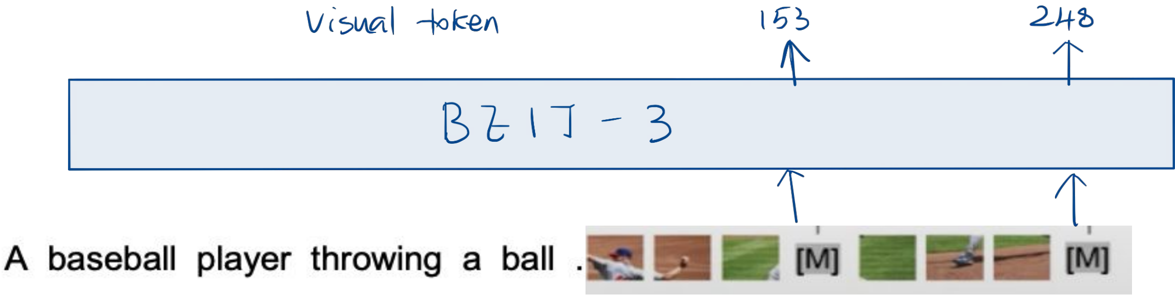
Masked Image Modeling



Pretraining tasks



Masked Vision - Language Modeling



Pretraining Setup and Scaling Up

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

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

Experiment - Vision-Language Downstream Tasks

- Objective: Assess BEIT-3's performance in tasks that require understanding both images and text.
- Tasks Include:
  - Image captioning
  - Text-to-image synthesis
  - Visual question answering

Experiment - Vision - Language Task

Model	VQAv2		NLVR2		COCO Captioning			
	test-dev	test-std	dev	test-P	B@4	M	C	S
Oscar [LYL+20]	73.61	73.82	79.12	80.37	37.4	30.7	127.8	23.5
VinVL [ZLH+21]	76.52	76.60	82.67	83.98	38.5	30.4	130.8	23.4
ALBEF [LSG+21]	75.84	76.04	82.55	83.14	-	-	-	-
BLIP [LLXH22]	78.25	78.32	82.15	82.24	40.4	-	136.7	-
SimVLM [WYY+21]	80.03	80.34	84.53	85.15	40.6	33.7	143.3	**25.4**
Florence [YCC+21]	80.16	80.36	-	-	-	-	-	-
OFA [WYM+22]	82.00	82.00	-	-	43.9	31.8	145.3	24.8
Flamingo [ADL+22]	82.00	82.10	-	-	-	-	138.1	-
CoCa [YWV+22]	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**

Visual Question Answering

Experiment - Vision Downstream Tasks

- Objective: Evaluate BEIT-3's prowess in purely visual tasks.
- Tasks Include:
  - Object detection
  - Semantic segmentation
  - Image classification

Experiment - Vision Downstream Tasks

Model	Extra OD Data	Maximum Image Size	COCO test-dev	
			AP^{box}	AP^{mask}
ViT-Adapter [CDW+22]	-	1600	60.1	52.1
DyHead [DCX+21]	ImageNet-Pseudo Labels	2000	60.6	-
Soft Teacher [XZH+21]	Object365	-	61.3	53.0
GLIP [LZZ+21]	FourODs	-	61.5	-
GLIPv2 [ZZH+22]	FourODs	-	62.4	-
Florence [YCC+21]	FLOD-9M	2500	62.4	-
SwinV2-G [LHL+21]	Object365	1536	63.1	54.4
Mask DINO [LZX+22]	Object365	1280	-	54.7
DINO [ZLL+22]	Object365	2000	63.3	-
BEiT-3	Object365	1280	**63.7**	**54.8**

Experiment - Vision Downstream Tasks

Model	Crop Size	ADE20K mIoU	ADE20K +MS
HorNet [RZT+22]	640^2	57.5	57.9
SeMask [JSO+21]	640^2	57.0	58.3
SwinV2-G [LHL+21]	896^2	59.3	59.9
ViT-Adapter [CDW+22]	896^2	59.4	60.5
Mask DINO [LZX+22]	-	59.5	60.8
FD-SwinV2-G [WHX+22]	896^2	-	61.4
BEiT-3	896^2	**62.0**	**62.8**

Experiment - Vision Downstream Tasks

Model	Extra Data	Image Size	ImageNet
*With extra **private** image-tag data*			
SwinV2-G [LHL+21]	IN-22K-ext-70M	640^2	90.2
ViT-G [ZKHB21]	JFT-3B	518^2	90.5
CoAtNet-7 [DLLT21]	JFT-3B	512^2	90.9
Model Soups [WIG+22]	JFT-3B	500^2	91.0
CoCa [YWV+22]	JFT-3B	576^2	91.0
*With only **public** image-tag data*			
BEIT [BDPW22]	IN-21K	512^2	88.6
CoAtNet-4 [DLLT21]	IN-21K	512^2	88.6
MaxViT [TTZ+22]	IN-21K	512^2	88.7
MViTv2 [LWF+22]	IN-21K	512^2	88.8
FD-CLIP [WHX+22]	IN-21K	336^2	89.0
BEIT-3	IN-21K	336^2	**89.6**

Experiment - Ablation Studies

Transformer	VQA	NLVR2	F30K
Standard	76.1	80.8	82.8
Multiway	**76.8**	**81.4**	**84.4**

(a) Multiway Transformer improves the performance over the conventional one.

Strategy	VQA	NLVR2	F30K
Joint	75.7	79.0	83.1
Separate	**76.8**	**81.4**	**84.4**

(b) Separate masking in MVLM is helpful.

Mono	Multi	VQA	NLVR2	F30K
✓	✗	71.3	64.6	79.3
✗	✓	75.8	79.3	81.1
✓	✓	**76.8**	**81.4**	**84.4**

(c) Whether we conduct masked prediction for monomodal (mono) and multimodal (multi) data.

Target	VQA	NLVR2	F30K
DALL-E [47]	73.2	77.7	76.6
Pixel (w/ norm) [19]	73.3	77.1	75.9
VQ-KD$_{\text{CLIP}}$ [43]	**76.8**	**81.4**	**84.4**

(d) Targets used for image reconstruction. VQ-KD$_{\text{CLIP}}$ [43] works the best.

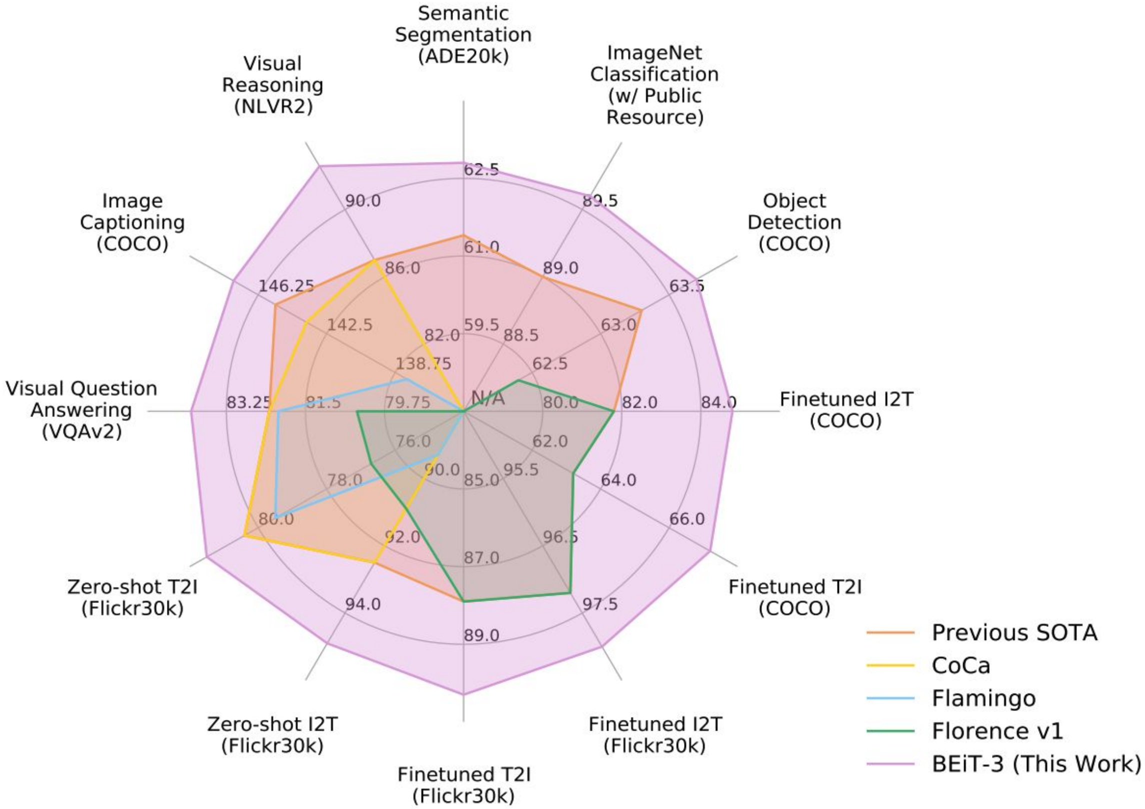
Mono	Multi	VQA	NLVR2	F30K
✗	✗	71.5	69.3	77.8
✓	✗	73.2	76.4	81.3
✗	✓	76.5	80.6	82.7
✓	✓	**76.8**	**81.4**	**84.4**

(e) Whether we enable text reconstruction for monomodal (mono) and multimodal (multi) data.

Mono	Multi	VQA	NLVR2	F30K
✗	✗	71.6	74.3	71.7
✓	✗	75.8	79.8	82.0
✗	✓	75.6	79.5	81.9
✓	✓	**76.8**	**81.4**	**84.4**

(f) Whether we enable image reconstruction for monomodal (mono) and multimodal (multi) data.

Experiment - Summary



Conclusion

- **BEIT-3, a general-purpose multimodal foundation model:**
  - achieves state-of-the-art performance across a wide range of vision and vision-language benchmarks.
- **Innovative Approach:**
  - monomodal(images, texts) and multimodal (image-text pair)
- **Multiway Transformer:**
  - emphasizes the efficiency of Multiway Transformers in addressing a variety of vision and vision-language tasks.
- **Future Direction**

Q & A

Which of the following best describes the purpose of "Unified Architectures" in the context of BEiT-3?

A) It allows BEiT-3 to process only image data.

B) It enables separate specialized architectures for each data type.

C) It provides BEiT-3 the ability to handle both individual (monomodal) and combined (multimodal) data types within a single framework.

D) It restricts BEiT-3 to text-only tasks.

Which of the following best describes the purpose of "Unified Architectures" in the context of BEiT-3?

A) It allows BEiT-3 to process only image data.

B) It enables separate specialized architectures for each data type.

C) It provides BEiT-3 the ability to handle both individual (monomodal) and combined (multimodal) data types within a single framework.

D) It restricts BEiT-3 to text-only tasks.

Why is scaling up models and systems often considered beneficial in deep learning?

A) To improve model generalization and performance on complex tasks.

B) To reduce the amount of training data required.

C) To make models more interpretable.

D) To decrease computational resources and speed up training.

Why is scaling up models and systems often considered beneficial in deep learning?

A) To improve model generalization and performance on complex tasks.

B) To reduce the amount of training data required.

C) To make models more interpretable.

D) To decrease computational resources and speed up training.