Rethinking the Effectiveness of Masked Adapter: Can Unimodality Assist Multimodality?

CSCI-566: Deep Learning and Its Applications

Team: DODO Bird

Apr 21, 2023

USC

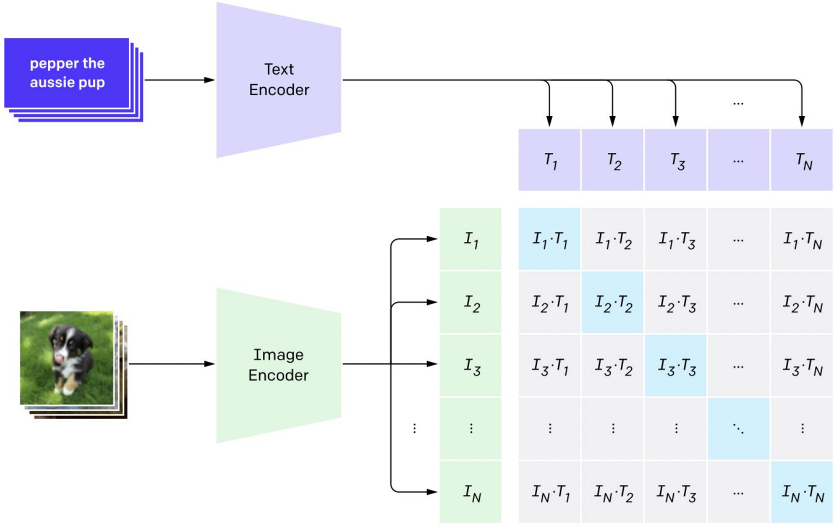
Outline

- Motivations
- Overall Methods
- Experiments
- Future Work

Motivations

Motivations

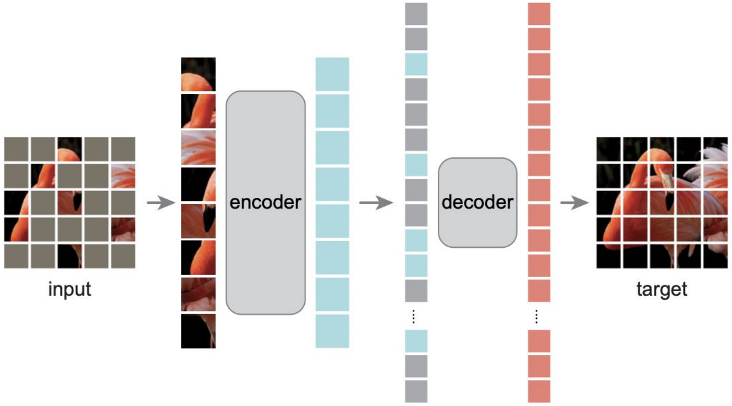
- Prevalent multimodal models, like CLIP, have shown outstanding performance on multimodal tasks. However, since CLIP is only trained on image-text contrastive loss, it suffers from poor unimodal representation.
- With strengthened unimodal representation, we hypothesize the model can have a better performance on multimodal tasks



Source: OpenAI CLIP
(Radford et al., 2021)

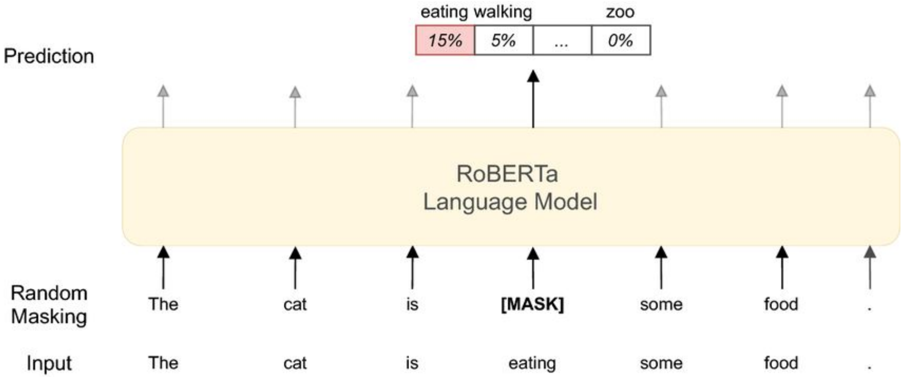
Motivations

- How to gain a better unimodal representation? Masked modeling!



Masked Image Modeling

Source: MAE (He et al., 2022)

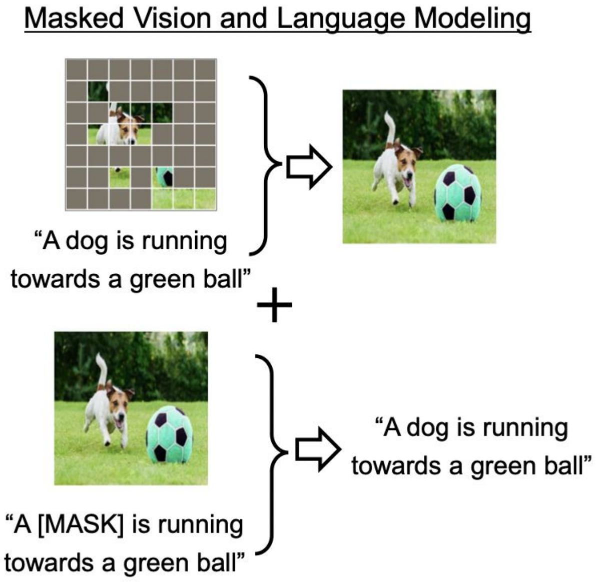


Masked Language Modeling

Source: RoBERTa (Liu et al., 2019)

Motivations

- Recent works like MaskVLM and BEiT-3 adopt Masked Image and Language Modeling in pretraining and achieve SOTA performance.

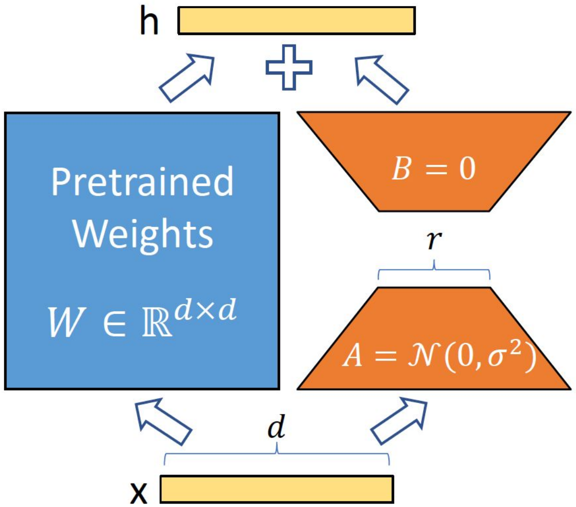


Source: MaskVLM (Kwon et al., 2022)

- Computation-expensive: Models like CLIP, ViLT have large scale parameters, making it difficult to fine-tune the entire model on specific downstream tasks.
- Can we gain better unimodal representation in a parameter-efficient way?

Motivations

- How to parameter-efficient fine-tune? With Adapter.
- Add a few trainable parameters on model. New tasks can be added without revisiting previous ones.

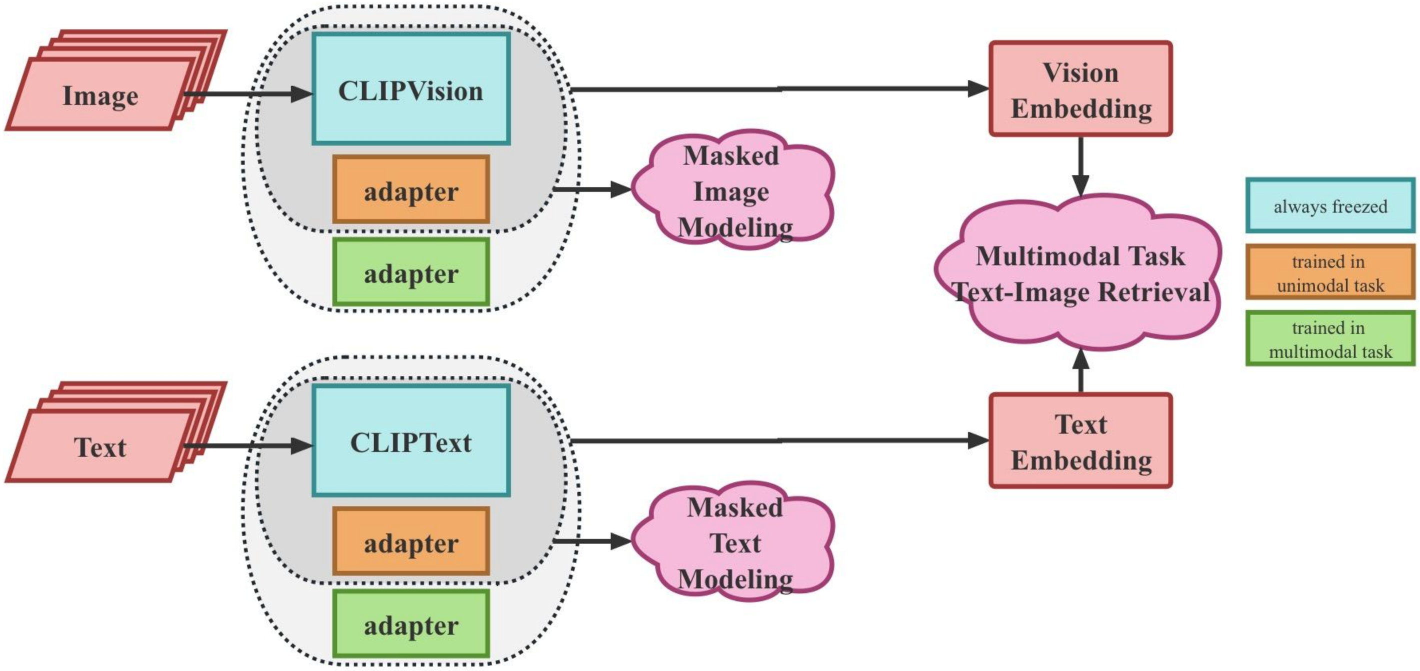


Source: LoRA
(Hu et al., 2021)

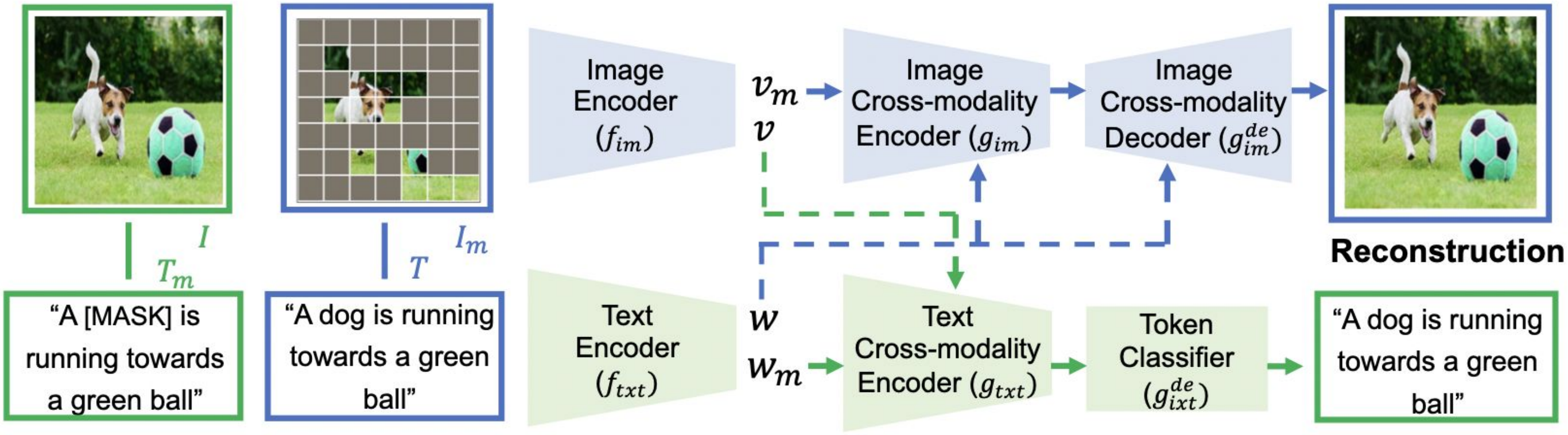
- Add Adapters in CLIP: Compared to contrastive V-L learning, it get a better representation in unimodality.
- Transfer: from trained unimodal adapter to fine-tune multi-modal tasks.

Overall Methods

Structure



USC



I
T_m
"A [MASK] is running towards a green ball"
I_m
T
"A dog is running towards a green ball"
Image Encoder (f_{im})
v_m
v
Image Cross-modality Encoder (g_{im})
Image Cross-modality Decoder (g_{im}^{de})
Reconstruction
Text Encoder (f_{txt})
w
w_m
Text Cross-modality Encoder (g_{txt})
Token Classifier (g_{txt}^{de})
"A dog is running towards a green ball"

MaskVLM

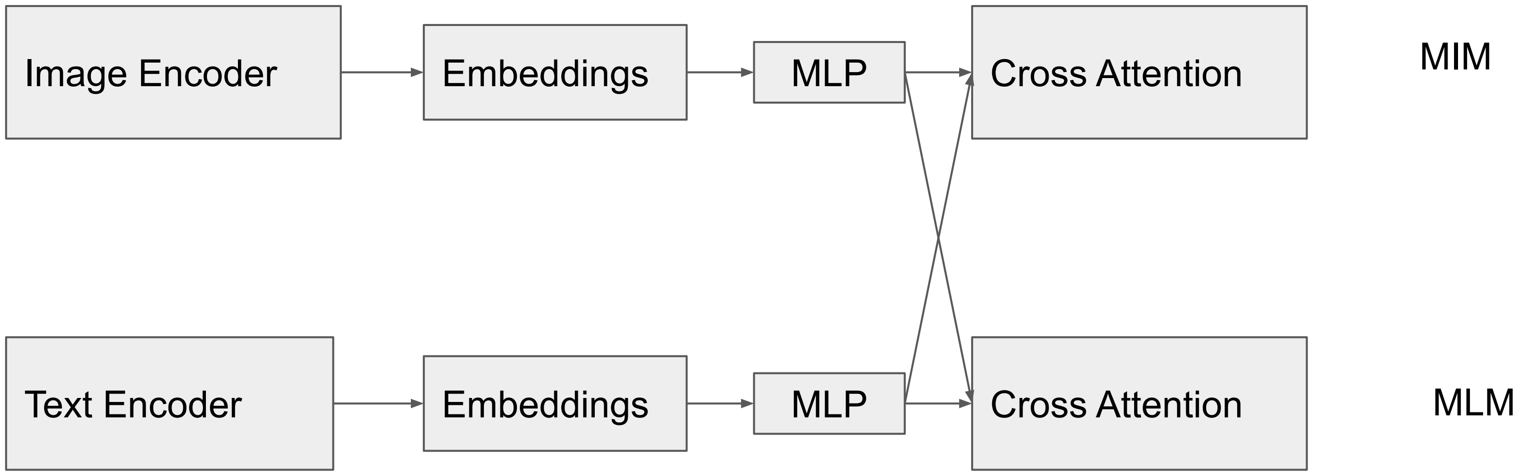
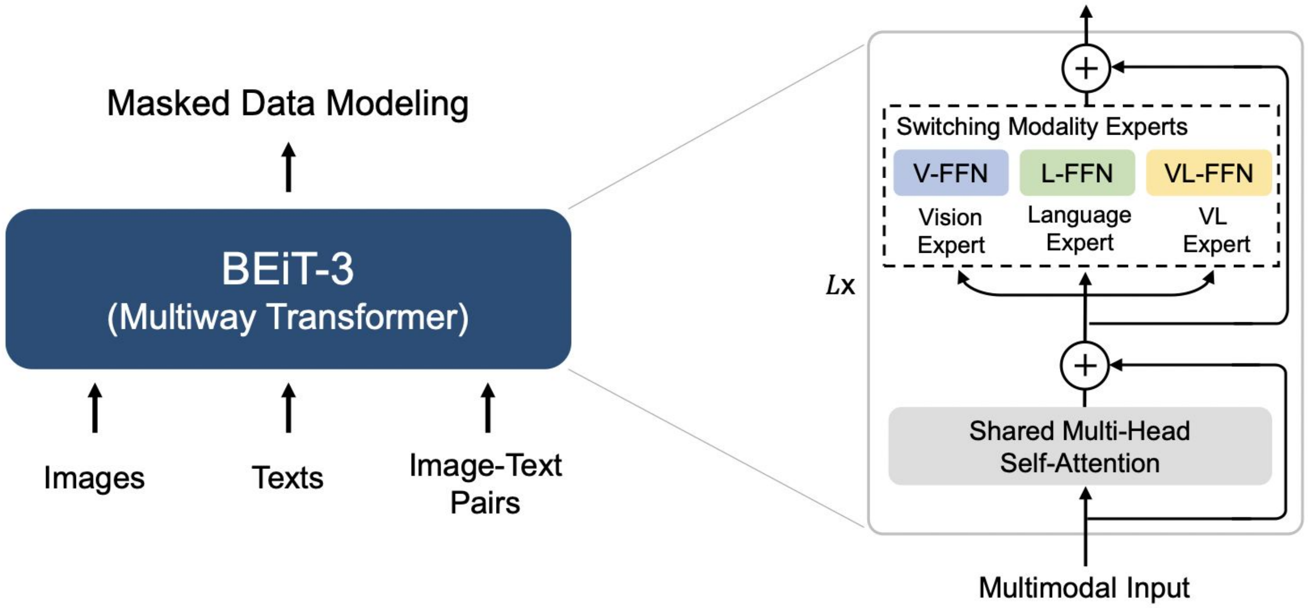
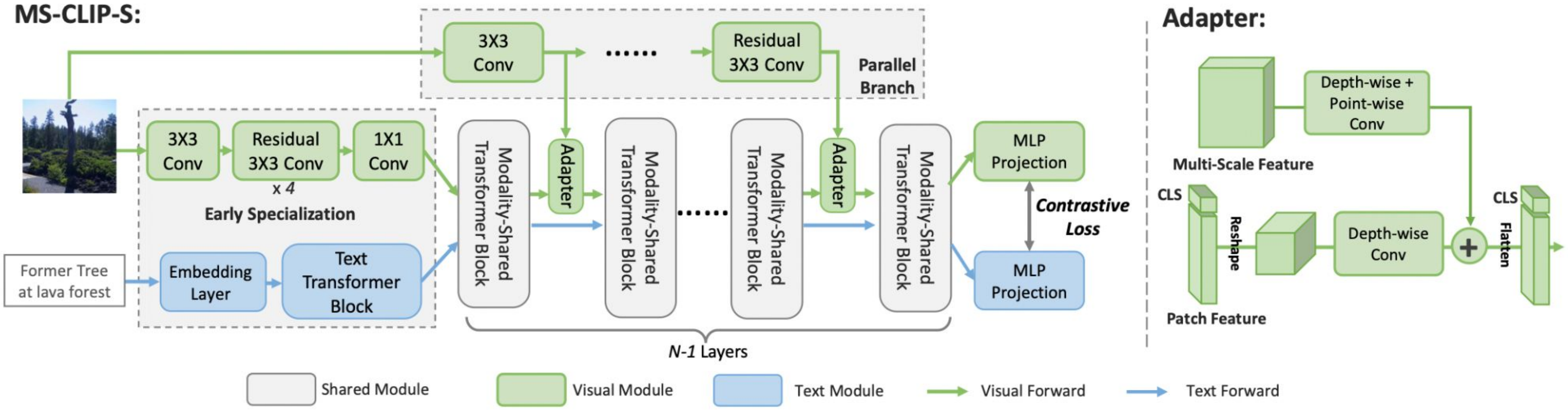


Image Encoder
Embeddings
MLP
Cross Attention
MIM
Text Encoder
Embeddings
MLP
Cross Attention
MLM

USC


Masked Data Modeling
BEiT-3
(Multiway Transformer)
Images
Texts
Image-Text Pairs
Switching Modality Experts
V-FFN
L-FFN
VL-FFN
Vision Expert
Language Expert
VL Expert
Lx
Shared Multi-Head Self-Attention
Multimodal Input

BEiT-3


MS-CLIP-S:
3X3 Conv
......
Residual 3X3 Conv
Parallel Branch
3X3 Conv
Residual 3X3 Conv
x 4
1X1 Conv
Early Specialization
Former Tree at lava forest
Embedding Layer
Text Transformer Block
Modality-Shared Transformer Block
Adapter
Modality-Shared Transformer Block
......
Modality-Shared Transformer Block
Adapter
Modality-Shared Transformer Block
N-1 Layers
MLP Projection
Contrastive Loss
MLP Projection
Adapter:
Depth-wise + Point-wise Conv
Multi-Scale Feature
CLS
Reshape
Depth-wise Conv
CLS
Flatten
Patch Feature
Shared Module
Visual Module
Text Module
Visual Forward
Text Forward

MS-CLIP

Methods

- To retain the alignment advantages of CLIP, our architecture adopts it globally and the pretrained weight is always frozen.
- To gain better unimodal representation, we add adapters on both CLIPVision and CLIPText model doing MIM and MLM.
- To re-align image and text embeddings, while preserving unimodal features, we freeze all current parameters and continue to add two new adapters for fine-tuning on the downstream task.

Experiments

Experiments

- Pretrained model: CLIP.
- Vision modality: Train LoRA Adapter + MIM on ImageNet-mini dataset.
- Language modality: Train MAM Adapter + MLM on Bookcorpus dataset.
- Multi-modality: Transfer these two trained V & L Adapters to Flickr-30k Image-Text retrieval dataset.
- Comparative Experiments: Adapters + MIM + MLM; Adapters + MIM; Adapters + MLM; Only Adapters; Full Fine-tune CLIP (baseline).

Trained on Unimodality

Vision: Adapter + MIM			
Dataset	Adapter	Mask Ratio	L1 Loss
ImageNet-mini	LoRA	0.6	0.371
Language: Adapter + MLM			
Dataset	Adapter	Mask Ratio	Acc
Bookcorpus	MAM	0.15	0.20

- Vision Adapter shows good performance on reconstruction loss.
- Language Adapter shows relatively low accuracy.

Transfer Adapters to Multi-modality

Method	# Fine-tuned Parameters	Text-to-Image Retrieval			Image-to-Text Retrieval		
		R@1	R@5	R@10	R@1	R@5	R@10
Full Finetune	149M	78.42	94.98	97.66	92.4	98.7	99.6
CLIP+Adapter	12M	78.36	94.81	97.61	91.4	**99.2**	**99.8**
CLIP+Adapter +MIM+MLM	14M	**78.92**	**95.26**	**97.75**	**92.6**	99.0	**99.8**

- Compared to fine-tuning CLIP, Adapters can get the similar performance.
- Our approach outperforms directly fine-tuning with Adapters.

Masked Modeling Comparison

Method	Text-to-Image Retrieval			Image-to-Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Full Finetune	78.42	94.98	97.66	92.4	98.7	99.6
CLIP+Adapter +MIM	78.70	94.84	97.69	92.3	99.1	**99.8**
CLIP+Adapter +MLM	**79.1**	**95.31**	97.74	91.6	**99.3**	**99.8**
CLIP+Adapter +MIM+MLM	78.92	95.26	**97.75**	**92.6**	99.0	**99.8**

- Apply both MIM & MLM on Adapters achieves relatively better performance.

Future Work

Future Work

- Take a chance of one-stage method
  - Directly combine MIM and MLM into our current structure.
  - Find decent hyper-parameters to balance losses of MIM, MLM and alignment
- Gain better unimodal representation
  - MLM only get 20% accuracy in BookCorpus
  - More complex dataset, better data cleaning and preprocessing.
  - Try some other pretrain tasks like classification, object detection to get more robust representation.
- Better vision-text alignment
  - Try our method on a more unified model like ViLT / shared weight encoders
- More comparative experiments on downstream tasks using other datasets like COCO

Thanks for Your Time!

CSCI-566: Deep Learning and Its Applications

Team: DODO Bird

Apr 21, 2023