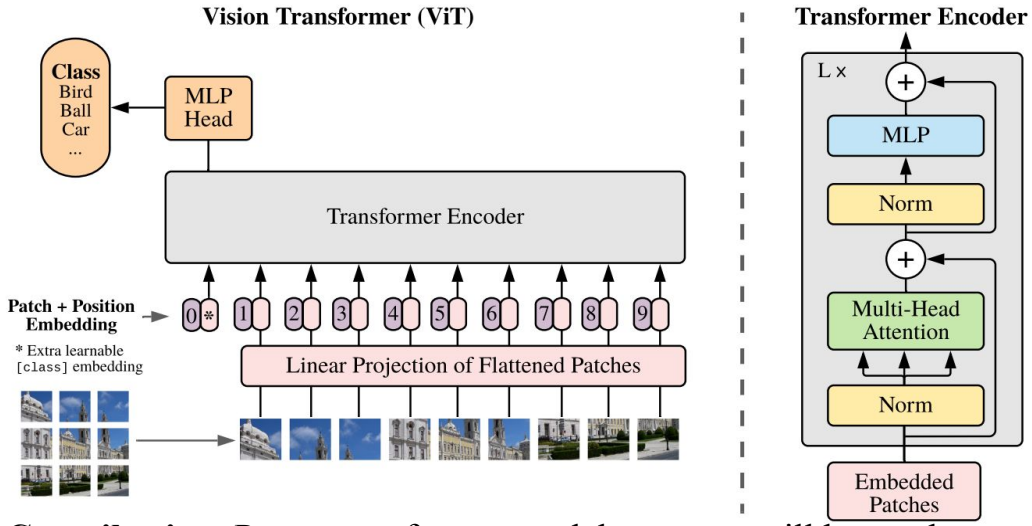
CV Pre-trained Model - Visual Transformer



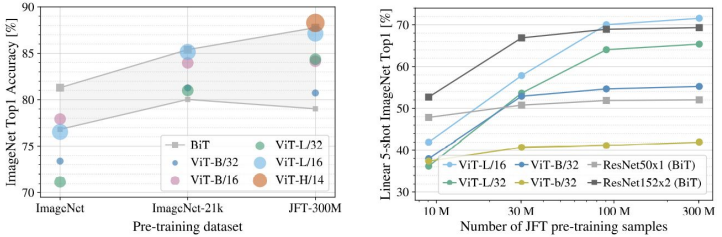


Fig 2. ViT vs ResNet on transfer learning and 5-shot

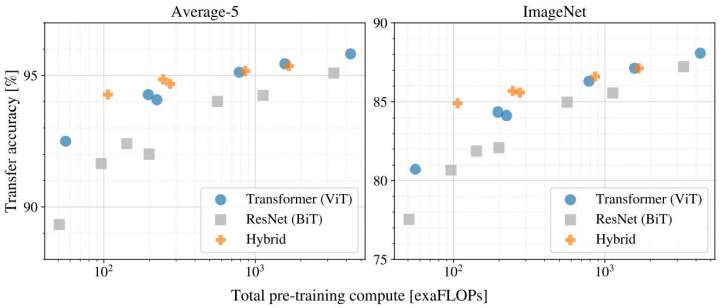


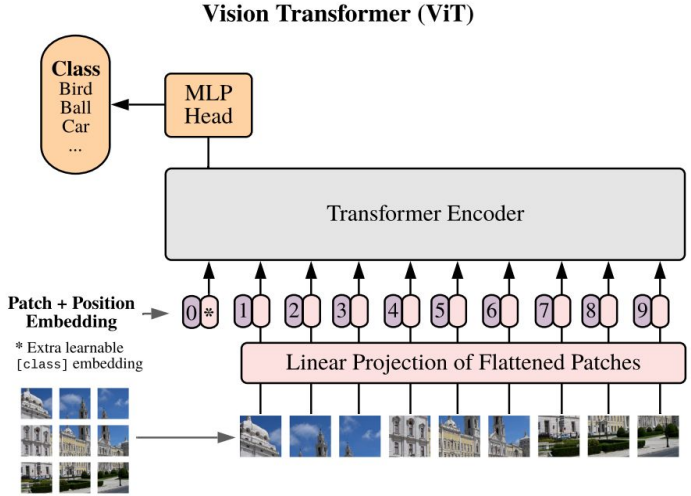
Fig 3. ViT, ResNet, Hybrid performance comparison

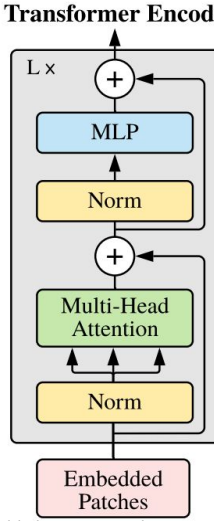
Contribution: Prove transformer model structure will better than CNN in huge data situation

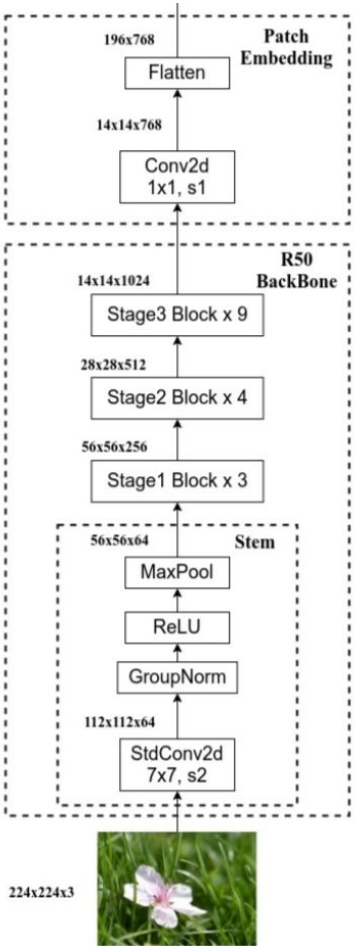
1. Resize input image to 224x224x3
2. Split into 196 patches with size 16x16x3
3. Flatten image to sequence length 196 with hidden size 768
4. Add position encoding
5. Concat [cls] token in front of the input sequence

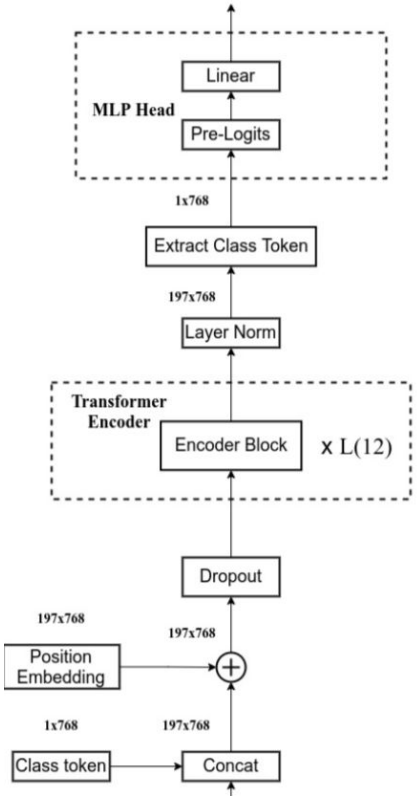
[ViT] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations. 2020.

CV Pre-trained Model - Visual Trans









Contribution: Prove transformer model structure will better than CNN in huge data situation

1. Resize input image to 224x224x3
2. Split into 196 patches with size 16x16x3
3. Flatten image to sequence length 196 with hidden size 768
4. Add position encoding
5. Concat [cls] token in front of the input sequence

[ViT] Dosovitskiy, Alexey, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." International Conference on Learning Representations. 2020.

CV Pre-trained Model - BEIT

Contribution: Apply MIM(Masked Image Model) task for self-supervised learning

1. Train d-VAE and restore the image (task1)
2. Apply d-VAE to get visual tokens
3. Patch and mask then follow ViT
4. Let BEIT predict the visual token (task2)

Models	Model Size	Labeled Data Size	ImageNet 384^2	ImageNet 512^2
Supervised Pre-Training on ImageNet-22K (using labeled data)				
ViT-B [DBK+20]	86M	14M	84.0	-
ViT-L [DBK+20]	307M	14M	85.2	85.30
ViT-H [DBK+20]	632M	14M	85.1	-
Supervised Pre-Training on Google JFT-300M (using labeled data)				
ViT-B [DBK+20]	86M	300M	84.2	-
ViT-L [DBK+20]	307M	300M	87.1	87.76
ViT-H [DBK+20]	632M	300M	88.0	88.55
Supervised Pre-Training on Google JFT-3B (using labeled data)				
ViT-B [ZKHB21]	86M	3000M	86.6	-
ViT-L [ZKHB21]	307M	3000M	88.5	-
Self-Supervised Pre-Training, and Intermediate Fine-Tuning on ImageNet-22K				
BEiT-B+ (ours)	86M	14M	86.8	-
BEiT-L+ (ours)	307M	14M	88.4	88.6
Self-Supervised Pre-Training, and Intermediate Fine-Tuning on In-House-70M				
BEiT-L+ (ours)	307M	70M	**89.3**	**89.5**

Fig 2. ViT vs ResNet on transfer learning and 5-shot

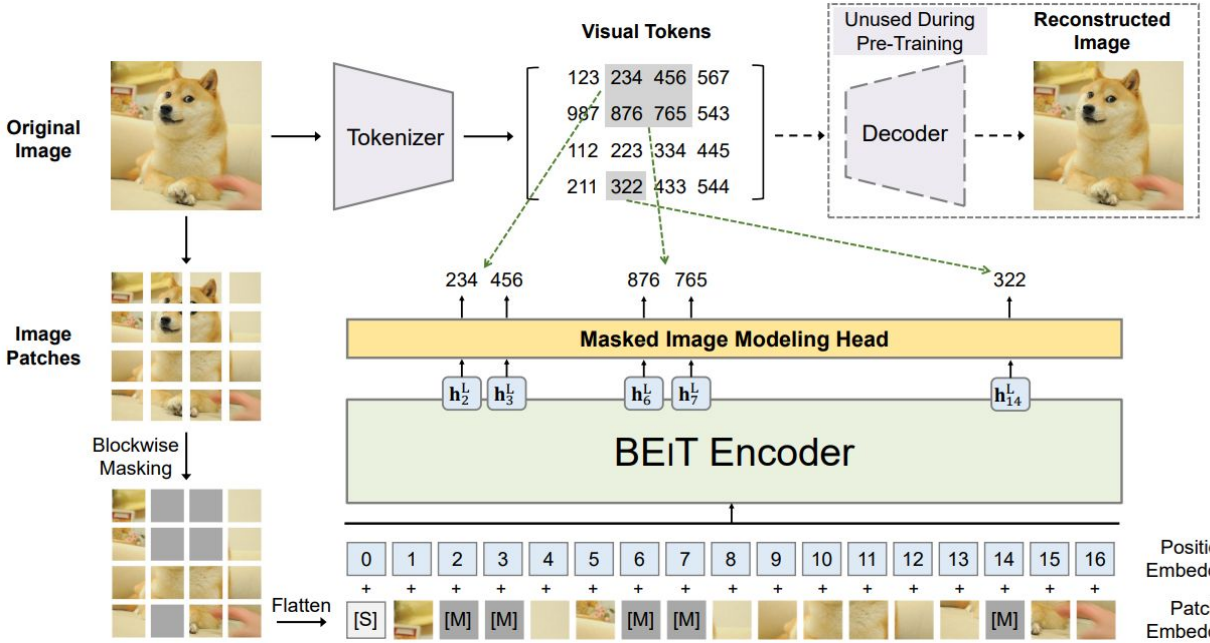


Fig 1. Model Structure

[BEIT] Bao, Hangbo, et al. "BEiT: BERT Pre-Training of Image Transformers." International Conference on Learning Representations. 2021.

CV Pre-trained Model - MAE

Contribution: MIM task without any visual token

1. Drop 75% patches
2. Shuffle remained 25% patches and input to Encoder
3. Reshuffle the Encoder output
4. Decode to restore the original image

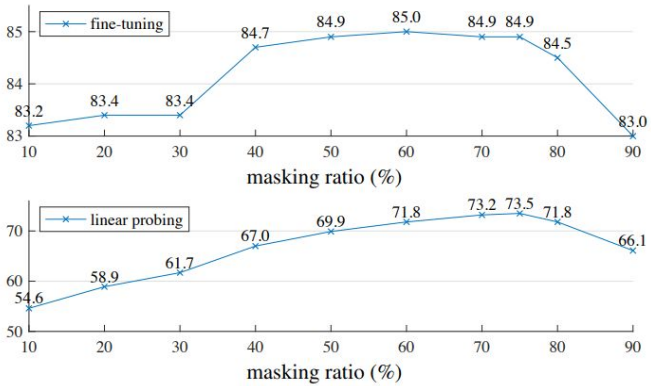


Fig 2. Mask about 75% image having the best performance

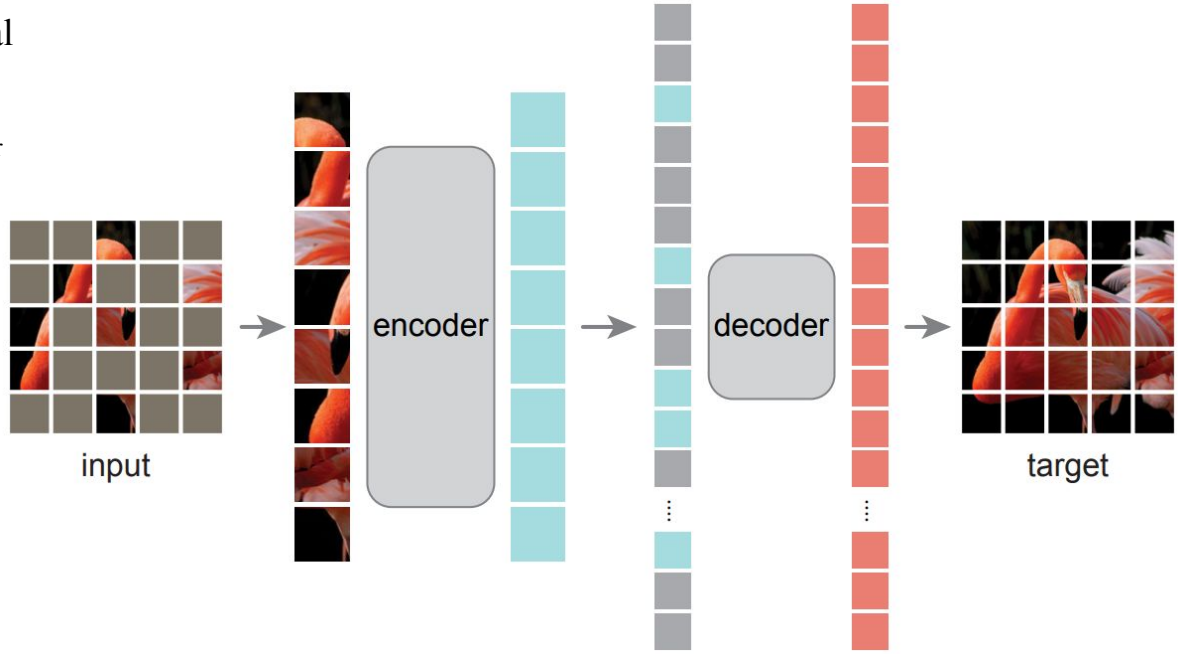


Fig 1. Asymmetric model structure (Encoder > Decoder)

[MAE] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

VL Model - CLIP

Contribution: Mapping text and image embedding to the same space

1. Collect many image-text pair from the Net
2. Separately encode the text and image
3. Image and text in the same pair as positive samples
4. Do contrastive learning

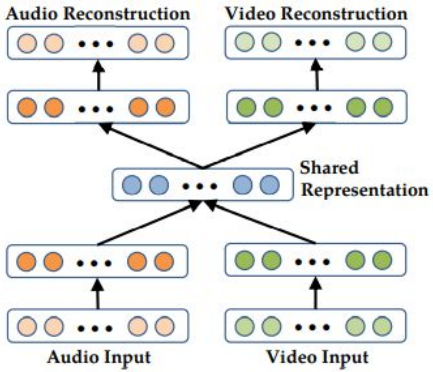


Fig 3. Bimodal Deep Autoencoder

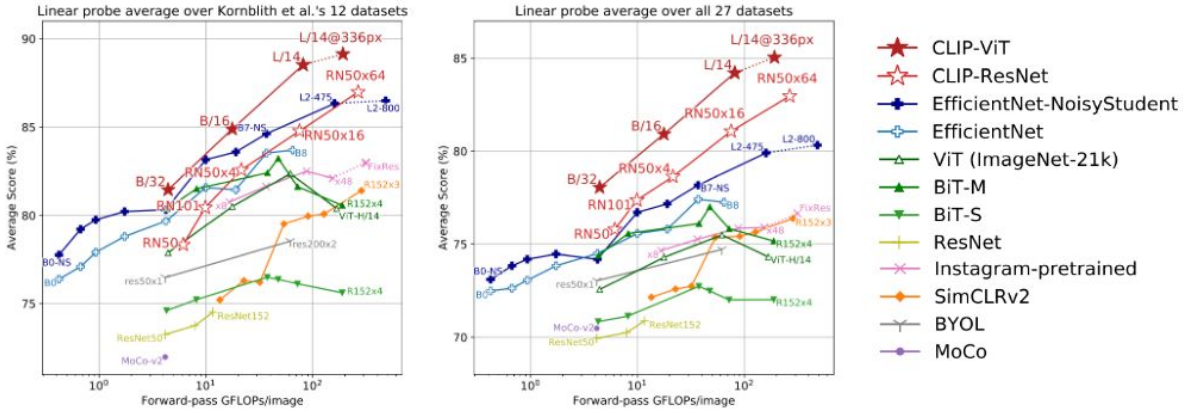


Fig 2. Performance of CLIP

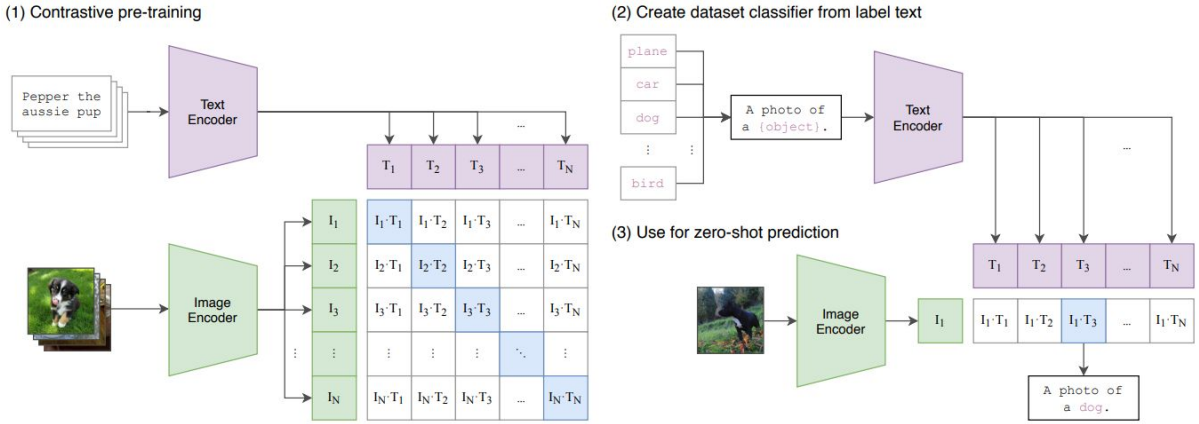


Fig 1. Training and inference of CLIP

[Bimodal] Ngiam, Jiquan, et al. "Multimodal deep learning." ICML. 2011.
[CLIP] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning. PMLR, 2021.
[distill] Gabriel Goh, et al. “Multimodal Neurons in Artificial Neural Networks” https://distill.pub/2021/multimodal-neurons/

VL Pre-trained Model - UNITER

Contribution: VL pre-trained model by 4 different tasks

1. Separately extract image(Faster R-CNN) and text(BERT) feature
2. Concat image and text feature as input for Transformer
3. Do following tasks
   1) Masked Language Model, MLM
   2) Mask Region Model, MRM
   3) Image Text Matching, ITM
   4) Word Region Alignment, WRA

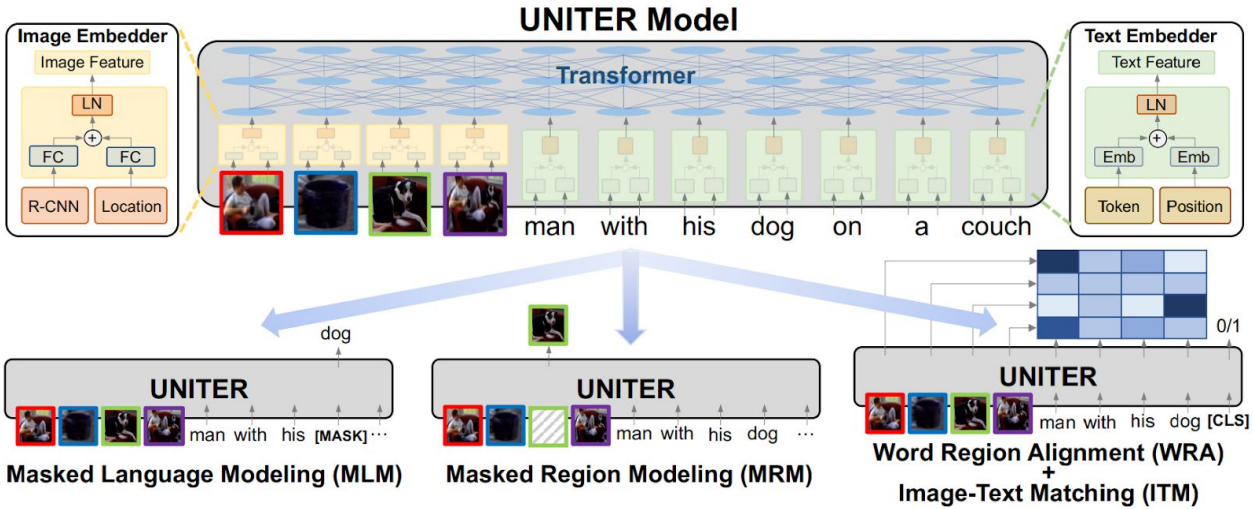


Fig 1. Training and inference of CLIP

[UNITER] Chen, Yen-Chun, et al. "Uniter: Universal image-text representation learning." European conference on computer vision. Springer, Cham, 2020.

VL Pre-trained Model - ViLT

Contribution: 1. Simplify the pre-training tasks 2. promote the training and inference speed 3. bypass region feature extractor

1. Embed the input text
2. Linear project the image patches
3. Concat [Class] token in text and image input
4. Add position encoding and modal-type encoding
5. Do ITM, MLM(wwm), WPA

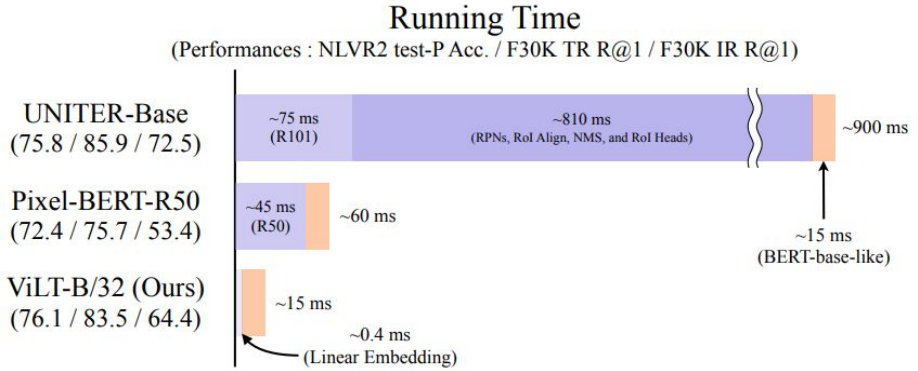


Fig 2. Performance comparison

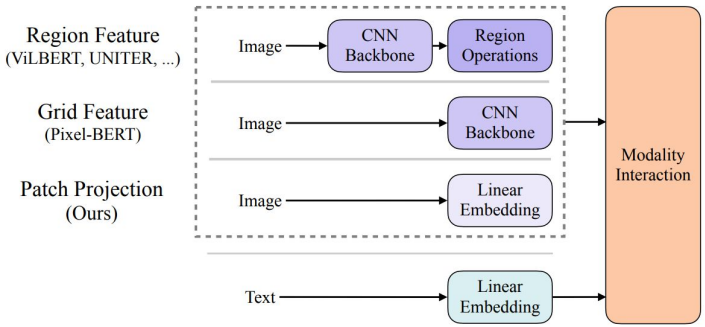


Fig 3. Different visual embedding schema

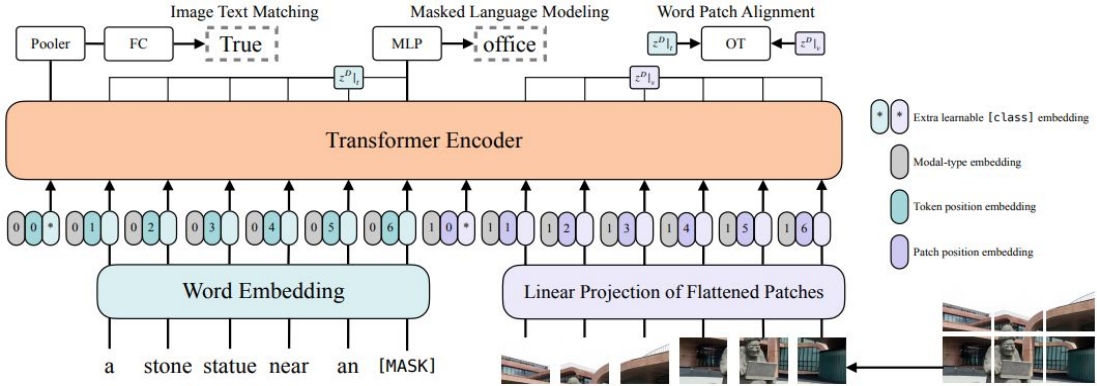


Fig 1. Input format and model structure

[ViLT] Kim, Wonjae, Bokyung Son, and Ildoo Kim. "Vilt: Vision-and-language transformer without convolution or region supervision." International Conference on Machine Learning. PMLR, 2021.

VL Pre-trained Model - CoCa

Contribution: Combine contrastive learning with as pre-training tasks to enhance zero-shot

1. PrefixLM as text embedder
2. ViT/ResNet as image encoder
3. Do contrastive learning
4. Do captioning task

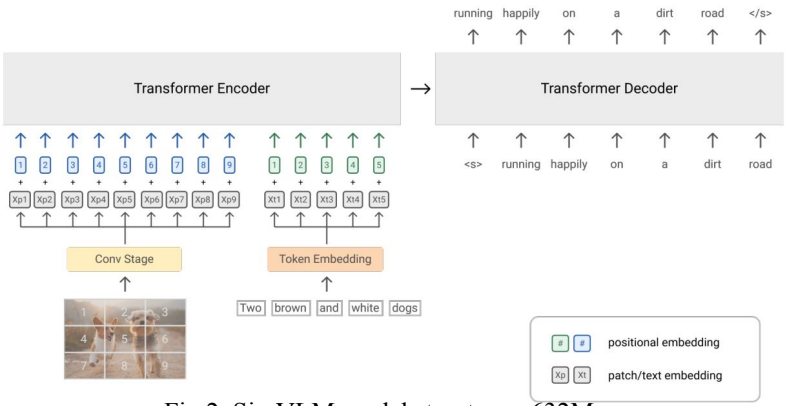


Fig 2. SimVLM model structure - 632M

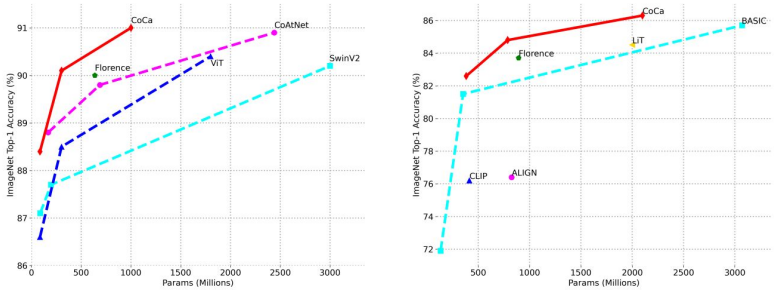


Fig 3. Performance on image classification

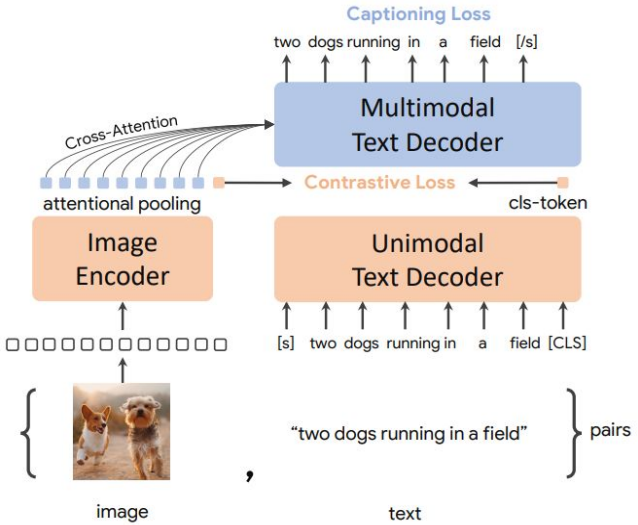


Fig 1. CoCa model structure - 2.1B

[SimVLM] Wang, Zirui, et al. "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision." International Conference on Learning Representations. 2022.
[CoCa] Yu, Jiahui, et al. "Coca: Contrastive captioners are image-text foundation models." arXiv preprint arXiv:2205.01917 (2022).