Multimodal Papers

Traditional multimodal tasks:

Image-Text Retrieval: image-to-text retrieval or text-to-image retrieval, evaluated by recall (R@1, R@5, R@10).

Visual Question Answering (VQA): to predict an answer given an image and a question; choose one answer (classification) or generate one (generation).

Visual Reasoning (VR): to predict whether a text describes a pair of images (classification).

Visual Entailment (VE): to predict whether the relationship between an image and a text is entailment, neutral, or contradictor (classification).

More modalities: speech, video, structured knowledge…

ViLT & **CLIP** see separate notes

ALBEF: **Align before Fuse: Vision and Language Representation Learning with Momentum Distillation** (2021) 289

VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts (2022) 65

BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (2022) 92

CoCa: Contrastive Captioners are Image-Text Foundation Models (2022) 118

BeiTv3: **Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks** (2022) 46

Align before Fuse: Vision and Language Representation Learning with Momentum Distillation

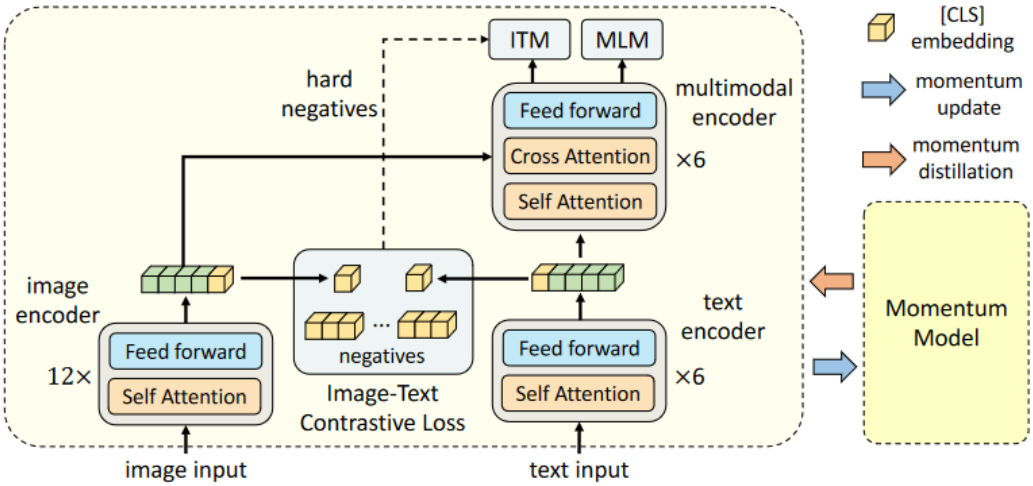


Figure 1: **Illustration of ALBEF.** It consists of an image encoder, a text encoder, and a multimodal encoder. We propose an image-text contrastive loss to align the unimodal representations of an image-text pair before fusion. An image-text matching loss (using in-batch hard negatives mined through contrastive similarity) and a masked-language-modeling loss are applied to learn multimodal interactions between image and text. In order to improve learning with noisy data, we generate pseudo-targets using the momentum model (a moving-average version of the base model) as additional supervision during training.

Image: ViT & Text: BERT → Align → Multimodal fusion

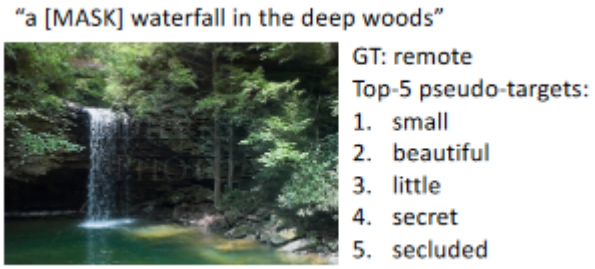
Align:

Image-Text Contrastive Loss: positive & hard negative (highest cosine similarity among others)

Momentum Distillation:

To address the problem of noisy image-text pairs collected from the web

Train the model such that its predictions not only match the ground truth one-hot label, but also match the pseudo-targets generated by the momentum model (exponential-moving-average)



VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts

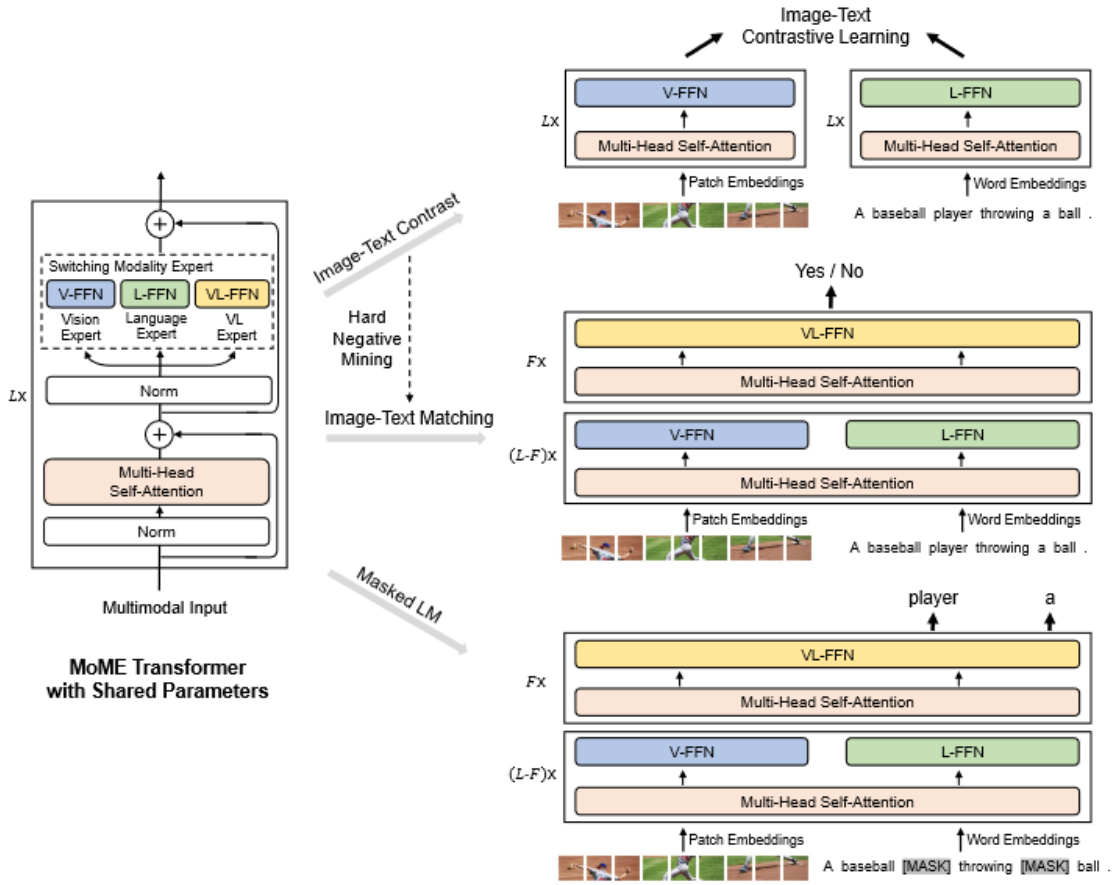


Figure 1: Overview of VLMO pre-training. We introduce mixture-of-modality-experts (MOME) Transformer to encode different modality input by modality-specific experts. The model parameters are shared across image-text contrastive learning, masked language modeling, and image-text matching pre-training tasks. During fine-tuning, the flexible modeling enables us to use VLMO as either a dual encoder (i.e., separately encode images and text for retrieval tasks) or a fusion encoder (i.e., jointly encode image-text pairs for better interaction across modalities).

Image encoder & text encoder → dual encoder

Multimodal fusion → fusion encoder

Mixture-of-modality-experts: jointly learns a dual encoder and a fusion encoder with a modular Transformer network

Self-attention layer: shared weights

FFN layer: vision expert, language expert, vision-language expert

BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

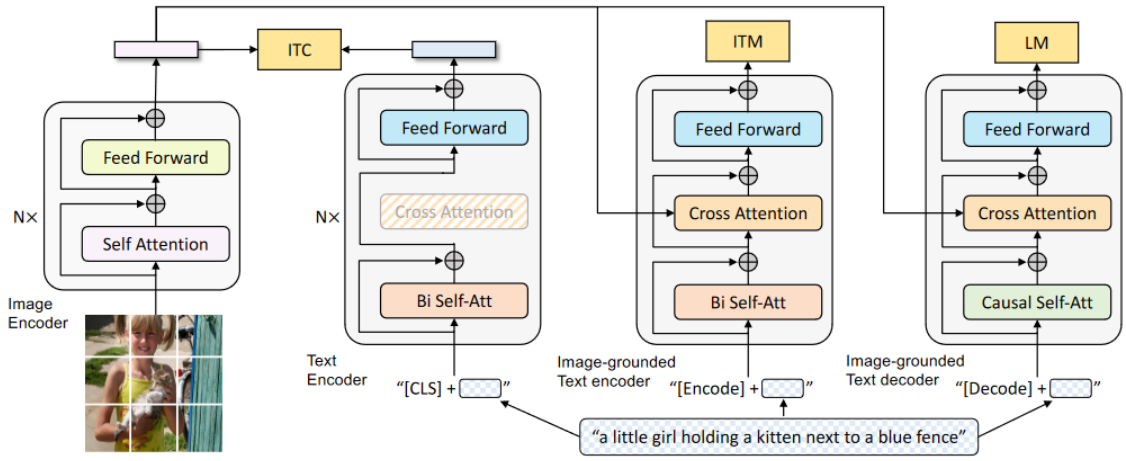


Figure 2. Pre-training model architecture and objectives of BLIP (same parameters have the same color). We propose multimodal mixture of encoder-decoder, a unified vision-language model which can operate in one of the three functionalities: (1) Unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations. (2) Image-grounded text encoder uses additional cross-attention layers to model vision-language interactions, and is trained with a image-text matching (ITM) loss to distinguish between positive and negative image-text pairs. (3) Image-grounded text decoder replaces the bi-directional self-attention layers with causal self-attention layers, and shares the same cross-attention layers and feed forward networks as the encoder. The decoder is trained with a language modeling (LM) loss to generate captions given images.

Model perspective — **Unified Vision-Language Understanding and Generation**

Mixture of encoder-decoder

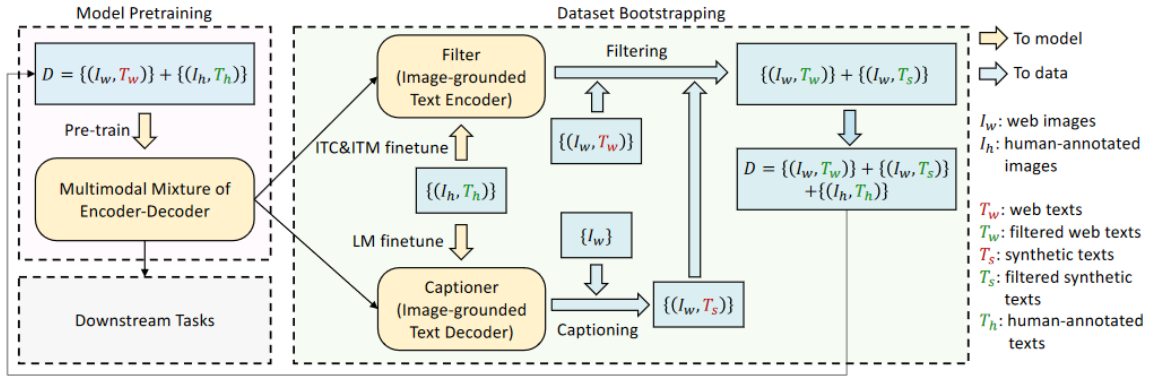


Figure 3. Learning framework of BLIP. We introduce a captioner to produce synthetic captions for web images, and a filter to remove noisy image-text pairs. The captioner and filter are initialized from the same pre-trained model and finetuned individually on a small-scale human-annotated dataset. The bootstrapped dataset is used to pre-train a new model.

Data perspective — **Bootstrapping**

Captioner: generate synthetic captions for images

Filter: remove noisy captions (remain match text-image pairs)

Accuracy: original 78.4%, +filtering 79.1%, +captioning 79.7%, +captioning & filtering 80.6%

General tool to get larger and higher-quality dataset

CoCa: Contrastive Captioners are Image-Text Foundation Models

Larger pre-training dataset → state-of-the-art performance on many tasks

Losses:

- contrastive loss between unimodal image and text embeddings
- captioning loss on the multimodal decoder outputs which predicts text tokens autoregressively

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

Background:

The big convergence of language, vision, and multimodal pretraining: By performing large-scale pretraining on massive data, we can easily transfer the models to various downstream tasks.

The success of transformers: dual-encoder architecture (CLIP) for efficient retrieval, encoder-decoder networks for generation tasks (BLIP, CoCa), fusion-encoder architecture for image-text encoding (ALBEF, VLMo)

General-purpose multimodal foundation model BEIT-3, regarding Image as a Foreign Language: images (Imglish), texts (English), and image-text pairs (“parallel sentences”)

Single model: Multiway Transformers

Single objective function: Masked data modeling

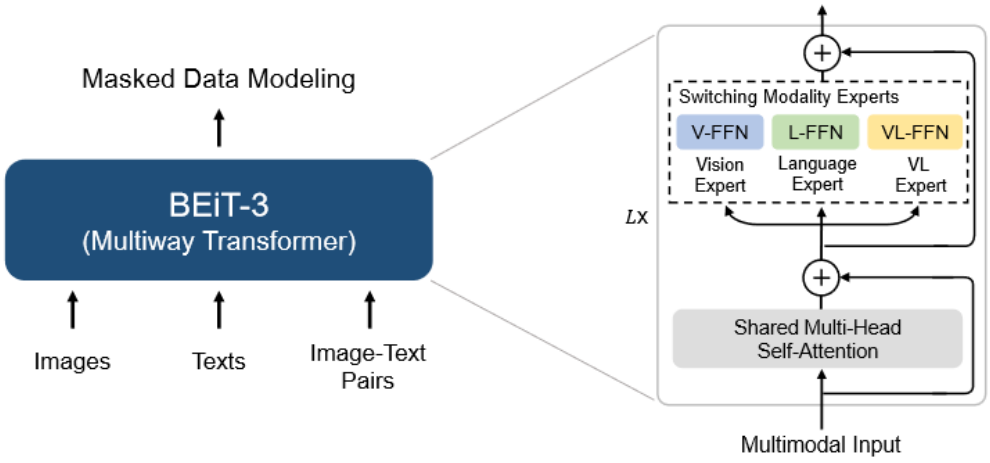


Figure 2: Overview of BEIT-3 pretraining. We perform masked data modeling on monomodal (i.e., images, and texts) and multimodal (i.e., image-text pairs) data with a shared Multiway Transformer as the backbone network.

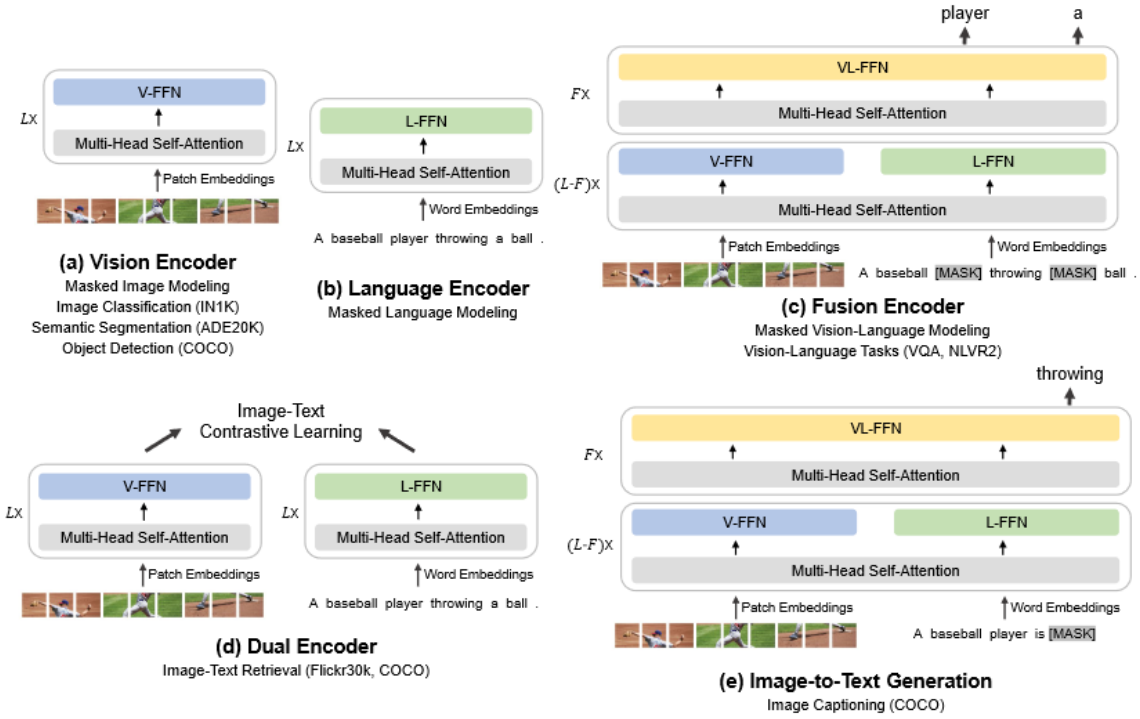


Figure 3: BEIT-3 can be transferred to various vision and vision-language downstream tasks. With a shared Multiway Transformer, we can reuse the model as (a)(b) vision or language encoders; (c) fusion encoders that jointly encode image-text pairs for deep interaction; (d) dual encoders that separately encode modalities for efficient retrieval; (e) sequence-to-sequence learning for image-to-text generation.