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

Hangbo Bao,* Wenhui Wang,* Li Dong, Qiang Liu
Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Furu Wei†
Microsoft
https://aka.ms/vlmo

arXiv'21

221018 Chia-Chun Ho

Outline
of LIIRM

Introduction

Methodology

Experiments

Conclusion

Comments

Introduction

Vision-Language Pre-training (VLP)

- VLP learns generic cross-modal representations from large-scale image-text pairs.
- Two mainstream architectures are widely used in previous work.
  - *Dual encoder* to encode images and text separately.
    - However, the shallow interaction between images and text is not enough to handle complex VL classification tasks.
  - *Fusion encoder* with cross-modal attention to model image-text pairs.
    - The fusion-encoder architecture achieves superior performance on VL classification tasks.
    - But it requires to jointly encode all possible image-text pairs to compute similarity scores for retrieval tasks.

Introduction

Vision-Language pretrained Model (VLMo)

- Proposed VLMo that can be used as either
  - a dual encoder to separately encode images and text for retrieval tasks,
  - or used as a fusion encoder to model the deep interaction of image-text pair for classification tasks.
- This's achieved by introducing Mixture-of-Modality-Experts (MoME) Transformer that can encode various modalities (image, text, and image-text pairs) within a Transformer block.

Introduction

Mixture-of-Modality-Experts (MoME)

- MoME employs a pool of modality experts to replace the feed-forward network in standard Transformer.
- It captures modality-specific information by switching to different modality experts, and use the shared self-attention across modalities to align visual and linguistic information.
- MoMe Transformer consists of three modality experts (vision, language, vision-language).
  - Thanks to the modeling flexibility, that can reuse MOME Transformer with the shared parameters for different purposes, i.e., text-only encoder, image-only encoder, and image-text fusion encoder.

Introduction

Pre-training Tasks

- VLMo is jointly learned with three pre-training tasks:
  - Image-text contrastive learning
  - Image-text matching
  - Masked language modeling
- In addition, propose a stagewise pre-training strategy to effectively leverage large-scale image-only and text-only corpus besides image-text pairs in VLMo pre-training.
  - It helps VLMo to learn more generalizable representations.

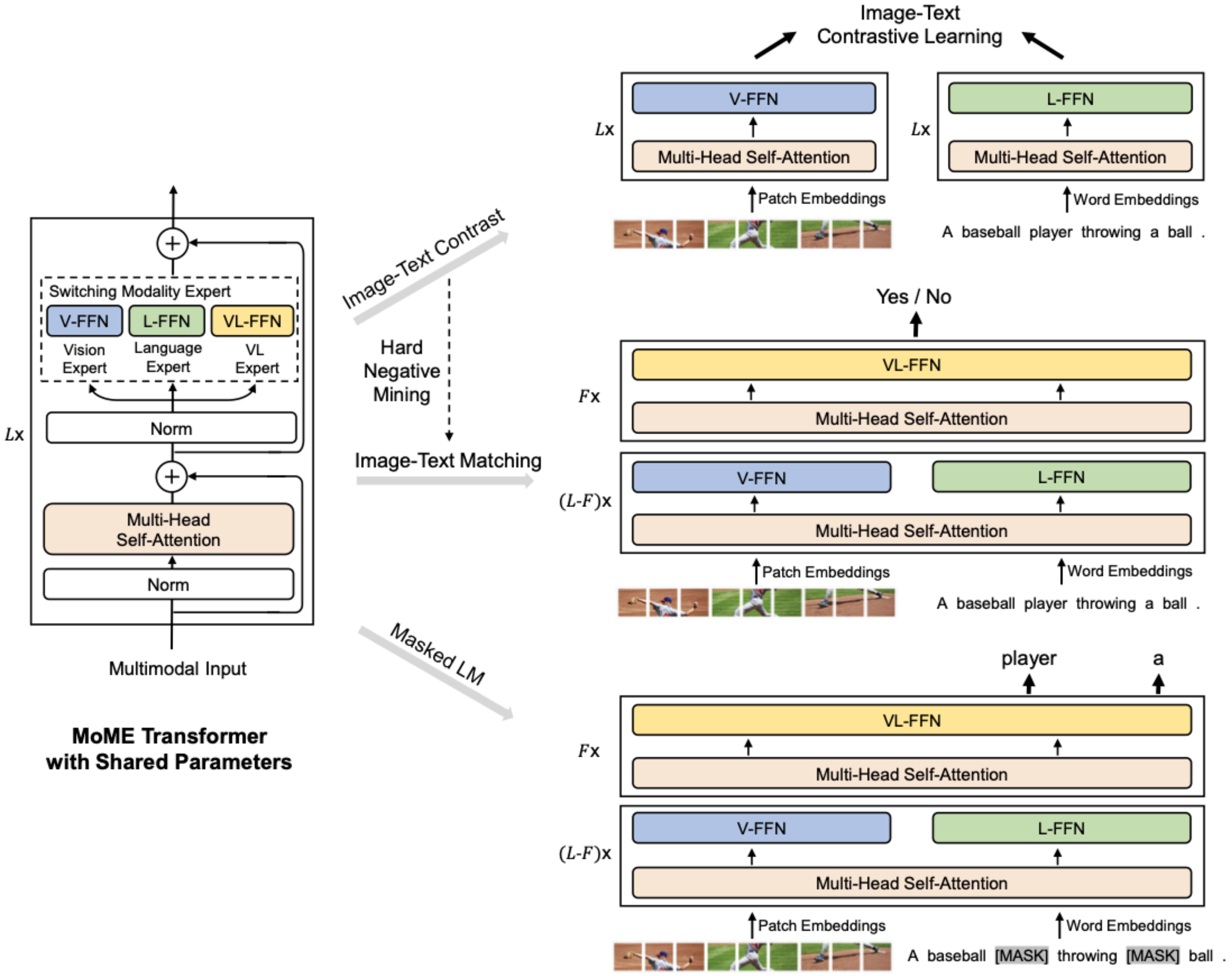
Introduction

Contributions

- Propose a unified vision-language pretrained model VLMo that can be used as a fusion encoder for classification tasks, or fine-tuned as a dual encoder for retrieval tasks.
- Introduce a general-purpose multimodal Transformer for vision-language tasks, namely MoME Transformer, to encode different modalities.
  - It captures modality-specific information by modality experts, and aligns contents of different modalities by the self-attention module shared across modalities.
- Showing that stagewise pre-training using large amounts of image-only and text-only data greatly improves our vision-language pretrained model.

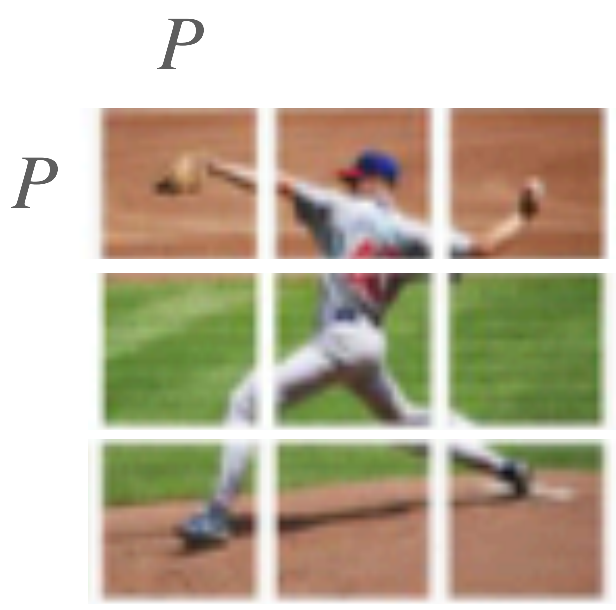
Methodology

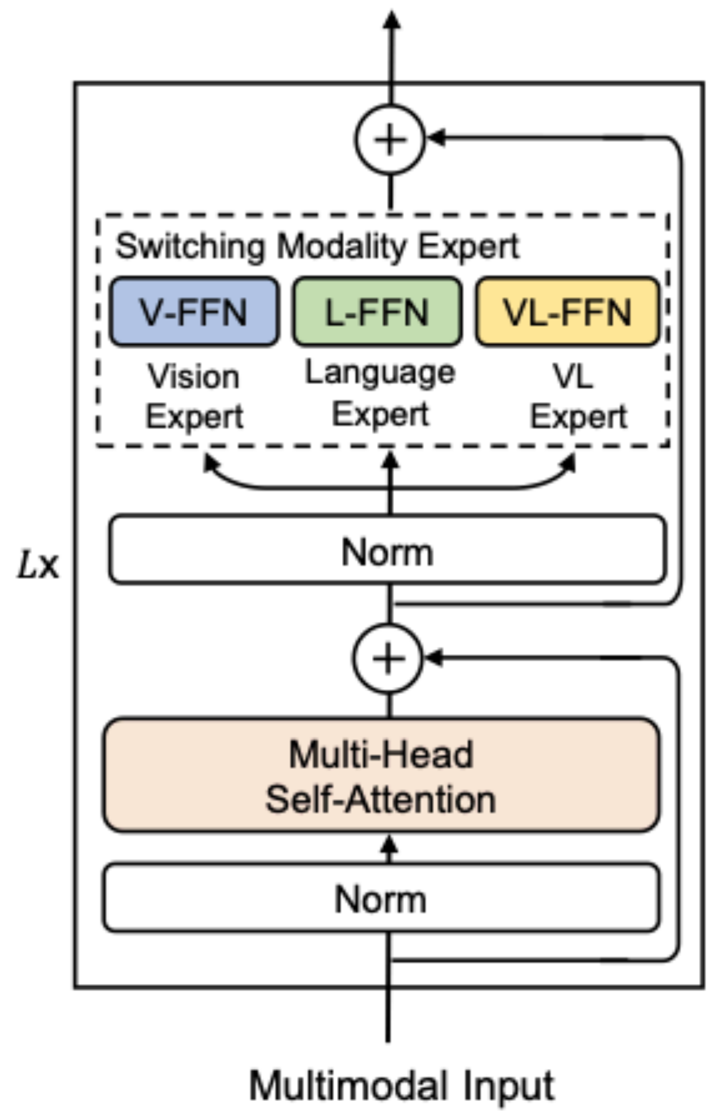
Proposed model



Methodology

Input Representations

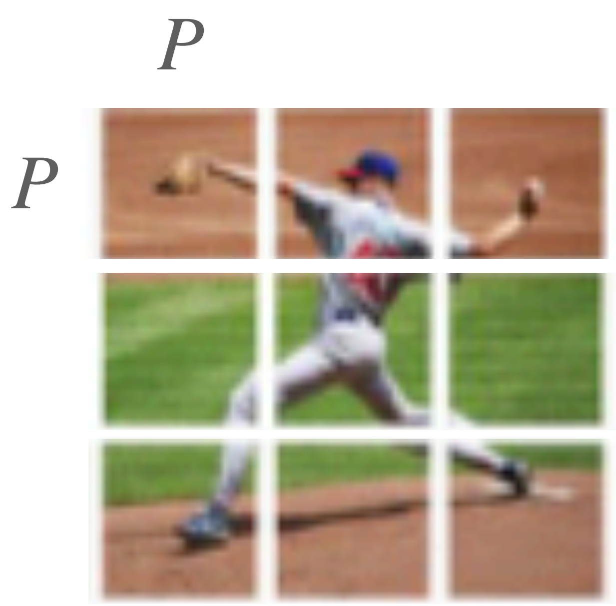


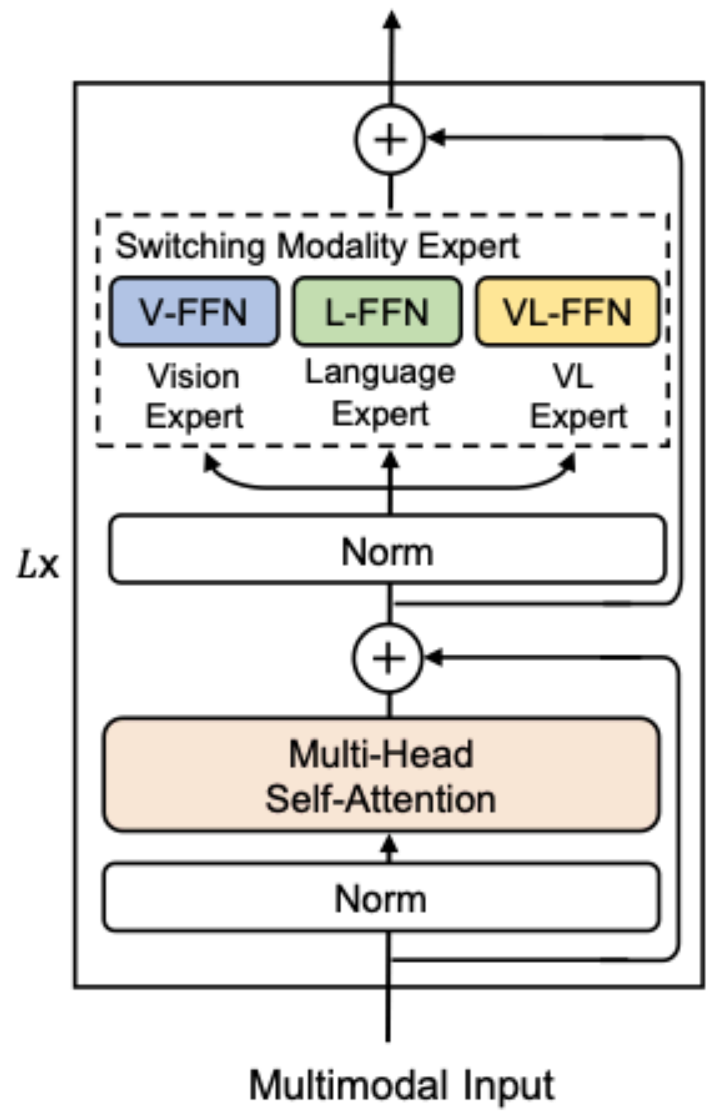


- Image Representations
  - 2D image $\boldsymbol{v} \in \mathbb{R}^{H \times W \times C}$ is split and reshaped into $N = NW/P^2$ patches $\boldsymbol{v}^p \in \mathbb{R}^{N \times (P^2 C)}$.
    - C: # of channels, (H, W): resolution of the input image, (P, P): patch resolution
  - The image patches are then flattened into vectors and are linearly projected to obtain patch embeddings, and also prepend a learnable special token $[I\_CLS]$ to the sequence.
  - Finally, image input representations are obtained via summing patch embeddings, learnable 1D position embeddings $V_{pos} \in \mathbb{R}^{(N+1) \times D}$ and image type embedding $V_{type} \in \mathbb{R}^{D}$
    - $\boldsymbol{H}_0^v = [\boldsymbol{v}_{[I\_CLS]}, \boldsymbol{V}\boldsymbol{v}_i^p, \ldots, \boldsymbol{V}\boldsymbol{v}_N^p] + \boldsymbol{V}_{pos} + \boldsymbol{V}_{type}$
    - $\boldsymbol{H}_0^v \in \mathbb{R}^{(N+1) \times D}$, linear projection $\boldsymbol{V} \in \mathbb{R}^{(P^2 C) \times D}$

Methodology

Input Representations



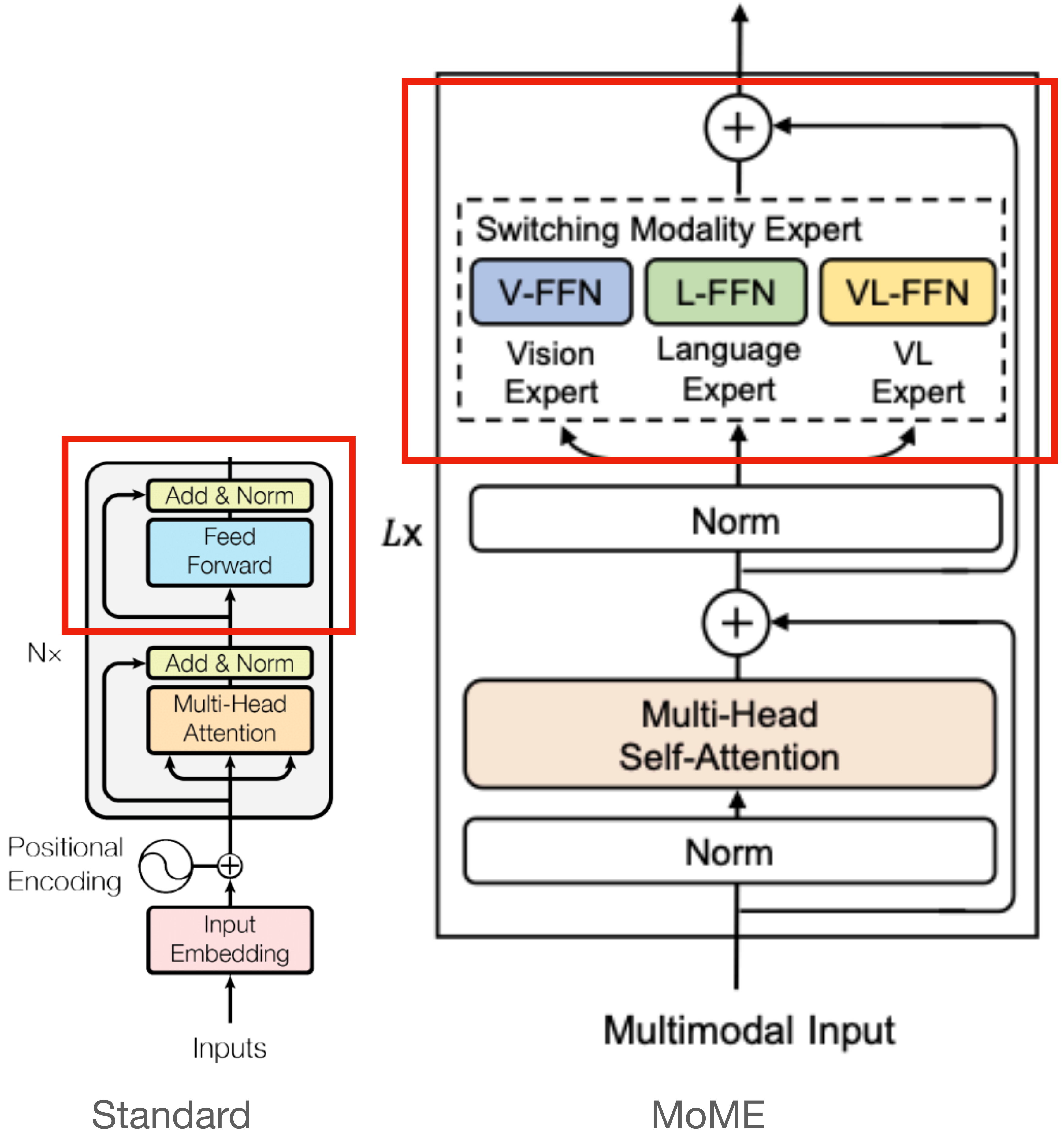


- Text Representation
  - Following BERT, tokenize the text to subword units by Word-Piece.
  - Add $[T\_CLS]$ & $[T\_SEP]$ to the text sequence.
  - Text input representation $\boldsymbol{H}_0^w \in \mathbb{R}^{(M+2)\times D}$ are computed via summing the corresponding word embedding, text position embedding and text type embedding.
    - $\boldsymbol{H}_0^w = [\boldsymbol{w}_{[T\_CLS]}, \boldsymbol{w}_i, \ldots, \boldsymbol{w}_M, \boldsymbol{w}_{[T\_SEP]}] + \boldsymbol{T}_{pos} + \boldsymbol{T}_{type}$, M: length of tokenized subword units
- Image-Text Representation
  - Concatenate image & text input vectors to form the image-text input representations $\boldsymbol{H}_0^{vl} = [\boldsymbol{H}_0^w; \boldsymbol{H}_0^v]$

Methodology

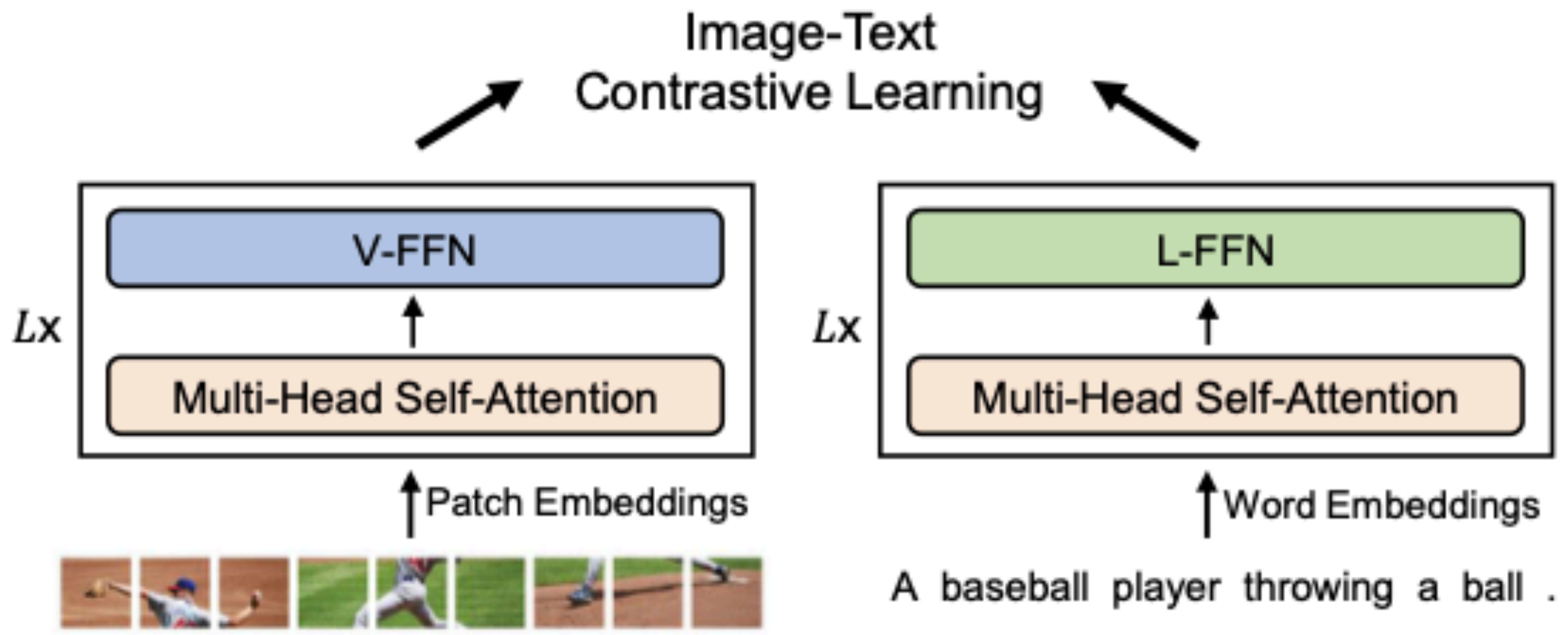
Mixture-of-Modality Experts Transformer

- MoME Transformer introduces mixture of modality experts as a substitute of the feed forward network of standard Transformer.
- Given previous layer's output vectors $\boldsymbol{H}_{l-1}, l \in [1,L]$.
- Each MoME Transformer block captures modality-specific information by switching to different modality expert, and employs multi-head self-attention (MSA) shared across modalities to align visual and linguistic contents.
- LN is short for layer normalization.
  - $\boldsymbol{H}_l' = \mathrm{MSA}(\mathrm{LN}(\boldsymbol{H}_{l-1})) + \boldsymbol{H}_{l-1}$
  - $\boldsymbol{H}_l = \mathrm{MoME-FFN}(\mathrm{LN}(\boldsymbol{H}_l')) + \boldsymbol{H}_l'$



Methodology

Pre-Training Tasks - Image-Text Contrast



- Given a batch of N image-text pairs, image-text contrastive learning aims to predict the matched pairs from $N \times N$ possible image-text pairs. There are $N^2 - N$ negative image-text pairs within a training batch.
- The final output vectors of $[I\_CLS]$ & $[T\_CLS]$ are used as the aggregated representation of the image and text, respectively.
- Followed by a linear projection and normalization, obtain image vectors $\{\hat{\boldsymbol{h}}_i^v\}_{i=1}^N$ and text vectors $\{\hat{\boldsymbol{h}}_i^w\}_{i=1}^N$ in a training batch to compute image-to-text and text-to-image similarities:

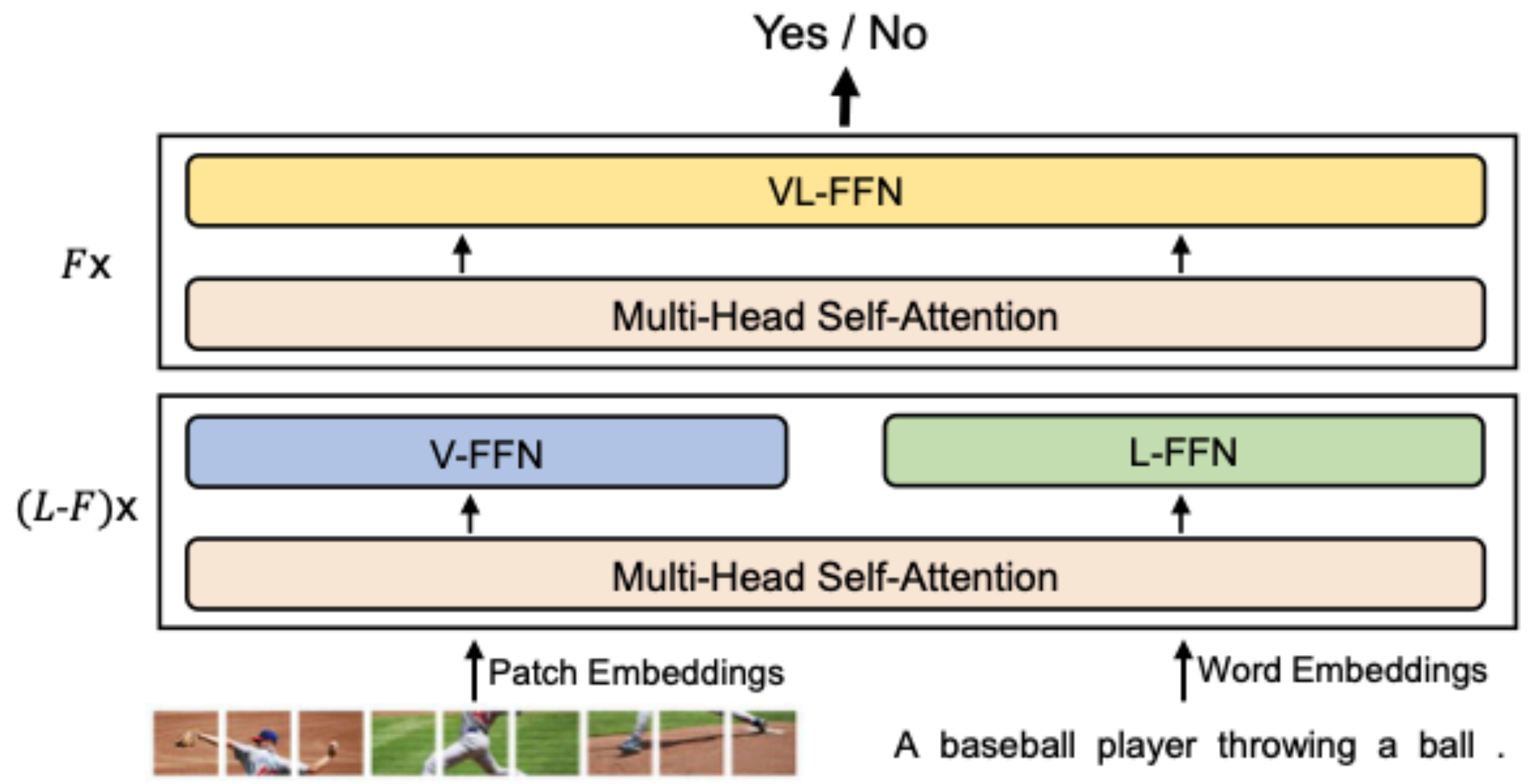
$$s_{i,j}^{i2t} = \hat{\boldsymbol{h}}_i^{v\top}\hat{\boldsymbol{h}}_j^w, s_{i,j}^{t2i} = \hat{\boldsymbol{h}}_i^{w\top}\hat{\boldsymbol{h}}_j^v$$

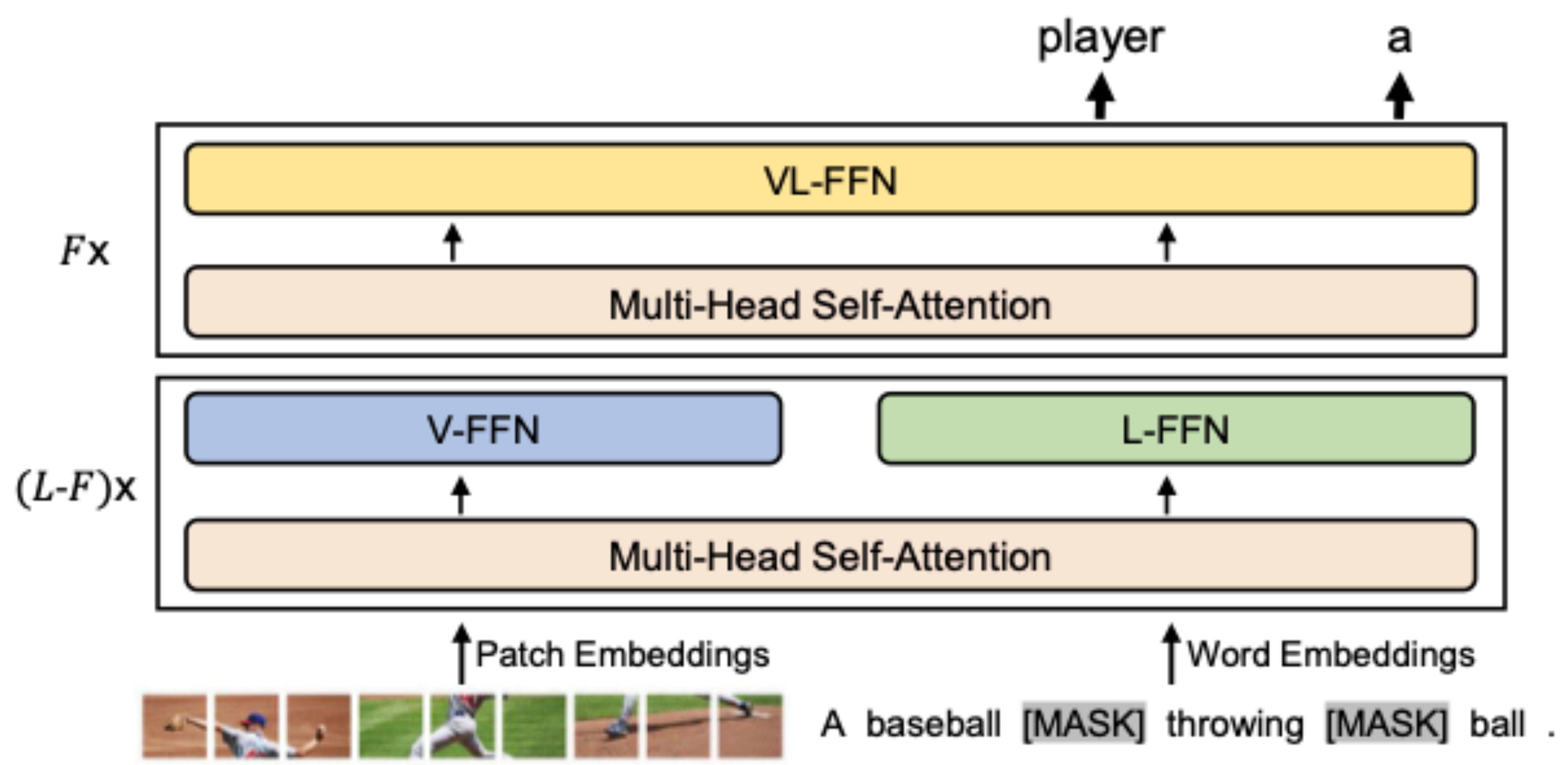
$$p_i^{i2t} = \frac{\exp(s_{i,i}^{i2t}/\sigma)}{\sum_{j=1}^N \exp(s_{i,j}^{i2t}/\sigma)}, p_i^{t2i} = \frac{\exp(s_{i,i}^{t2i}/\sigma)}{\sum_{j=1}^N \exp(s_{i,j}^{t2i}/\sigma)}$$

Methodology

Other Pre-Training Tasks

- Image-Text Matching
  - Image-text matching aims to predict whether the image and text is matched.
  - Using the final hidden vector of the $[T\_CLS]$ token to represent the image-text pair, and feed the vector into a classifier with cross-entropy loss for binary classification
- Masked Language Modeling
  - Following BERT, randomly choose tokens in the text sequence, and replace them with the $[MASK]$ token.
  - The model is trained to predict these masked tokens from all the other unmasked tokens and vision clues.





Methodology

Stagewise Pre-Training

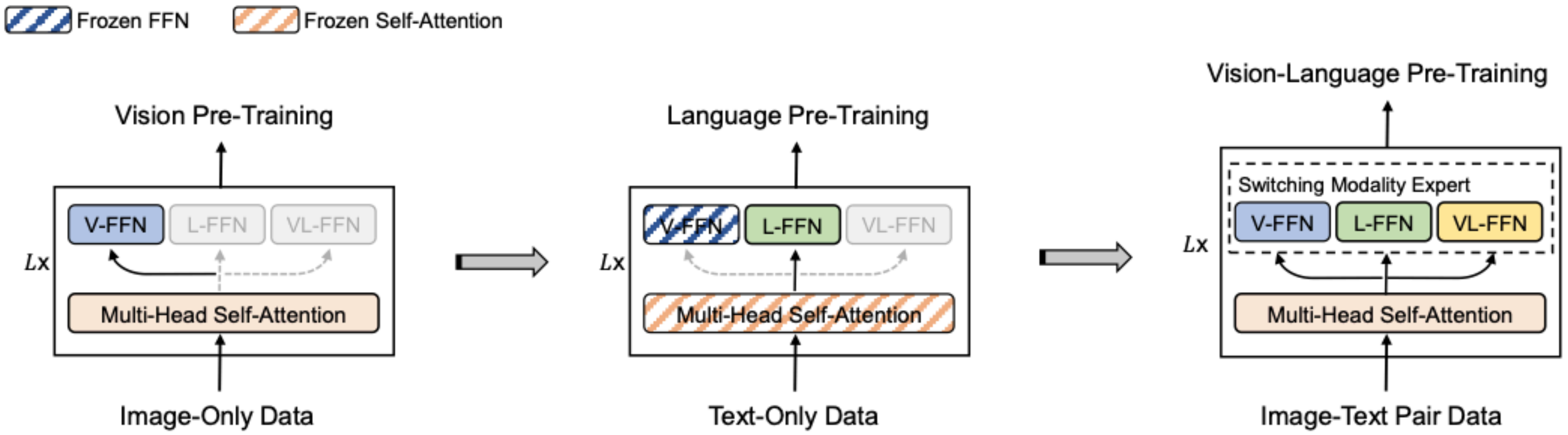
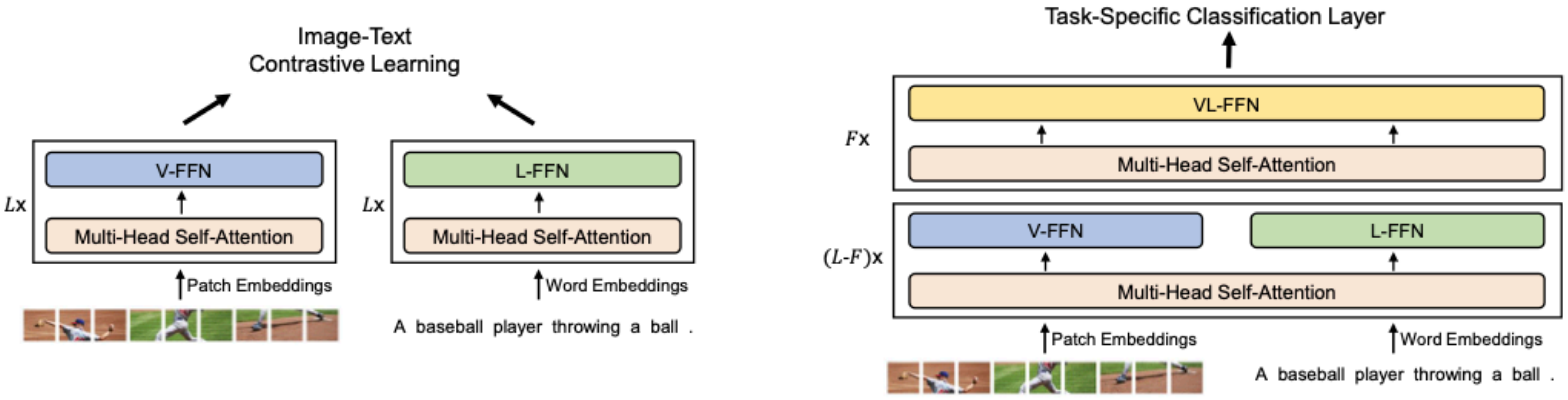


Figure 2: Stagewise pre-training using image-only and text-only corpora. We first pretrain the vision expert (V-FFN) and self-attention module on large-scale image-only data as in BEIT [2]. Then the parameters of vision expert and self-attention module are frozen, and we train the language expert (L-FFN) by masked language modeling on large amounts of text-only data. Finally, we train the whole model with vision-language pre-training.

Methodology

Fine-Tuning VLMo on Downstream Tasks



(a) Fine-Tuning on VL Retrieval Tasks

(b) Fine-Tuning on VL Classification Tasks

Figure 3: Fine-tuning VLMo on vision-language retrieval and classification tasks. The model can be fine-tuned as a dual encoder to separately encode image and text for retrieval tasks. VLMo can also be used as a fusion encoder to handle interaction of image-text pairs for classification tasks.

Experiments

Evaluation on Classification Tasks

- Visual Question Answering (VQA)
  - For VQA, a natural image and a question are given, the task is to generate/choose the correct answer. Train and evaluate the model on VQA 2.0 dataset.
  - Using the final encoding vector of the $[T\_CLS]$ token as the representation of the image-question pair and feed it to a classifier layer to predict the answer.
- Natural Language for Visual Reasoning (NLVR2)
  - The NLVR2 dataset requires the model to predict whether a text description is true about a pair of images.
  - Concatenate the final output vectors of the $[T\_CLS]$ token of the two input pairs. The concatenated vector is then fed into a classification layer to predict the label.

Experiments
Result

Model	# Pretrain Images	VQA test-dev	VQA test-std	NLVR2 dev	NLVR2 test-P
Base-Size Models Pretrained on COCO, VG, SBU and CC datasets					
UNITER-Base [3]	4M	72.70	72.91	77.18	77.85
VILLA-Base [14]	4M	73.59	73.67	78.39	79.30
UNIMO-Base [25]	4M	73.79	74.02	-	-
ViLT-Base [20]	4M	71.26	-	75.70	76.13
ALBEF-Base [23]	4M	74.54	74.70	80.24	80.50
VLMo-Base	4M	**76.64**	**76.89**	**82.77**	**83.34**
Large-Size Models Pretrained on COCO, VG, SBU and CC datasets					
UNITER-Large [3]	4M	73.82	74.02	79.12	79.98
VILLA-Large [14]	4M	74.69	74.87	79.76	81.47
UNIMO-Large [25]	4M	75.06	75.27	-	-
VLMo-Large	4M	**79.94**	**79.98**	**85.64**	**86.86**
Models Pretrained on More Data					
VinVL-Large [49]	5.7M	76.52	76.60	82.67	83.98
SimVLM-Large [46]	1.8B	79.32	79.56	84.13	84.84
SimVLM-Huge [46]	1.8B	80.03	80.34	84.53	85.15
Florence-Huge [48]	900M	80.16	80.36	-	-
VLMo-Large++	1.0B	**82.88**	**82.78**	**88.62**	**89.54**

Conclusion
of VLMo

- Propose a unified vision-language pretrained model VLMo.
  - which jointly learns a dual encoder and a fusion encoder with a shared MoME Transformer backbone.
- MoME introduces a pool of modality experts to encode modality-specific information, and aligns different modalities using the shared self-attention module.
- The unified pre-training with MOME enables the model to be
  - used as a dual encoder for efficient vision-language retrieval,
  - or as a fusion encoder to model cross-modal interactions for classification tasks.
- Showing that stagewise pre-training that leverages large-scale image-only and text-only corpus greatly improves vision-language pre-training.

Comments

of VLMo

- Not release checkpoint until now.
- Stagewise pretraining solve the lacking image-text pair problem.
  - Also let the unimodal encoder learn more generalizable representations.
- May can utilize concept of mixture of expert (MoE) to my work.