VLMo

- is a pretrained Vision-Language Model
- first work to introduce the Mixture-of-Modality-Experts (MoME) architecture for Transformer models
- here modality encoders, so image and text encoders, are not seperate
- image and text encoder are build using a Fusion Encoder
  - encoder consists of, as usual, Transformer blocks, but with two MLPs per block instead of one
  - one MLP for image and one for text
  - inspired by the Mixture-of-Experts (MoE) architecture
  - first introduced for NLP with LSTM models in 2017 [1] and just recently adapted to Transformer models in 2022
- for NLP idea was to have multiple (MLP) experts in one Transformer block, and route each token to one or more experts, through a learnable routing, and computing the weighted sum of the experts' outputs [2]
- in MoME, there are as many experts as there are modalities, for us two, as we use image and text
- routing is not learned during training, but based on the input modality
- if text is the input, then the text MLP is used, and if image is the input, then the image MLP is used
- Self-Attention is shared between image and text
- in upper layers, there is just one MLP, the Vision-Language Expert [3]
  - for VLMo-Base, 12 layers and 768 hidden dim, oriented on ViT-Base architecture, upper two layers are the Vision-Language Expert
  - for VLMo-Large, 24 layers and 1024 hidden dim, oriented on ViT-Large architecture, upper three layers are the Vision-Language Expert
- as in FLAVA [4], in layers where there are image and text experts, image and text are encoded seperately
  - but again, Self-Attention is shared between image and text
  - means Self-Attention has to be able to compute attention between text tokens, and attention between image patches seperately
- for the Vision-Language Expert layers, embeddings of text and image are concatenated and passed through the upper layers together
  - means Self-Attention in upper two layers, VLMo-Base, can compute attention between text tokens and image patches

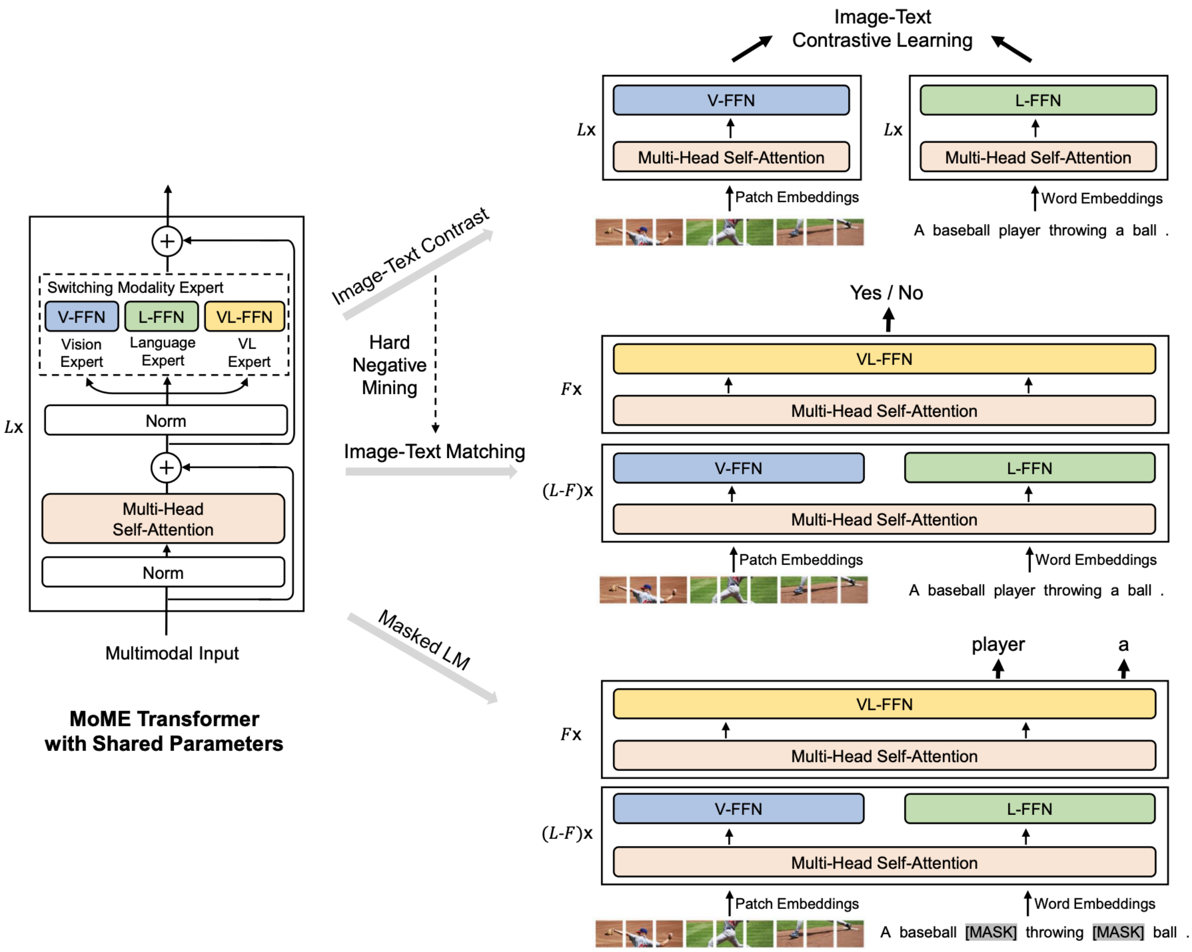

Figure 1: [3].

- trained using a combination of three losses:
  - masked language modeling (MLM)
    - 15% of text tokens are masked and the model has to predict them based on the image patches and unmasked text tokens
  - contrastive learning over all image-text pairs in the current batch, collected from all GPUs
    - total batch size is 1024, so they have 1023 negative samples
  - image-text matching
    - binary classification -> does an image and text belong together?
    - no cosine similarity between image embedding and text embedding performed, as in contrastive learning
    - because text and image embeddings are concatenated for VL expert layers, and passed as one input to the model, the CLS token of this joined embedding is taken, passed into a linear classifier, and use binary cross-entropy as the loss function
    - as in contrastive learning, negative examples (non-matching image-text pairs) also have to be used
    - negative samples created by hard negative mining
    - we take cosine similarities of contrastive learning from current batch, for each image, one text/caption is sampled based on a multinomial distribution over the cosine similarities between the image and all text/captions in the batch
    - text/caption with higher cosine similarity to the image has higher probability of being sampled -> makes the task more challenging

- for retrieval downstream task, whether it is zero-shot or finetuning, cls token output of image and text encoders are used
- not the cls token output of the Vision-Language Expert, as seen in Figure 2 (a)
- for finetuning on vision-language tasks, cls token output of the Vision-Language Expert is passed to a classification head, which is just a linear layer

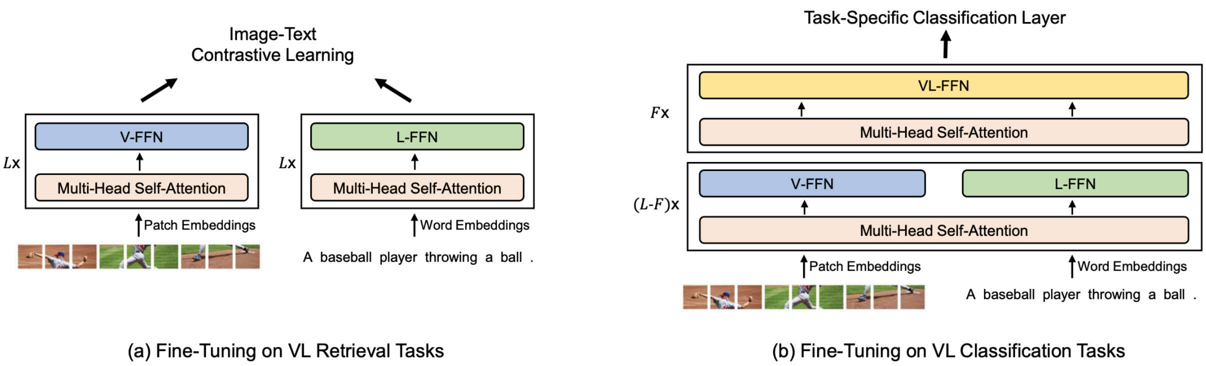

Figure 2: [3].

Bibliography

[1] N. Shazeer *et al.*, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, 2017.

[2] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: scaling to trillion parameter models with simple and efficient sparsity," *J. Mach. Learn. Res.*, vol. 23, no. 1, Jan. 2022.

[3] H. Bao *et al.*, "VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts," in *Advances in Neural Information Processing Systems*, 2022. [Online]. Available: https://openreview.net/forum?id=bydKs84JEyw

[4] A. Singh *et al.*, "FLAVA: A foundational language and vision alignment model," *CoRR*, 2021, [Online]. Available: https://arxiv.org/abs/2112.04482