JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Efficient Multimodal Fusion via Interactive Prompting

Yaowei Li[1], Ruijie Quan[2], Linchao Zhu[2] , Yi Yang[2]

[1] ReLER, AAII, University of Technology Sydney [2] CCAI, Zhejiang University

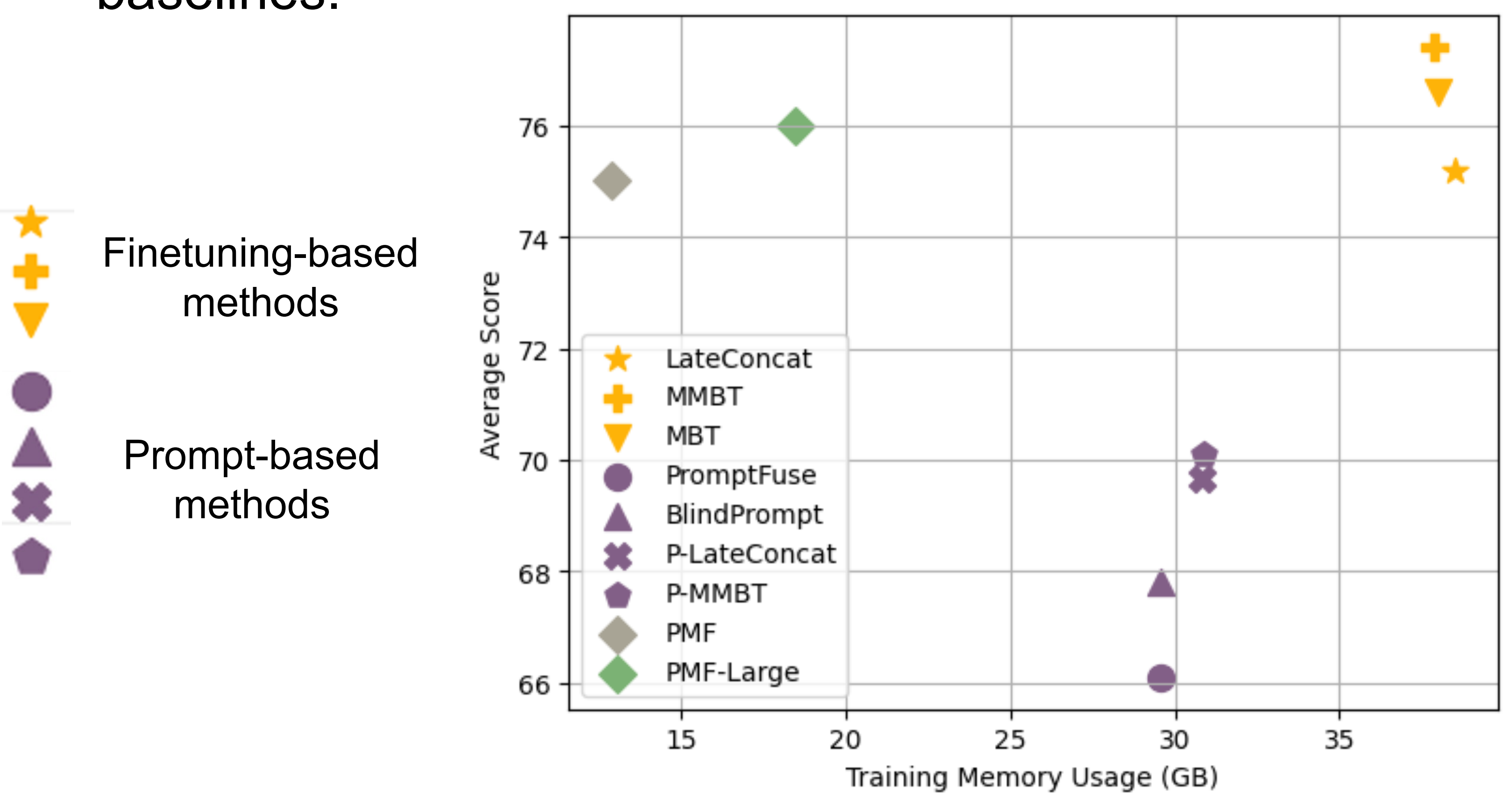
ReLER Recognition, LEarning, Reasoning

Motivation

- Existing multimodal learning methods are mostly finetuning-based, which requires huge computational costs since the gradients and optimizer states for all parameters of multimodal models have to store.
- Prompt-based methods are parameter-efficient but not very memory-efficient.
- Prompt-based methods usually underperformed fine-tuning baselines by a large margin with full data.

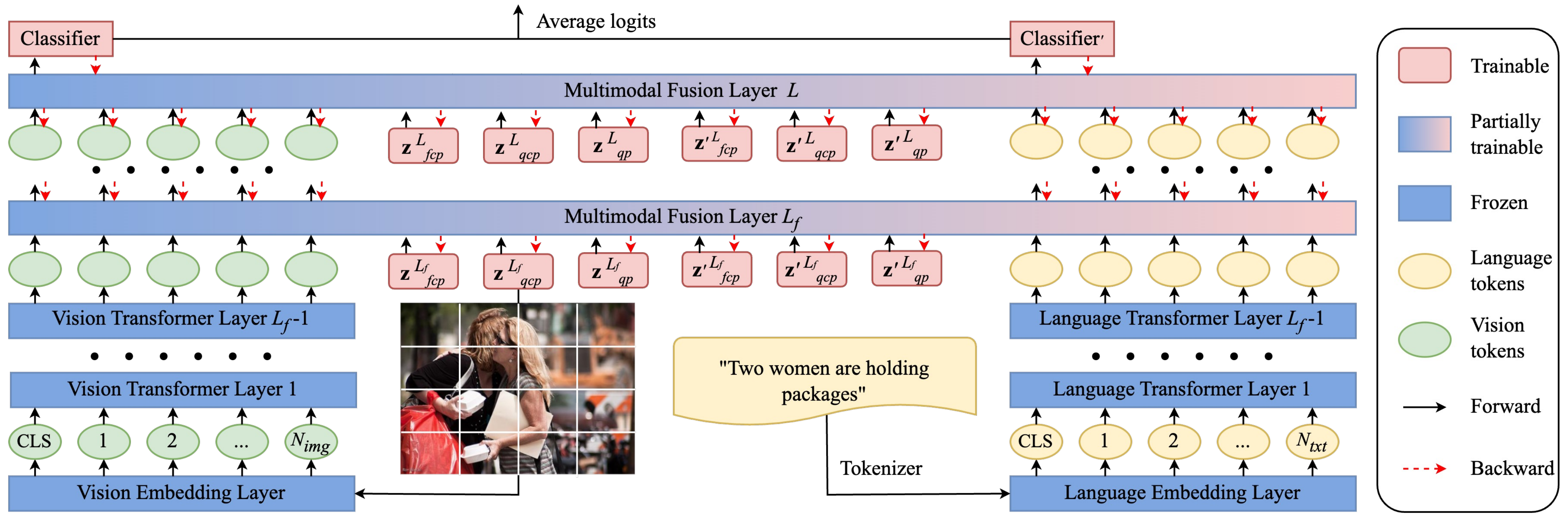
Contribution

- We proposed a Prompt-based Multimodal Fusion method (**PMF**) for fusing unimodally pretrained transformers.
- **PMF** is both parameter-efficient and memory-efficient. Compared with finetuning-based methods, PMF has less than 3% trainable parameters and saves more than 60% of training memory usage.
- **PMF** is modular and flexible. It can be applied to any unimodally pretrained transformers.
- **PMF** achieves comparable performance with finetuning baselines.



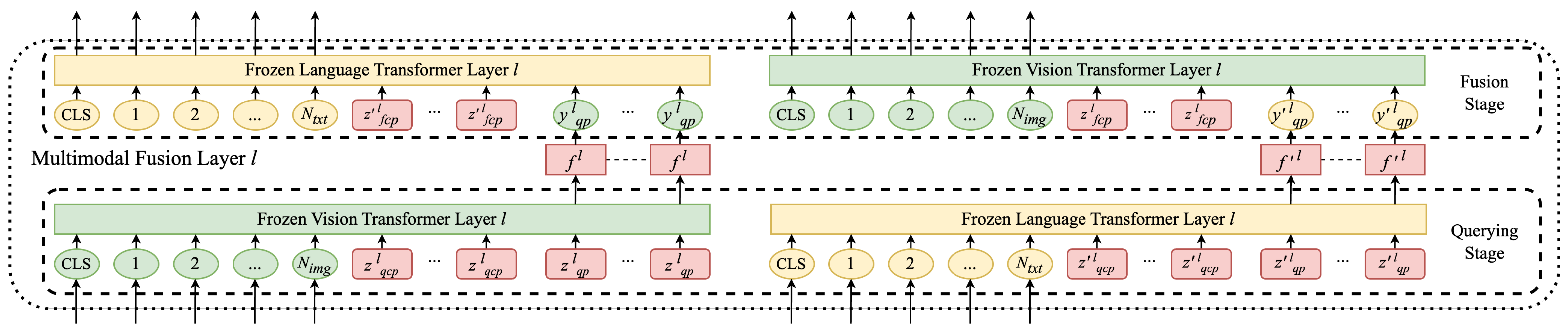
Proposed Algorithm

Prompt-based Multimodal Fusion Strategy



PMF applied to vision-language inputs. In the forward propagation, image and text inputs are first embedded into continuous token sequences and fed to the unimodal transformer layers for base feature extraction. The base features from both modalities then pass through multiple prompt-based multimodal fusion layers to get the feature of two CLS tokens for final classification. In the backward propagation, only multimodal fusion layers take part in the calculation of gradients, greatly saving memory usage during training.

Interactive Prompting in Multimodal Fusion Layers


Frozen Language Transformer Layer l
Frozen Vision Transformer Layer l
Fusion Stage
CLS
1
2
...
N_{txt}
z'^l_{fcp}
y^l_{qp}
N_{img}
z^l_{fcp}
y'^l_{qp}
Multimodal Fusion Layer l
f'^l
f^l
Frozen Vision Transformer Layer l
Frozen Language Transformer Layer l
Querying Stage
z^l_{qcp}
z^l_{qp}
z'^l_{qcp}
z'^l_{qp}

Prompt-based multimodal fusion layer. Intuitively, the query context prompt (z_{qcp}, z'_{qcp}) and query prompt (z_{qp}, z'_{qp}) can be seen as a pair of 'question' and 'answer' with the aim of extracting necessary information for exchange between two modalities. After being translated by a non-linear mapping 'translator' (f^l, f'^l), the 'answer' is then delivered to the other modality for better cross-modal understanding. Finally, the fusion context prompts '(z_{fcp}, z'_{fcp}) then provide the context to the delivered answer to facilitate the fusion.

Experiments

Main Results

Method	Updated Param. (Million)	Memory Usage (GB) Train/Inference	SNLI-VE	Food-101	MM-IMDB	Avg.
Linear	-	3.76 / 3.23	50.05	78.96	49.76 / 56.83	60.77
ViT	86.5	9.36 / 1.99	33.33	74.69	38.39 / 49.88	50.72
BERT	109.0	30.82 / 2.79	69.82	87.44	58.91 / 64.31	72.96
LateConcat	196.0	38.54 / 3.36	70.01	93.29	59.56 / 64.92	75.18
MMBT*	196.5	37.87 / 3.48	74.69	94.10	60.80 / 66.10	77.41
MBT*	196.0	38.00 / 4.06	74.02	93.56	59.60 / 64.81	76.60
VPT	-	6.12 / 2.01	33.33	72.55	35.22 / 44.49	48.58
P-BERT	-	28.13 / 2.99	63.28	81.07	48.67 / 54.58	65.33
PromptFuse	-	29.57 / 3.55	64.53	82.21	48.59 / 54.49	66.09
BlindPrompt	-	29.57 / 3.65	65.54	84.56	50.18 / 56.46	67.81
P-LateConcat	0.3	30.82 / 3.43	63.05	89.03	53.91 / 59.93	69.67
P-MMBT	0.9	30.90 / 3.48	67.58	86.58	52.95 / 59.30	70.10
PMF (M=4, L_f=L-2)	2.5	12.84 / 4.08	71.92	91.51	58.77 / 64.51	75.02
PMF-large (M=4, L_f=L-2)	4.5	18.44 / 6.42	72.10	91.68	61.66 / 66.72	75.99

Multimodal classification performance. PMF achieve comparable performance to the finetuning baselines with less than 3% of trainable parameters and up to 66% of training memory usage. PMF-Large uses bert-large less and vit-large models (24 hidden layers) while others use bert-base and vit-base models (12 hidden layers).

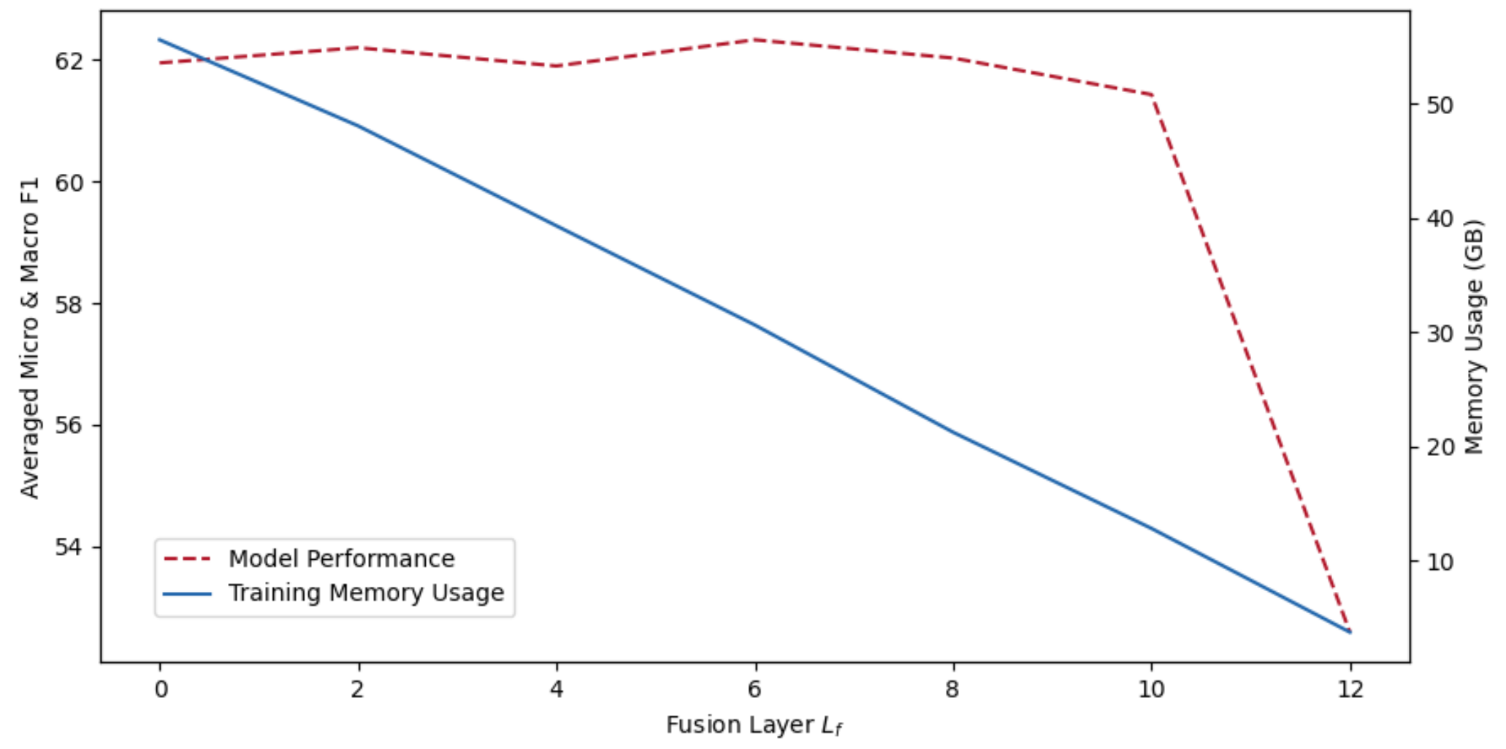
Text Encoder	Image Encoder	Memory Usage Train/Inference	MM-IMDB
bert-base	vit-base	12.84 / 4.08	58.77 / 64.51
bert-base	vit-large	14.16 / 4.89	59.70 / 65.20
bert-large	vit-base	17.17 / 5.53	60.08 / 65.41
bert-large	vit-large	18.44 / 6.42	61.66 / 66.72

PMF applying to different unimodal transformers. PMF can be applied to unimodal transformers at different scales with more memory savings.

Training Memory	SNLI-VE	Food-101	MM-IMDB	Avg.
33.36 GB	72.27	92.1	59.67 / 65.57	75.66

PMF applied with Network Aarchitecture Search (NAS). PMF can be further boosted with NAS algorithms to save time on searching introduced hyperparameters.

Ablation Study



Model Performance and Training Memory Usage under different fusion layers L_f .

- The training memory usage keeps decreasing as the fusion starts later.
- Performance is relatively consistent with fusion layer $L_f \leq 10$.