CLIP-FMoE: Scalable CLIP via Fused Mixture-of-Experts with Enforced Specialization

Luong Tran[1], Lan-Cuong Nguyen[1], Huynh Dang Nguyen[1], Dat Nguyen Cong[1], Dung D. Le[2], Van Nguyen[1*]

[1]FPT Software AI Center
[2]Center for AI Research, VinUniversity

ICLR

Overview

CLIP-FMoE scales CLIP with specialized experts while preserving zero-shot generalization through a two-stage fusion design.

- **Isolated Constrained Contrastive Learning (ICCL):** two-level semantic clustering and parallel expert training.
- **Fusion Gate:** blends pretrained CLIP and domain experts token-wise.
- **Broad gains:** stronger classification vs CLIP adaptation baselines, with best retrieval and long-context retrieval.

Why CLIP-FMoE?

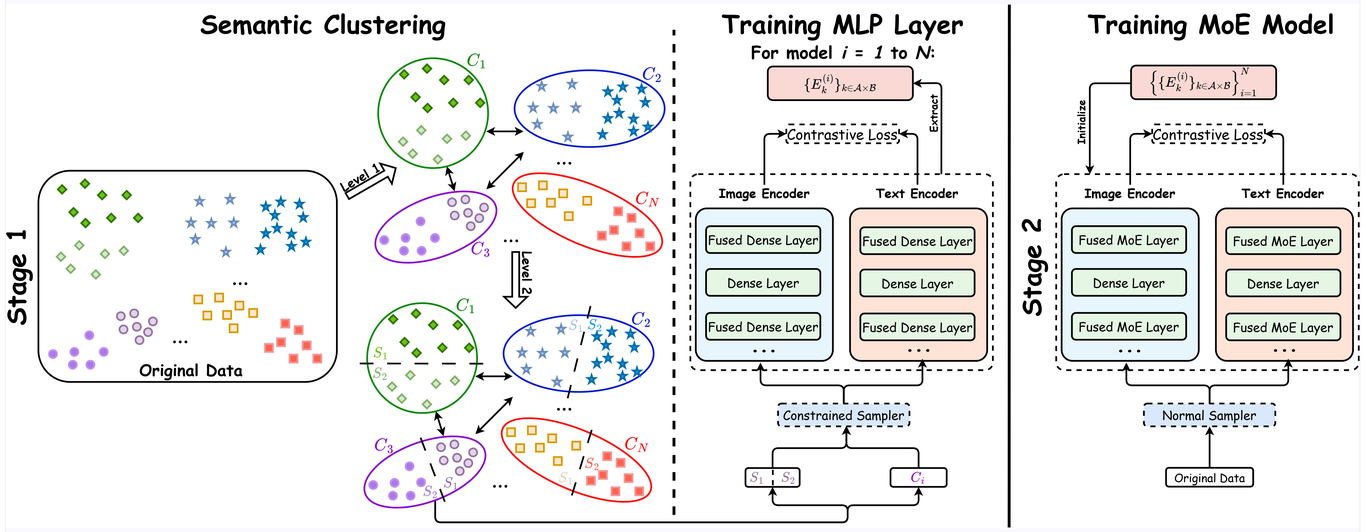
- Pure CLIP scaling is expensive: larger models require longer schedules and larger GPU clusters.
- Existing CLIP-MoE variants often suffer from weak expert specialization or costly sequential training.
- Fine-tuning on fine-grained captions can hurt original zero-shot performance (catastrophic forgetting).

Goal: increase capacity and specialization while keeping pretrained CLIP knowledge.

Takeaways

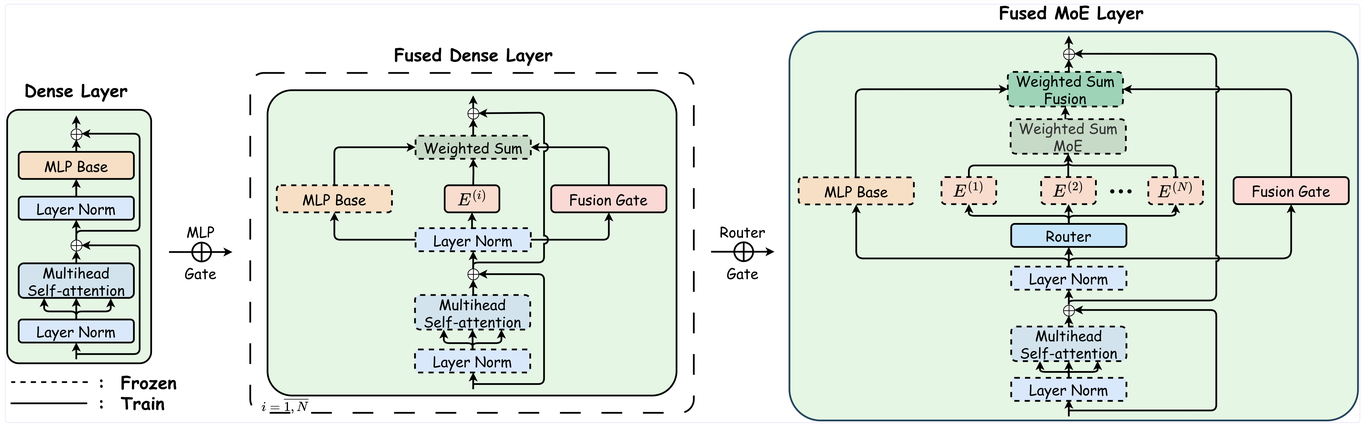
- CLIP-FMoE offers a strong efficiency/performance trade-off for CLIP adaptation.
- Future work: broader multimodal extension (video-text, audio-text, multilingual VLMs).

Two-Stage ICCL Pipeline



- **Stage 1 (ICCL):** each expert learns from one semantic cluster with constrained sub-cluster sampling.
- **Stage 2 (Unification):** freeze experts, train router and Fusion Gate on full data.

Fused MoE Block



Architecture. Expert outputs are fused with the base MLP through a learnable gate.

$$E_{agg}(x) = G(x)E_{base}(x) + (1 - G(x))E_{new}(x), \quad G(x) = \sigma(Linear(x))$$

This gate keeps pretrained representations accessible while allowing expert-specific adaptation.

Efficiency and Scalability

72.5%	4.15x	~80%
less total training time vs CLIP-MoE	faster clustering vs CLIP-MoE	fewer peak trainable params vs Up-Cycling

Model	Total Train (min)	Infer FLOPs
CLIP-MoE	193.51	112.44
CLIP-FMoE	**53.24**	128.22

Zero-Shot Classification

Method	Avg (11)	Cars	FGVC	IN-1K
OpenAI CLIP	69.62	77.93	31.77	75.54
Fine-tuning	66.08	69.12	25.83	72.89
Up-cycling	66.69	68.98	27.66	72.92
CLIP-MoE	66.47	71.20	27.81	73.19
CLIP-FMoE	**68.65**	**74.62**	**29.58**	**74.87**

Note: Cars/FGVC/IN-1K are representative datasets; Avg (11) is averaged over all 11 benchmarks reported in the paper.

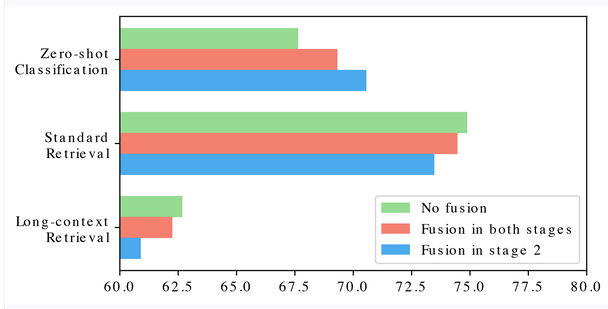
CLIP-FMoE reduces the average drop from pretrained CLIP to < 1%, while clearly outperforming other adaptation baselines.

Long Context Understanding

Method	MS COCO IR	MS COCO TR	IIW-Long IR	IIW-Long TR
OpenAI CLIP	36.5	56.4	90.5	90.2
Fine-tuning	46.6	64.3	**97.5**	96.8
LongCLIP	47.0	63.3	97.2	94.5
TULIP	45.7	61.2	97.2	96.2
CLIP-FMoE	**47.7**	**65.6**	97.3	**98.0**

Setting: CLIP ViT-L/14 with 248-token context (except OpenAI CLIP), trained on ShareGPT4V.

Ablation



Fusion Gate ablation. We compare 3 variants: no gate, gate in both stages (chosen design), and gate only in Stage 2 across classification and retrieval settings.