Hyper-Align: Efficient Modality Alignment via Hypernetworks

Jaisidh Singh[1,2,3,5] Diganta Misra[2,3] Boris Knyazev[6] Antonio Orvieto[2,3,4]

[1]University of Tübingen [2]ELLIS Institute Tübingen [3]MPI for Intelligent Systems, Tübingen
[4]Tübingen AI Center [5]Zuse School ELIZA [6]SAIT AI Lab Montreal

ICLR 2025 Workshop on Weight Space Learning

SUMMARY

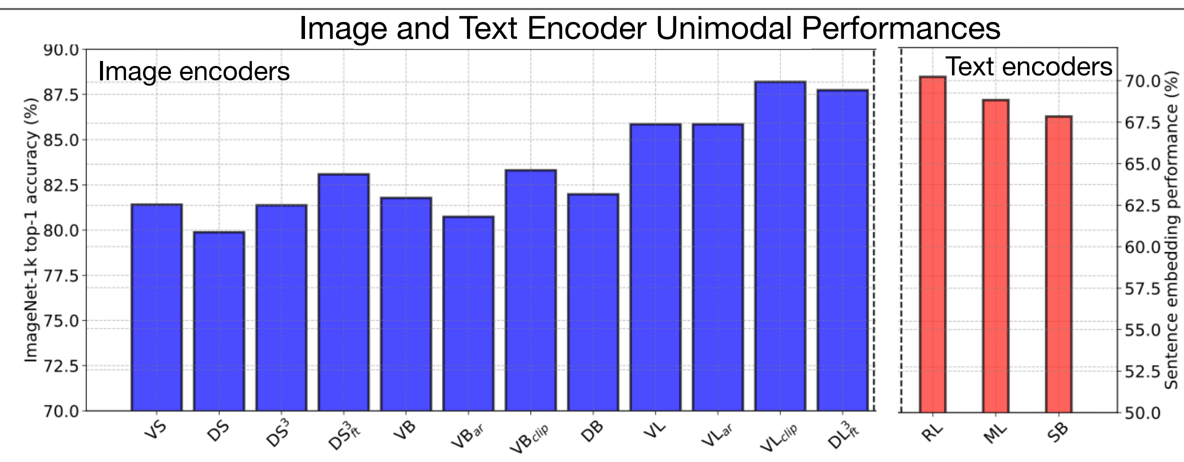
Contrastive vision-language models (VLMs) like CLIP: Align encoders of image-text modalities via an InfoNCE loss.

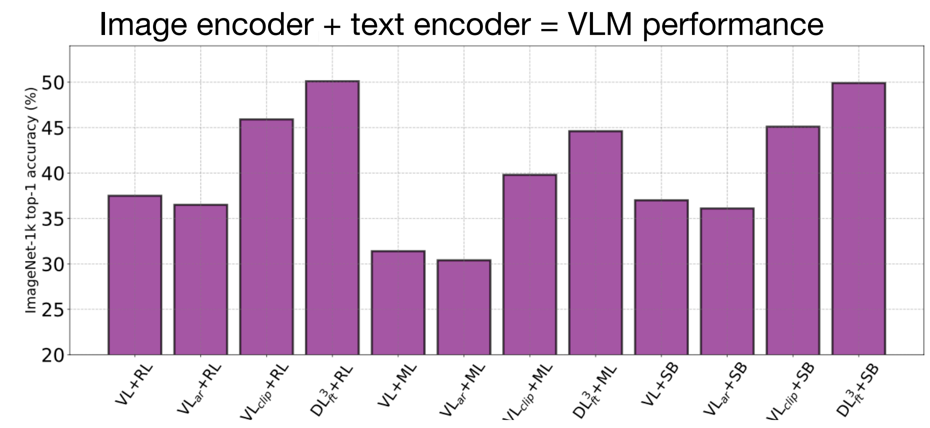
Background: Instead of training VLMs end-to-end,
- APE trains a modality connector (MLP) between pretrained encoders
- outperforms CLIP at significantly lower costs

Problem: unimodal performance ≠ multimodal performance.
- Finding the optimal pair in N image and M text encoders requires searching all $N \times M$ combinations
- Training any one combination needs massive data volumes
- Hence, training all combinations individually becomes unfeasible

Proposed solution (Hyper-Align): Use a hypernetwork to learn all $N \times M$ modality connectors together, instead of learning them individually via APE.

Result: With linear layers as the modality connectors, Hyper-Align is 8x cheaper than APE in terms of FLOP costs, at negligible performance drop.





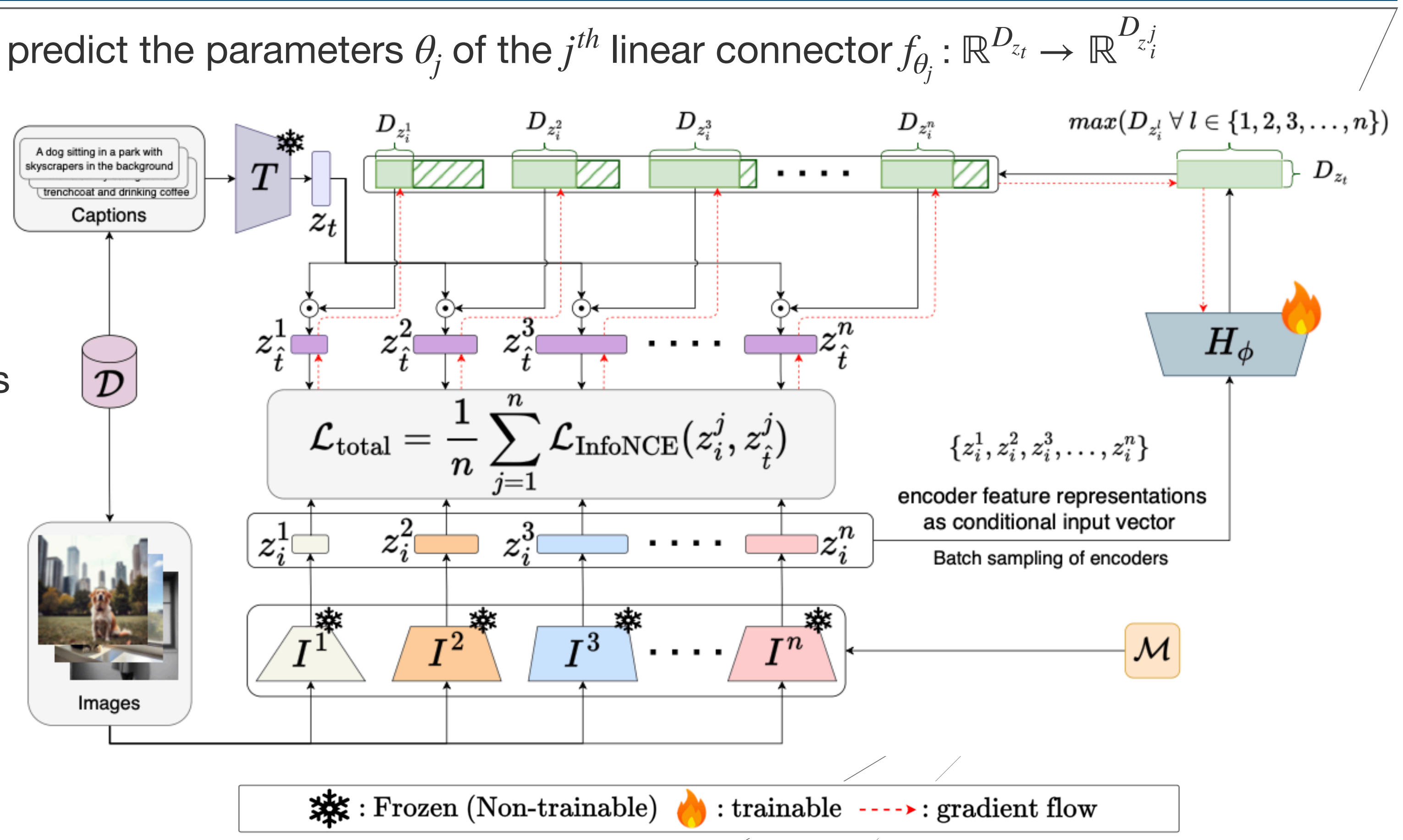
METHODOLOGY

Overview: Hypernetwork H_ϕ uses a conditional input c_j to predict the parameters θ_j of the j^{th} linear connector $f_{\theta_j} : \mathbb{R}^{D_{z_t}} \rightarrow \mathbb{R}^{D_{z_i^j}}$
- n image encoders and 1 text encoders
- $H_\phi(c_j) = \theta_j$ where $c_j =$ padded-batch-average(z_i^j)
- Training objective is L_{total}

Hypernetwork design: H_ϕ is an MLP that
- observes image features of different dimensions
- to predict connector parameters of variable dimensions

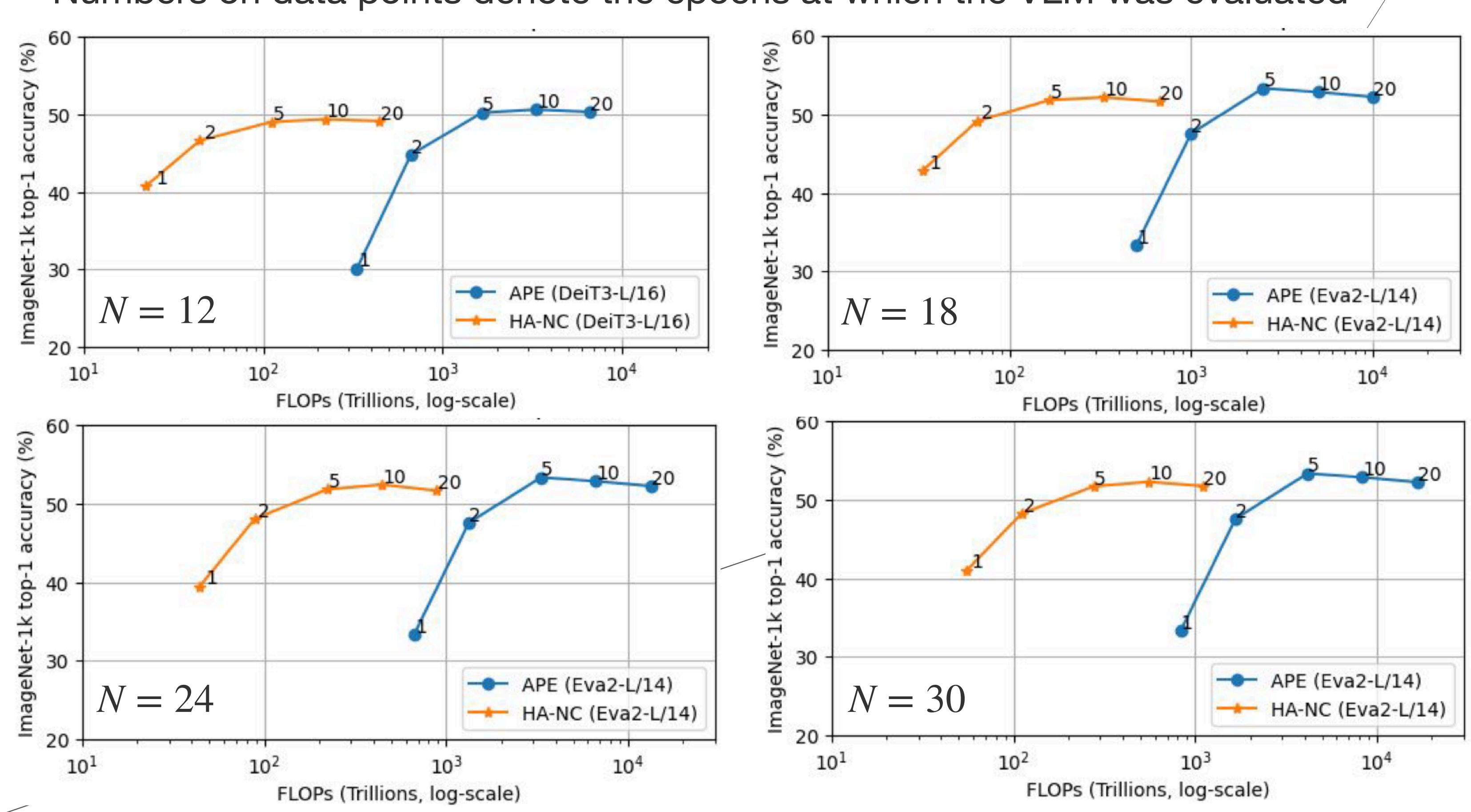
Efficient training via model-batching:
- train on a mini-batch $B_m < n$ of image encoders per step to efficiently scale up the no. of combinations (n)

$\mathcal{D}$: Multimodal Dataset
$\mathcal{M}$: Image Encoder Zoo
H_ϕ : HyperNetwork
I^k : k^{th} image encoder
T : text encoder
z_i^k : k^{th} image feature vector
z_t : text feature vector
$z_{\hat{t}}^k$: k^{th} mapped text feature vector
$D_{z_i^k}$: dimensionality of k^{th} image feature vector



EXPERIMENTS & RESULTS

Scaling up no. of combinations:
- $M = 1$ (sentence-t5-base) and N varies from 12 to 30
- Best image encoder reported at each value of N
- Numbers on data points denote the epochs at which the VLM was evaluated



Search over various image encoder scales:
- $N = 30$ equally split among 3 feature dims
- Parameter count ↑ as feature dim ↑

Best ImageNet accuracy shown per scale

Scale type	Range	Method: Ours	Method: APE
Feature dim	384	36.75	38.36
	768	42.83	45.44
	1024	51.92	53.86
Param. count	< 30M	36.75	38.36
	30M − 120M	43.04	44.84
	> 120M	51.92	53.86

CONCLUSION

Parameter prediction via hypernetworks can
- **efficiently search image-text encoder pairs** for optimal VLMs, under constraints
- Future work can use Hyper-Align on **image encoders and LLMs** to create MLLMs

Corresponding author: jaisidh.singh@student.uni-tuebingen.de