Adapting Large-Scale Vision-and-Language Models for Downstream Classification

Konwoo Kim, advised by Professor Deepak Pathak

Carnegie Mellon University

Introduction

Vision-and-language pre-training methods have rapidly shown increasing promise. Although fine-tuning methods have emerged, many have only been studied for language models and it's unclear how they perform on diverse downstream settings.

Contrastive Language-Image Pre-Training

In our analysis, we focus on CLIP [1] (contrastive language-image pre-training), a large-scale model trained on **over 400 million** online images and text. Given a batch of image-text data, CLIP is trained with a contrastive loss to maximize the similarity of paired data and minimize the similarity of unpaired data.

$$p(\boldsymbol{T}_j \mid \boldsymbol{I}_i) = \frac{\exp\left(\cos(\boldsymbol{T}_j, \boldsymbol{I}_i)/\tau\right)}{\exp\left(\cos(\boldsymbol{T}_j, \boldsymbol{I}_i)/\tau\right) + \sum_{k \neq j} \exp\left(\cos(\boldsymbol{T}_k, \boldsymbol{I}_i)/\tau\right)}$$

Figure 1. Prediction probability for a single image-text pair

For our experiments, we consider three classes of fine-tuning methods which can act on CLIP.

- Methods which only fine-tune **existing parameters**
  - Full-model fine-tuning, LayerNorm tuning
- Methods which **add new parameters** at the beginning, middle, or end of the model
  - Prompt-tuning [2], Adapter/Compacter [3], Linear probe
- Methods which combine both ideas

We analyze these on four settings determined by two factors: **amount and distribution of downstream data.**

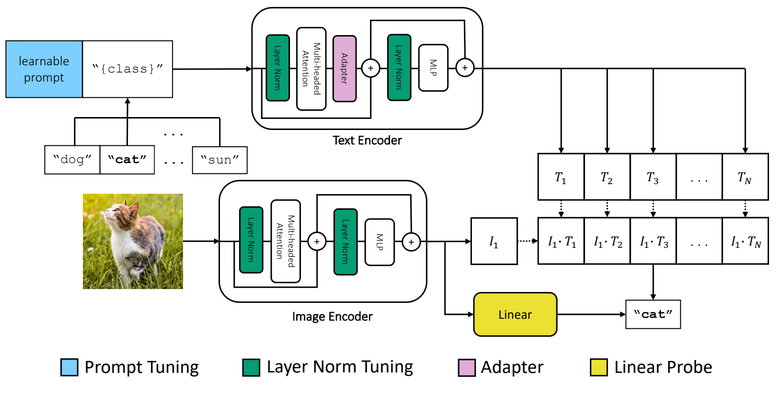

Figure 2. Illustration of multiple methods for adapting CLIP to downstream image classification tasks

Effectiveness of LayerNorm

We find that fine-tuning only the layer normalization parameters is an effective baseline. These parameters exist within intermediate layers of CLIP and apply per-element normalization across batches.

$$\boldsymbol{y} = \frac{\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}]}{\sqrt{\mathrm{Var}[\boldsymbol{x}] + \epsilon}} \cdot \gamma + \beta$$

Figure 3. Layer normalization transformation on a single mini-batch

LayerNorm tuning performs the best in all four regimes, across amount and distribution of downstream data.

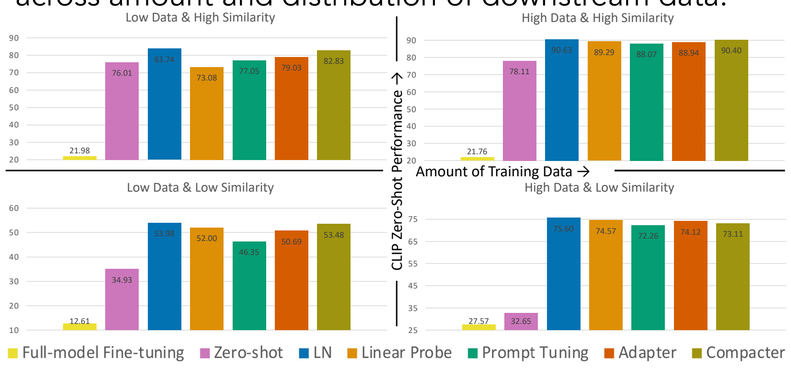

Figure 4. Comparison of fine-tuning methods across different regimes of training data and CLIP zero-shot performance

Because LayerNorm tuning only affects existing parameters, we propose two ways of combining it with existing fine-tuning methods.

- Simultaneously fine-tuning the parameters of both methods
- Using the weights of a pre-trained LayerNorm-tuned model as initialization

Across the low and high-data regimes, both combinations consistently improve upon the baseline performance of other fine-tuning methods.

	Low-Data Regime			High-Data Regime		
Type of LN Tuning	None	Normal	As Initialization	None	Normal	As Initialization
Linear Probe	62.54	63.55	**66.61**	81.93	83.84	**84.36**
Prompt Tuning	61.70	62.82	**63.95**	80.26	**83.69**	83.17
Adapter	64.82	66.23	**66.63**	81.53	83.31	**83.63**
Compacter	68.15	**69.69**	68.76	81.75	**83.61**	83.17

Table 1. Effect of combining LayerNorm tuning with other methods

The strong performance of LayerNorm tuning shows the benefit of **learning representations grounded in multiple modalities**.

Domain and Class Generalization

We evaluate our methods within a few-shot setting and on a domain generalization benchmark. We find that:

- LayerNorm tuning alone is competitive with current, task-specific state of the art methods
- With a linear probe, it produces representations which generalize across classes and domains

	FMoW	Camelyon17	iWildCam
Zero-shot	19.71	67.46	3.73
LayerNorm Tuning	47.59	90.47	18.52
Linear Probe + LN as Initialization	48.98	89.98	23.80
Best leaderboard result	**55.5**	**91.6**	**38.5**

	Mini-ImageNet (1-shot)	Mini-ImageNet (5-shot)
Zero-shot	86.20	96.56
LayerNorm	89.24	96.46
Prompt Tuning + LN	89.61	97.05
Adapter + LN	91.17	97.39
Linear Probe + LN	**92.08**	**97.94**
Best leaderboard result	82.99	91.50

Table 2. Domain generalization and few-shot results

Discussion

Our analysis provides two key findings:

- LayerNorm tuning is a simple but effective baseline for adapting to downstream classification tasks.
- Combining LayerNorm tuning with existing methods improves performance in diverse settings.

Our results show the importance of **joint vision-and-language training for adaptation and robust visual representations**. Future directions of work include:

- Examining alternative pre-training schemes
- Evaluating fine-tuning methods on more general tasks across vision and language
- Applying fine-tuning methods to other domains like reinforcement learning or robotics

References

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020.

[2] Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691.

[3] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning (pp. 2790-2799). PMLR.