Low-Rank Few-Shot Adaptation of Vision-Language Models

著者	Maxime Zanella
発行年	2024
学会名	CVPR
キーワード	LoRA Vision-Language Model
URL	https://openaccess.thecvf.com/content/CVPR2024W/PV/html/Zanella_Low-Rank_Few-Shot_Adaptation_of_Vision-Language_Models_CVPRW_2024_paper.html

概要

この研究は、Vision-Language Models (VLMs)におけるFew-Shot学習に対して、Low-Rank Adaptation (LoRA)を挿入することで、効率的かつ高精度な適応が可能であることを示したもの。

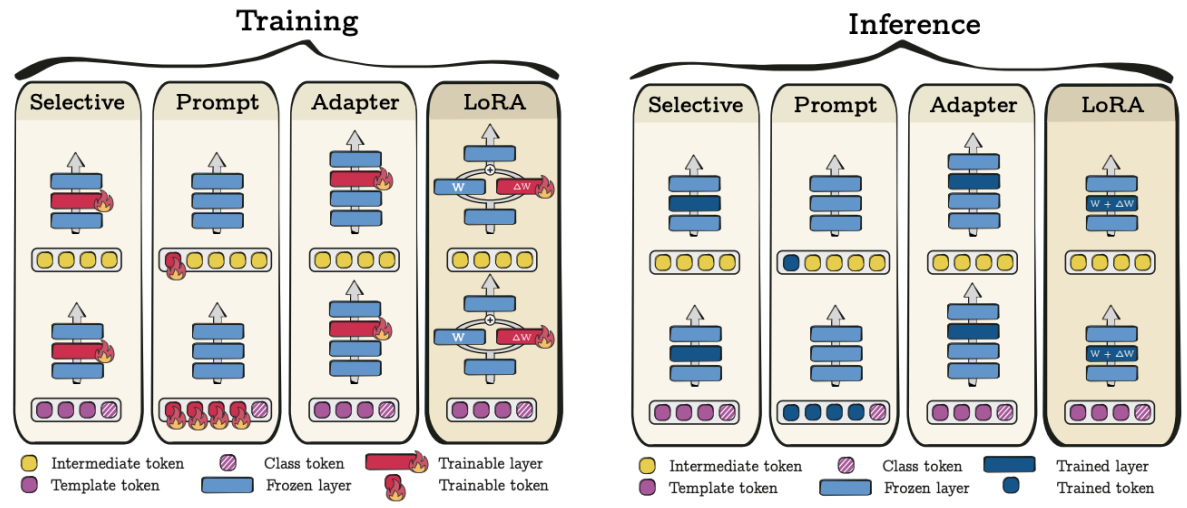
提案手法である"CLIP-LoRA"は、従来のプロンプト学習やアダプター手法よりも少ない計算資源で高い性能を達成している。

研究背景

- CLIPのようなVLMは画像と言語の共通の埋め込み空間を使ってゼロショットの分類が可能である
- Few-Shotでの適応には再学習が必要であり、パラメータ効率性が課題になっていた

    - 既存のFew-Shot手法は、プロンプト学習やアダプターに依存しており、計算コストやハイパラの調整が必要になってくる
- NLPの文脈では、LoRAのようなPEFT手法が普及している一方で、VLMの文脈では行われていなかった

⬇️ 各チューニング手法の説明



(a) Prompt, Adapter and Low-rank techniques introduce extra parameters for training, which may potentially extend training duration and/or memory footprint in comparison to selective methods. However, they have the advantage of being more flexible and are often easier to use.

(b) Prefix and Adapter methods result in extra parameters after adaptation, potentially increasing inference time and memory footprint relative to the vanilla model. Conversely, LoRA merges newly trained low-rank matrices with the original frozen ones, eliminating additional parameters at inference.

Figure 1. Different categories of Parameter-Efficient Fine-Tuning (PEFT) methods during (a) training, and (b) inference.

論文の肝

- CLIP+VLMの組み合わせで、Few-shotでも効率よくチューニングすることができた
- LoRAを展開する上で、適応させるエンコーダの選択や重み行列の選択、行列のランクなど徹底検証した
- LoRAの導入により、タスクに依存しないハイパラで、SoTAを上回ることができた

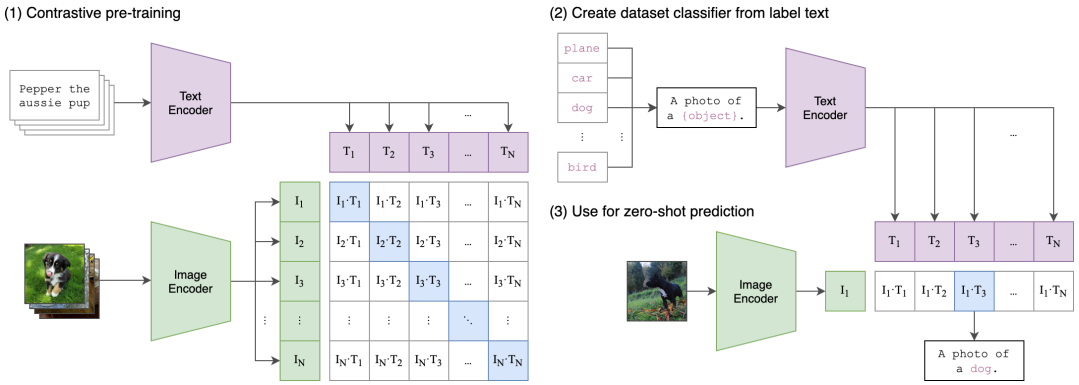
提案手法

📎 そもそもCLIPとは

OpenAIが提案した手法である。link

(1) 訓練時：画像x_iをVision Encoder θ_{vision}に、対応するプロンプトc_k (a photo of a dog)をText Encoder θ_{text}に埋め込み、対応させることで、画像とテキストの共通埋め込み空間を獲得する。

(2) 推論時：入力したテスト画像x_tの特徴量と、各クラスプロンプトの特徴量の類似度を測り、最も類似度が高いクラスプロンプトをその画像のクラスとして推論する。



CLIP-LoRA

CLIPの画像エンコーダとテキストエンコーダはTransformerで構成されている。その両方のTransformerの中のアテンションの重み ($W_{Query}, W_{Key}, W_{Value}$)を以下のように変更する。

$$h = Wx + \gamma \Delta W x = Wx + \gamma BAx$$

- xは入力、hは特徴
- $\Delta W \in R^{d_1 \times d_2}$の時、$A \in R^{r \times d_2}, B \in R^{d_1 \times r}$、$r = 2$で今回行っている
- γはスケーリング係数

以上のように、元の重みWは更新せず，低ランクなBAのみをチューニングする。BAを学習すべき差分と捉えることができ、安定かつ効率的に学習させることができる。

また、プロンプト文やハイパラを固定し、再現性やタスクの非依存性を確保した。

実験

精度評価

- 多様なデータセット
- 多様な比較手法

とともに精度評価

Table 2. Detailed results for 11 datasets with the ViT-B/16 as visual backbone. Top-1 accuracy averaged over 3 random seeds is reported. Highest value is highlighted in **bold**, and the second highest is underlined.

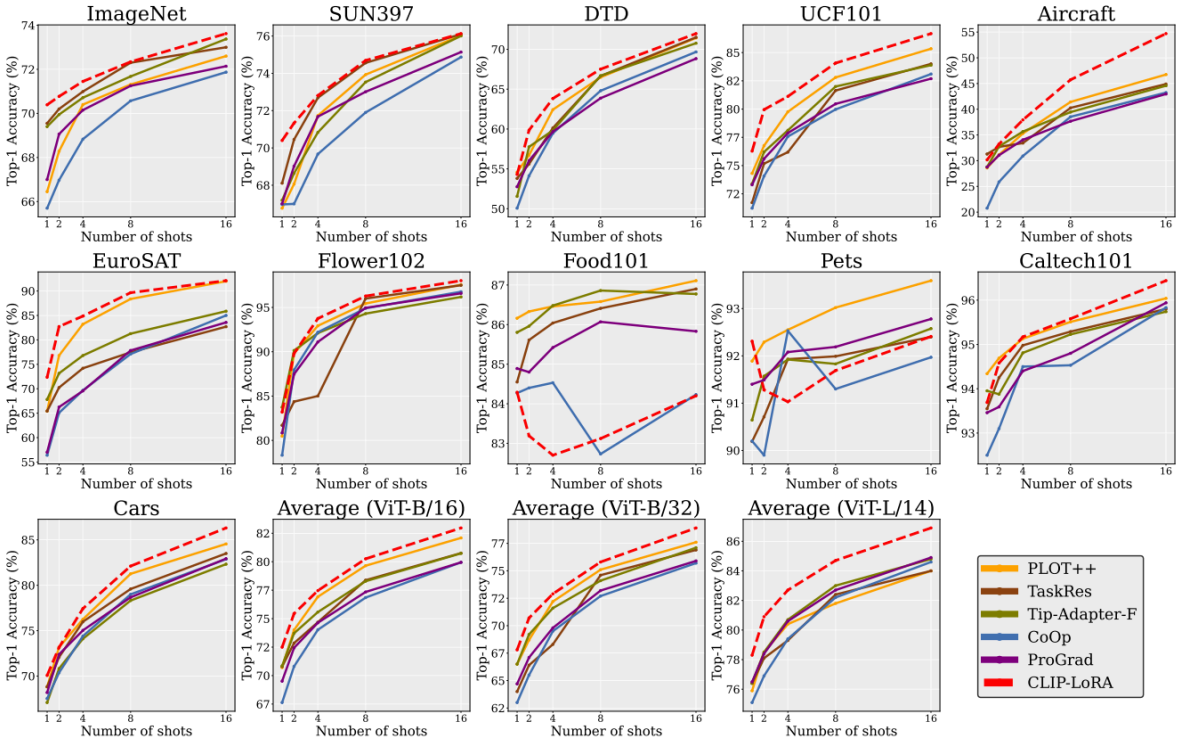
Shots	Method	ImageNet	SUN	Aircraft	EuroSAT	Cars	Food	Pets	Flowers	Caltech	DTD	UCF	Average
0	**CLIP** (ICML '21)	66.7	62.6	24.7	47.5	65.3	86.1	89.1	71.4	92.9	43.6	66.7	65.1
	CoOp (4) (IJCV '22)	68.0	67.3	26.2	50.9	67.1	82.6	90.3	72.7	93.2	50.1	70.7	67.2
	CoOp (16) (IJCV '22)	65.7	67.0	20.8	56.4	67.5	84.3	90.2	78.3	92.5	50.1	71.2	67.6
	CoCoOp (CVPR '22)	69.4	68.7	28.1	55.4	67.6	84.9	91.9	73.4	94.1	52.6	70.4	68.8
	TIP-Adapter-F (ECCV '22)	69.4	67.2	28.8	67.8	67.1	85.8	90.6	**83.8**	94.0	51.6	73.4	70.9
	CLIP-Adapter (IJCV '23)	67.9	65.4	25.2	49.3	65.7	86.1	89.0	71.3	92.0	44.2	66.9	65.7
1	PLOT++ (ICLR '23)	66.5	66.8	28.6	65.4	68.8	86.2	91.9	80.5	**94.3**	**54.6**	74.3	70.7
	KgCoOp (CVPR '23)	68.9	68.4	26.8	61.9	66.7	**86.4**	92.1	74.7	94.2	52.7	72.8	69.6
	TaskRes (CVPR '23)	69.6	68.1	**31.3**	65.4	68.8	84.6	90.2	81.7	93.6	53.8	71.7	70.8
	MaPLe (CVPR '23)	69.7	69.3	28.1	29.1	67.6	85.4	91.4	74.9	93.6	50.0	71.1	66.4
	ProGrad (ICCV '23)	67.0	67.0	28.8	57.0	68.2	84.9	91.4	80.9	93.5	52.8	73.3	69.5
	CLIP-LoRA (Ours)	**70.4**	**70.4**	30.2	**72.3**	**70.1**	84.3	**92.3**	83.2	93.7	54.3	**76.3**	**72.5**
	CoOp (4) (IJCV '22)	69.7	70.6	29.7	65.8	73.4	83.5	92.3	86.6	94.5	58.5	78.1	73.0
	CoOp (16) (IJCV '22)	68.8	69.7	30.9	69.7	74.4	84.5	92.5	92.2	94.5	59.5	77.6	74.0
	CoCoOp (CVPR '22)	70.6	70.4	30.6	61.7	69.5	86.3	92.7	81.5	94.8	55.7	75.3	71.7
	TIP-Adapter-F (ECCV '22)	70.7	70.8	35.7	76.8	74.1	86.5	91.9	92.1	94.8	59.8	78.1	75.6
	CLIP-Adapter (IJCV '23)	68.6	68.0	27.9	51.2	67.5	86.5	90.8	73.1	94.0	46.1	70.6	67.7
4	PLOT++ (ICLR '23)	70.4	71.7	35.3	83.2	76.3	86.5	92.6	92.9	95.1	62.4	79.8	76.9
	KgCoOp (CVPR '23)	69.9	71.5	32.2	71.8	69.5	**86.9**	92.6	87.0	95.0	58.7	77.6	73.9
	TaskRes (CVPR '23)	71.0	72.7	33.4	74.2	76.0	86.0	91.9	85.0	95.0	60.1	76.2	74.7
	MaPLe (CVPR '23)	70.6	71.4	30.1	69.9	70.1	86.7	**93.3**	84.9	95.0	59.0	77.1	73.5
	ProGrad (ICCV '23)	70.2	71.7	34.1	69.6	75.0	85.4	92.1	91.1	94.4	59.7	77.9	74.7
	CLIP-LoRA (Ours)	**71.4**	**72.8**	**37.9**	**84.9**	**77.4**	82.7	91.0	**93.7**	**95.2**	**63.8**	**81.1**	**77.4**
	CoOp (4) (IJCV '22)	71.5	74.6	40.1	83.5	79.1	85.1	92.4	96.4	95.5	69.2	81.9	79.0
	CoOp (16) (IJCV '22)	71.9	74.9	43.2	85.0	82.9	84.2	92.0	96.8	95.8	69.7	83.1	80.0
	CoCoOp (CVPR '22)	71.1	72.6	33.3	73.6	72.3	**87.4**	93.4	89.1	95.1	63.7	77.2	75.4
	TIP-Adapter-F (ECCV '22)	73.4	76.0	44.6	85.9	82.3	86.8	92.6	96.2	95.7	70.8	83.9	80.7
	CLIP-Adapter (IJCV '23)	69.8	74.2	34.2	71.4	74.0	87.1	92.3	92.9	94.9	59.4	80.2	75.5
16	PLOT++ (ICLR '23)	72.6	76.0	46.7	92.0	84.6	87.1	**93.6**	97.6	96.0	71.4	85.3	82.1
	KgCoOp (CVPR '23)	70.4	73.3	36.5	76.2	74.8	87.2	93.2	93.4	95.2	68.7	81.7	77.3
	TaskRes (CVPR '23)	73.0	**76.1**	44.9	82.7	83.5	86.9	92.4	97.5	95.8	71.5	84.0	80.8
	MaPLe (CVPR '23)	71.9	74.5	36.8	87.5	74.3	**87.4**	93.2	94.2	95.4	68.4	81.4	78.6
	ProGrad (ICCV '23)	72.1	75.1	43.0	83.6	82.9	85.8	92.8	96.6	95.9	68.8	82.7	79.9
	CLIP-LoRA (Ours)	**73.6**	**76.1**	**54.7**	**92.1**	**86.3**	84.2	92.4	**98.0**	**96.4**	**72.0**	**86.7**	**83.0**

Table 1. Training time on 16-shots ImageNet task. Experiments were conducted on a single A100 80Gb with the original code provided by the authors. For PLOT++ the time reported includes the 2 training stages.

Method	Training time
CoOp (16)	2h
PLOT++	15h30
ProGrad	3h20
CLIP-LoRA	50 min.

- 多くのデータセットでSoTAを出している
- FoodとPetsはKgCoOPが強かった
- 学習時間は他の手法の1/2以下

Shotごとの精度比較



- FoodとPetsでは学習が安定していない

アテンションLoRAごとの精度評価

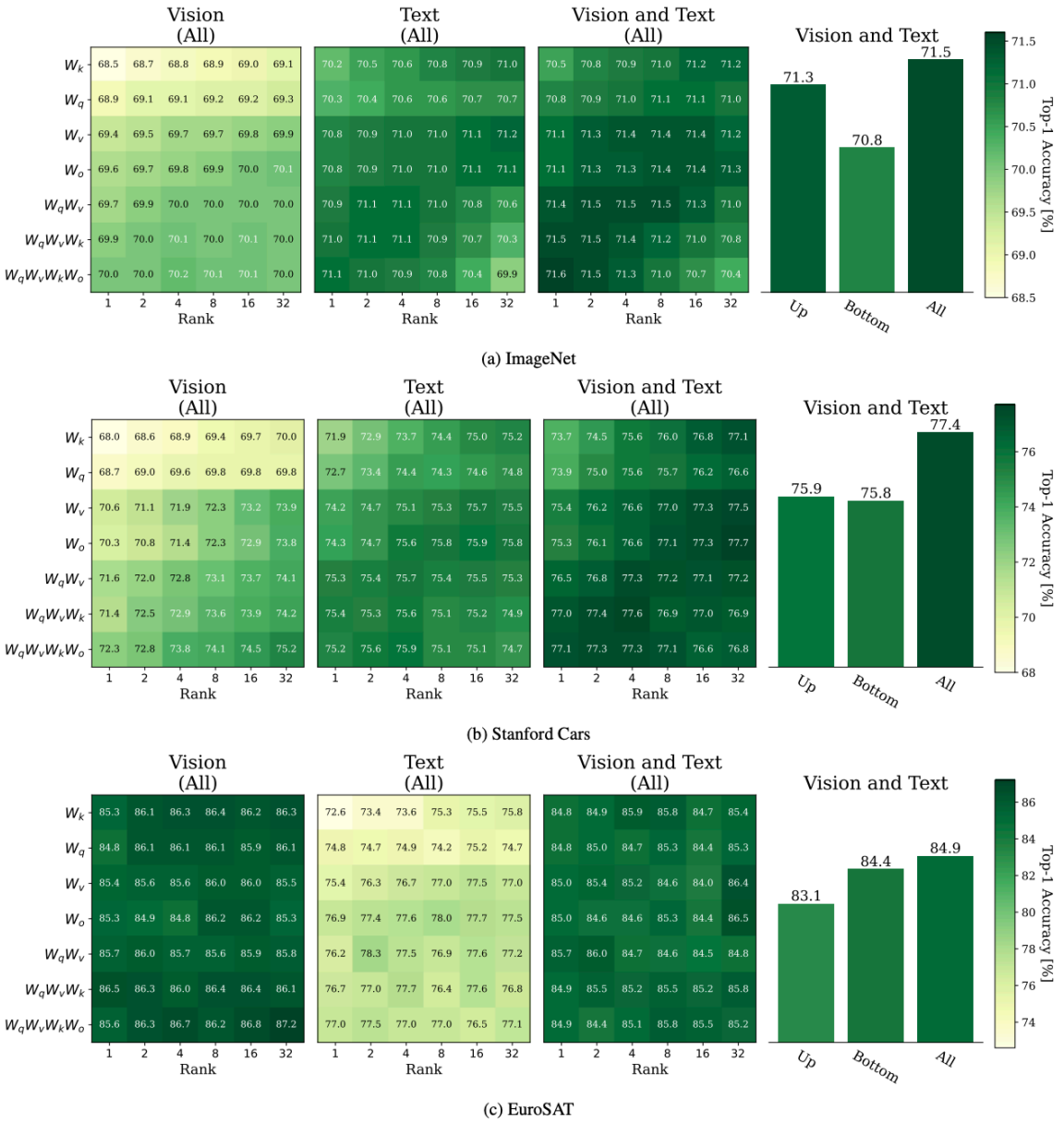


Figure 3. Top-1 accuracy with 4-shots for different matrices of the attention bloc and increasing rank, when the low-rank matrices are positioned at every level of the encoders (All). The fourth bar plot study the impact of positioning the low-rank matrices only on the half last levels (Up), the first half levels (Bottom), or at every level (All). Reported top-1 accuracy is averaged over 3 random seeds.

- CLIP-LoRAでLoRAのつけるところを何パターンか試した
- 総合的に見ると、VisionとTextの両方にLoRAをかましたほうが良さそう
- かつ、全体的にLoRAをつけたほうが精度は良い

結論

- VLMにおいてAdaptorベースやPromptベースではない新たなアプローチでのチューニング手法の提案

- 高効率かつ再現性の高い学習が可能
- 強力なベースラインの確立

所感

- 実験が非常に網羅的で参考になった。特にLoRAの適用位置に関する検証は興味深く、予想通りVisionとTextの両方に適用した方が安定して高い精度が出ていた
- 一方で、EuroSATのようにVision側だけでも高精度を発揮するデータセットもあり、ドメイン特化で運用する場合にはLoRAの適用位置を柔軟に調整する余地があると感じた
- 飲食系のドメインに応用したくてこの論文を読んだが、Food101で既存手法に精度で劣っていた点はやや残念だった。明示的な工夫が必要かもしれない
- しかし、計算効率がいいのは確かなので試してみる価値は十分にありそう
- 本研究はFew-Shot設定に焦点を当てていたが、数千枚規模のラベル付きデータがある場合の性能も気になる