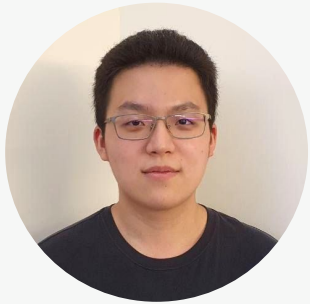
驚！不用兩分鐘就訓練完 Transformer！！

Multimodal Prompt Tuning:
Real-World Usage on NTU Tree Classification.

R11922147
林鴻儒

R11922A15
張仲喆

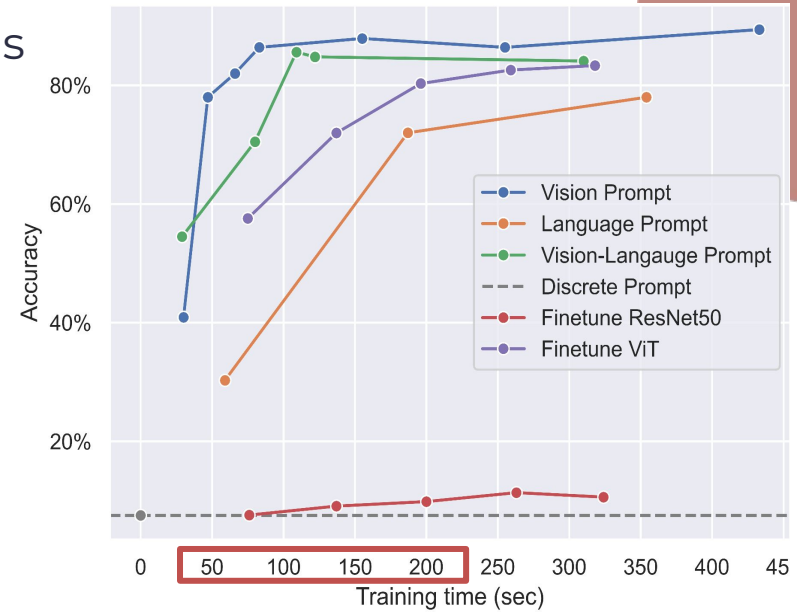


R11944004
李勝維

R11944021
廖金億

Abstract

- Collected a dataset: “Trees in NTU campus,” with >350 high quality images of 15 classes of tree in NTU for evaluation.
- Designed three different configurations of **soft-prompt on CLIP**.
- Evaluated the effectiveness of the **different methods under same training time budget**.



01

Introduction and Dataset

Dataset: Trees in NTU Campus

- Motivation: To know the environment is to know oneself.
- How to collect the dataset?
  - Using cellphone to capture images of trees in NTU campus.
  - Capturing the trees at different time and angle

    → maximize in-class variance

Dataset: Trees in NTU Campus

- Stem used as feature in 15-class dataset with ~350 images.
  - Texture classification task
- Training set: 8~16 images per class.

  Testing set: 8~10 images per class.

Royal Palm
大王椰子

Indigenous Cinnamon Tree
土肉桂

White Barkfig
垂榕

Hoop Pine
肯氏南洋杉

大王椰子	Royal Palm
土肉桂	Indigenous Cinnamon Tree
大葉桃花心木	Honduras Mahogany
小葉南洋杉	Araucaria Excelsa
石栗	Indian Walnut
朴樹	Chinese Hackberry
血桐	Macaranga
垂榕	White Barkfig
肯氏南洋杉	Hoop Pine
美人樹	Floss-silk Tree
烏桕	Chinese Tallow Tree
楓香	Formosan Sweet Gum
榕樹	Marabutan
蒲葵	Chinese Fan Palm
樟樹	Comphor Tree

All 15 classes

02

Related Work and Methodology

What Is Multimodal?

Modality: 模態

- 「现在我有冰淇淋」: Text is a kind of modality
-  : Image is a kind of modality
-  : Audio is a kind of modality
- Video, emotion, brainwave, optical flow, and more...

What Is Multimodal?

Human beings are multimodal in nature.

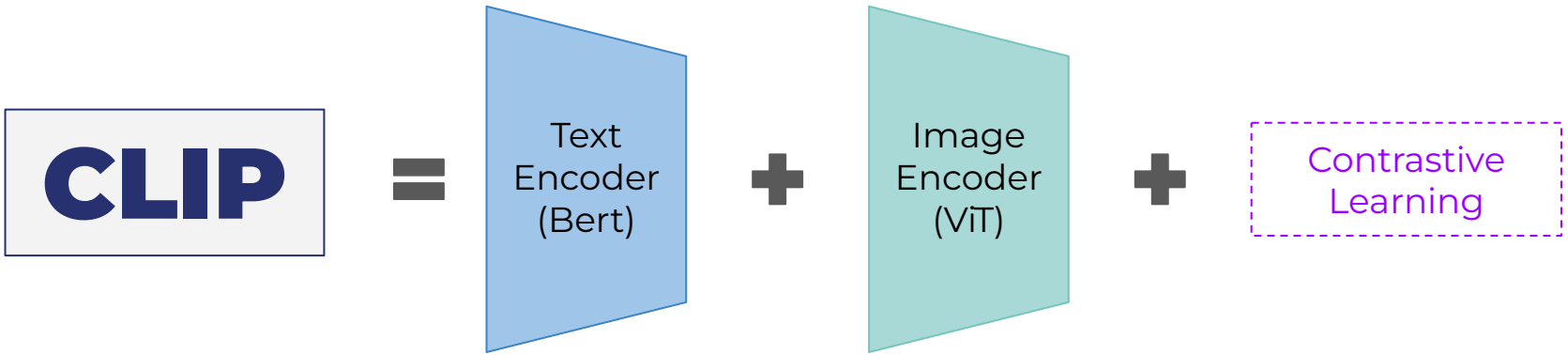
→ What about neural networks?

→ Transformers can process almost any modality.

Multimodal deep learning has been a growing trend:

- Generative model: Stable Diffusion, DALL-E, ...
- Contrative learning-based: CLIP, SimVLM, CoCa, ...
- V-L pretraining: BEiT, ViLBERT, ...
- You name it

CLIP: Connecting Vision and Language



Radford et al. Learning Transferable Visual Models From Natural Language Supervision. 2021

CLIP Training Objective

Note: CLIP is trained with
50M image-text pairs * **2500** V100 days

An image

Image Encoder

Image representation

Maximize cosine similarity

Paired text

A horse carrying a large load of hay and two people sitting on it

Text Encoder

Text representation

InfoNCE loss

Minimize cosine similarity

Unrelated image

Image Encoder

Image representation

CLIP Image Classification

Motivation:
CLIP shares the same semantic space behind different modalities.

ie: “a black cat” and “  ” has similar semantic/feature vectors

Idea:
CLIP can do zero-shot classification by giving it the **task description(prompt)**.

CLIP Image Classification

Text Encoder

Image Encoder

CLIP Image Classification

Handcrafted prompts

a photo of a flower

a photo of a Elon Musk

a photo of a dog

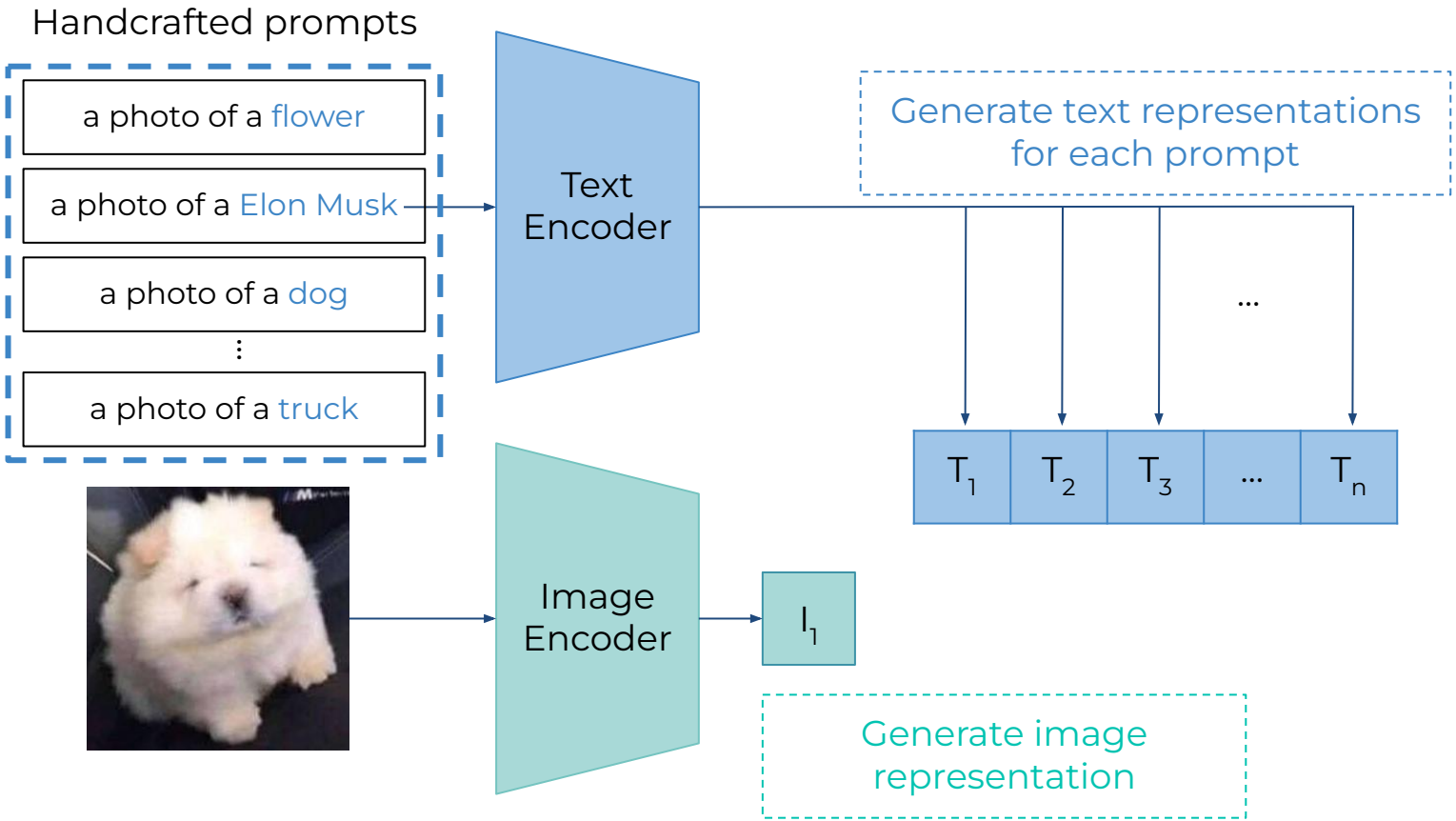
⋮

a photo of a truck

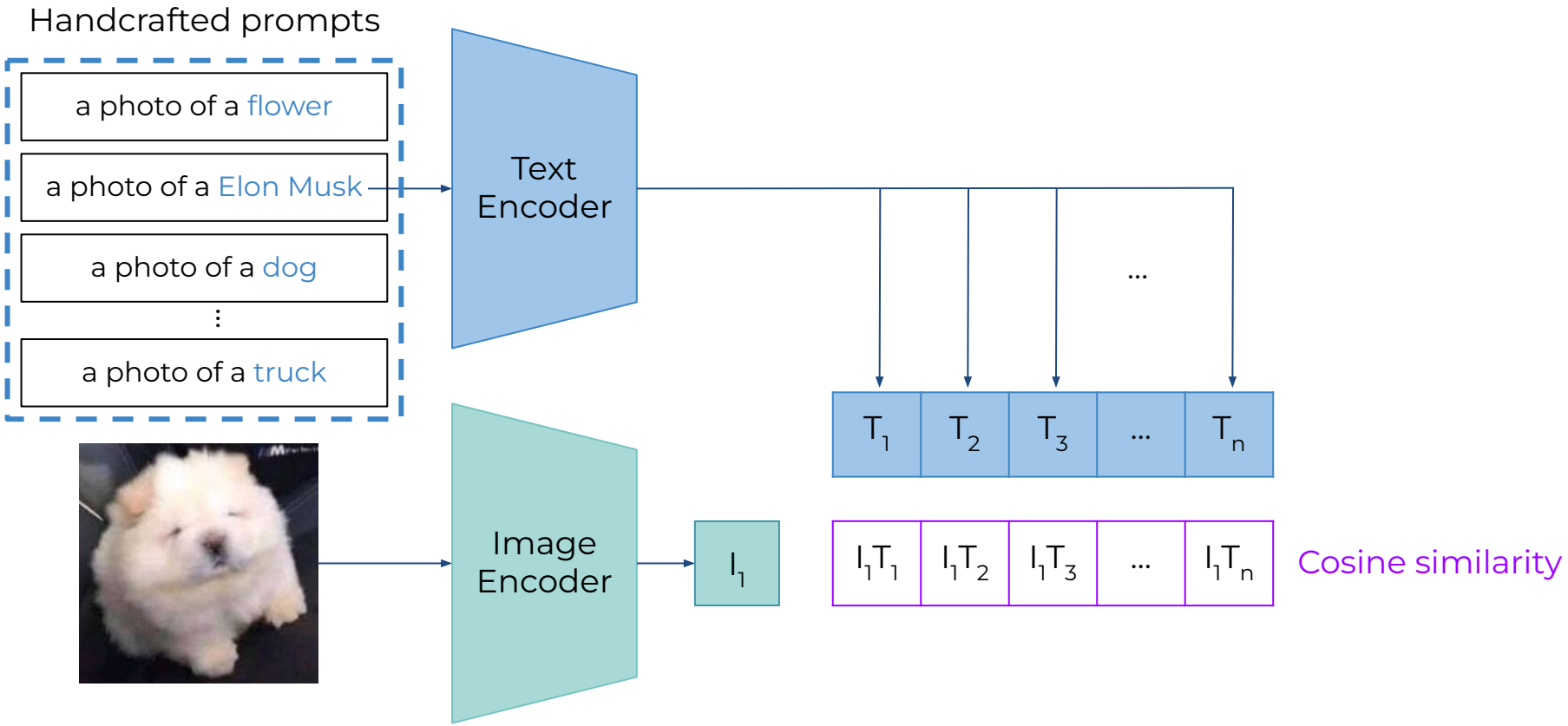
Text Encoder

Image Encoder

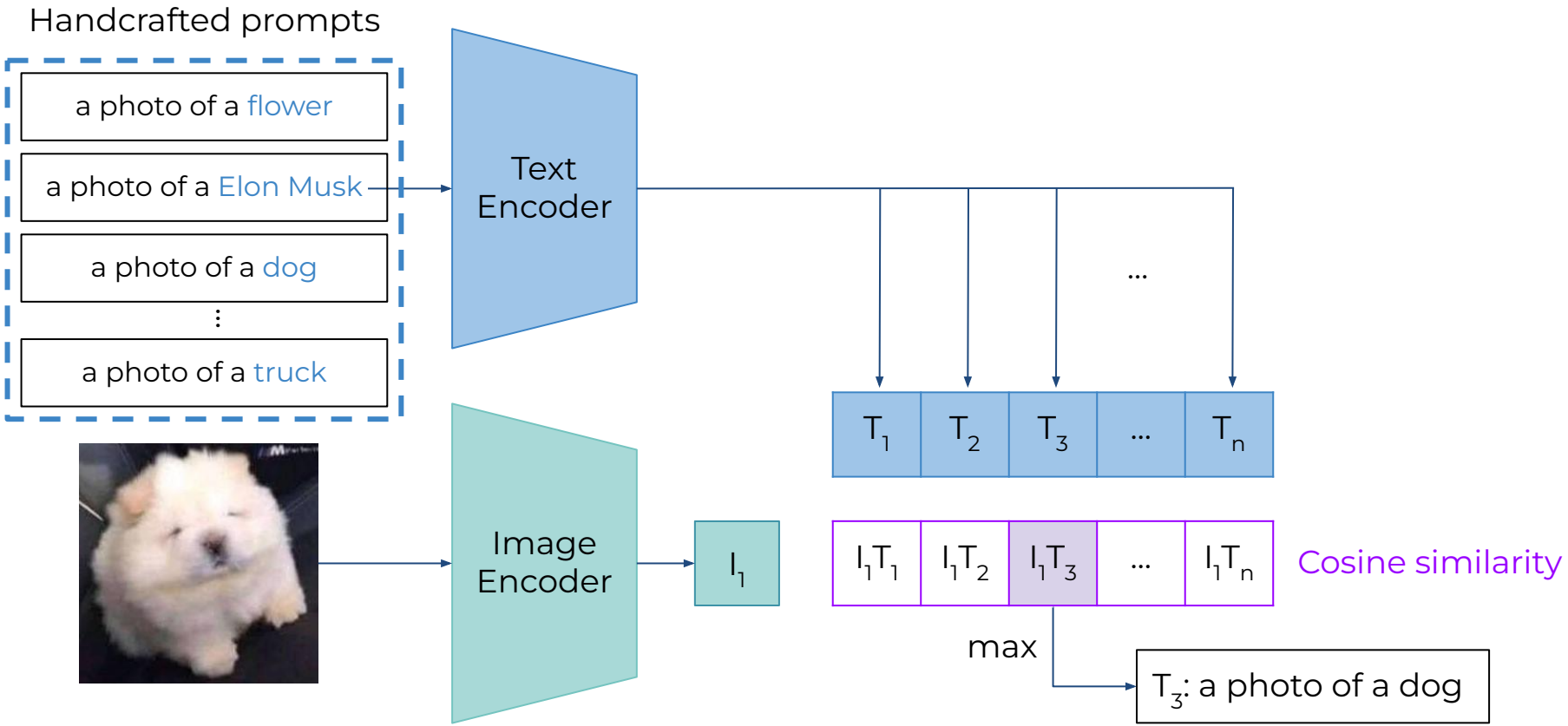
CLIP Image Classification



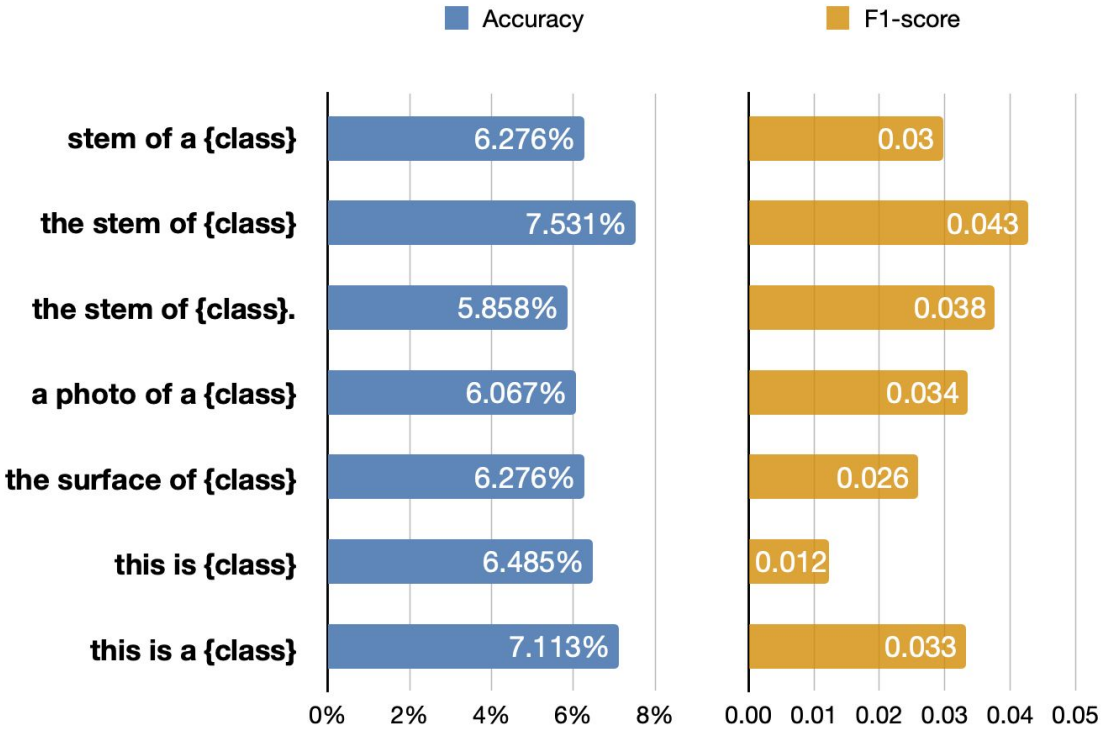
CLIP Image Classification



CLIP Image Classification



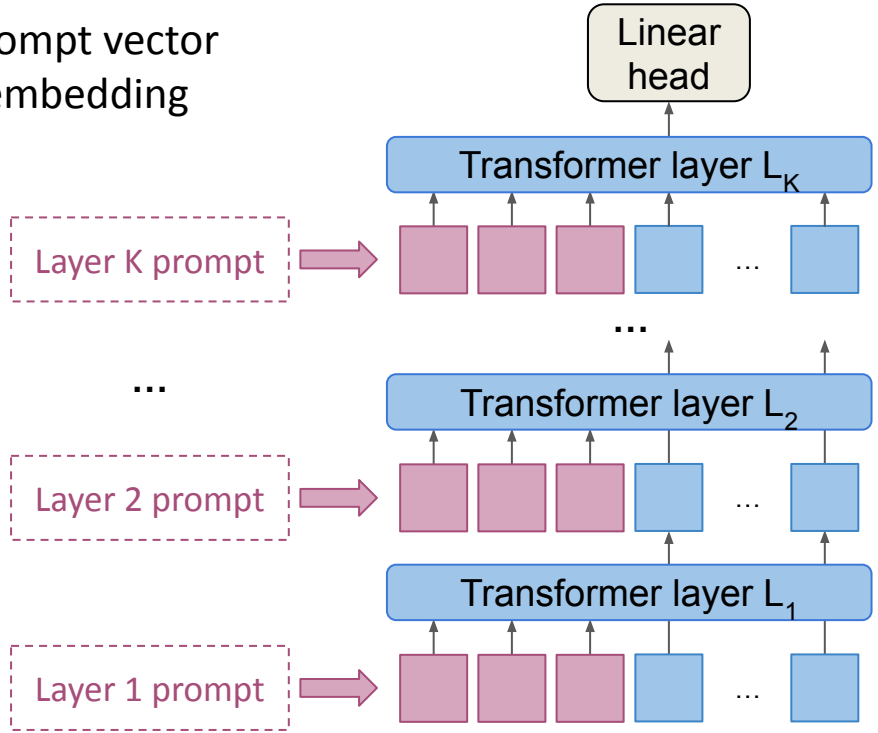
Handcrafted Prompts Are Suboptimal



LEARN the prompt
from data
→ Soft-prompt

Borrowing From NLP: P-Tuning v2

: Soft-Prompt vector
: Word embedding

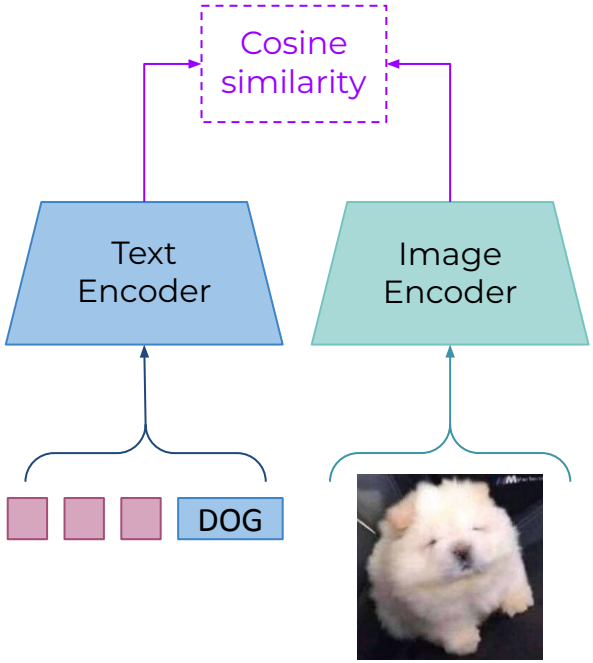


Liu et al. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks. ACL 2022

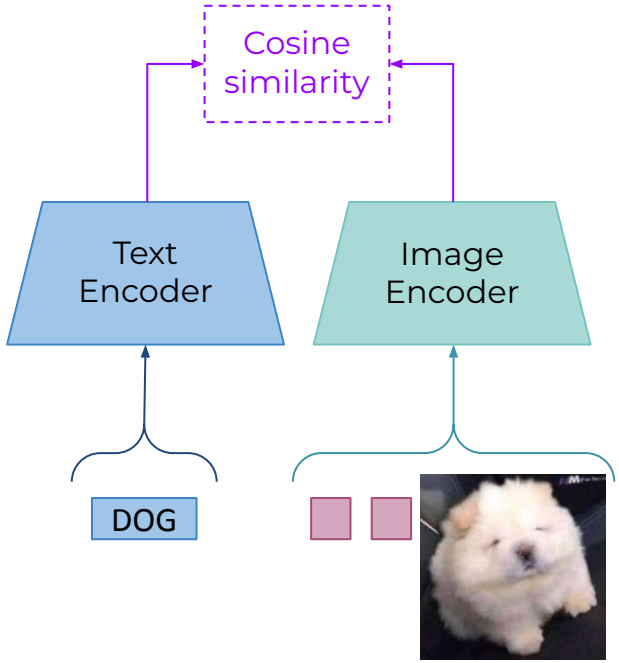
Three Soft-Prompting Strategies

: Soft-Prompt vector

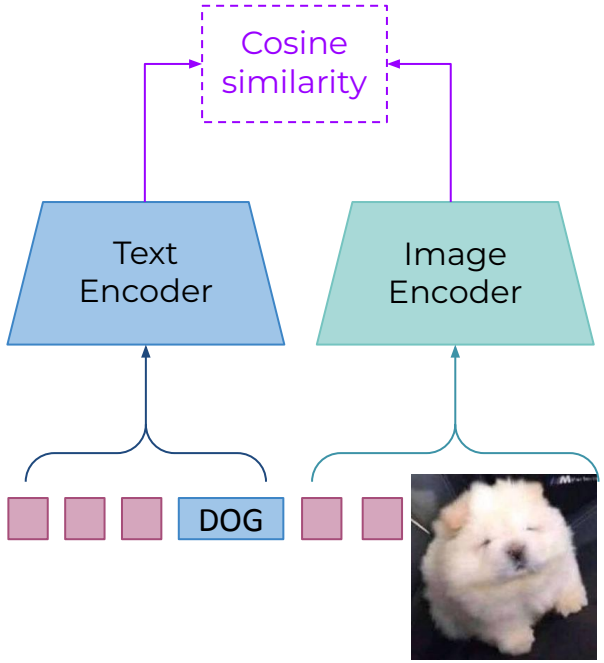
Language-only prompt



Vision-only prompt



Vision-Language prompt



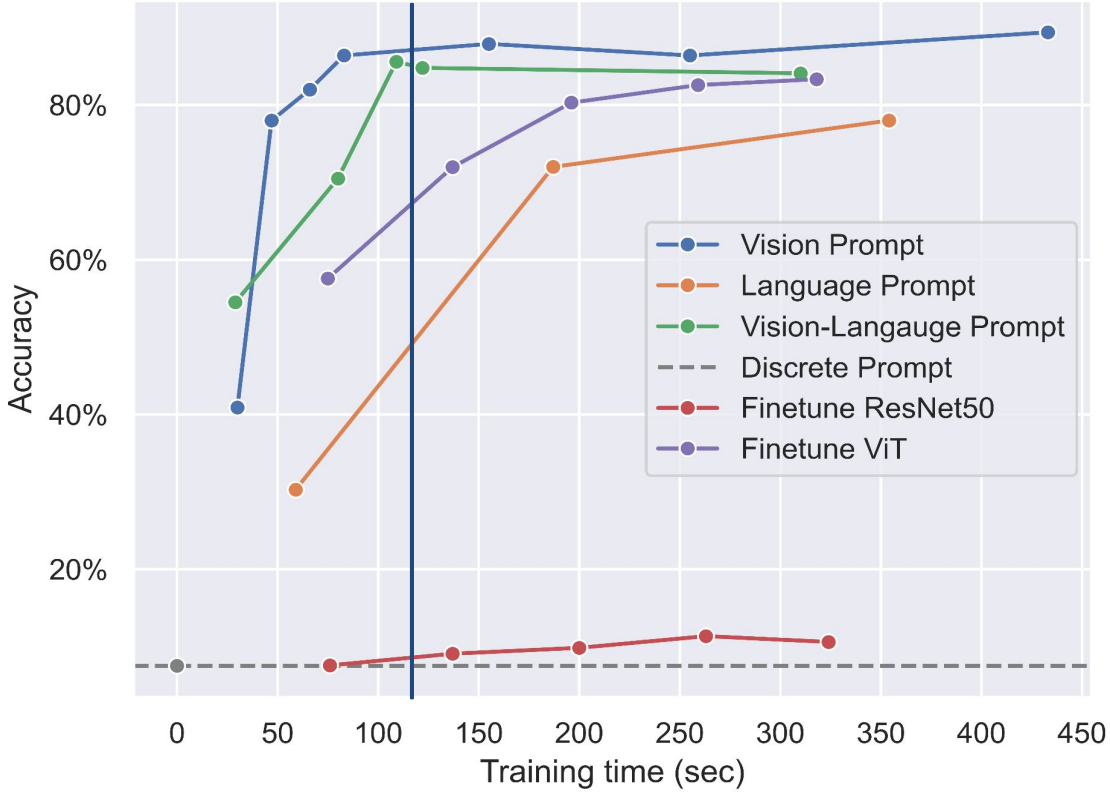
03 Experiments

Experiment Setting

- Comparing Four different prompting strategies:
  - Handcrafted discrete prompt
  - Vision-only soft prompt
  - Language-only soft prompt
  - Vision-Language soft prompt

  Plus two traditional fintuning methods:
  - Pretrained ResNet50
  - Pretrained Vision Transformer
- Every setting was trained multiple times to estimate its best performance under **various training time budgets**.

Vision Prompt Dominates



※ All trainings are conducted under the exact same hardware and software

Analysis

- Language prompt performs the worst:
  - Language does not provided useful informations in this instance.
  - Thus, V-L prompt performs slightly worse than vision prompt.
- Research* has shown that larger inter-class text variance makes vision prompt more effective.
  - The class names have very different word embeddings indeed.

Zang et al. Unified Vision and Language Prompt Learning. ICLR 2023 Submission

04 Conclusion

Conclusion

- Vision Prompt dominates this dataset.
- Soft-Prompt reach higher accuracy than finetuning with **limited computing resource and training data**.
- Soft-Prompt converges faster than finetuning.

Reference

- Radford et al. Learning Transferable Visual Models From Natural Language Supervision.
- Khattak et al. Maple: Multi-Modal Prompt Learning
- Zang et al. Unified Vision and Language Prompt Learning
- Paper under double-blind review. Rethinking the Value of Prompt Learning for Vision-Language Models
- Zhou et al. Learning to Prompt for Vision-Language Models
- Zhou et al. Conditional Prompt Learning for Vision-Language Models
- Liu et al. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing

Thank You