Project title: Visual Prompt Tuning
Team Name: MochaMinds
Members: Disha Hegde, Henry Wong, Vashisth Tiwari

Milestone Introduction to Deep Learning

Disha Hegde†, Henry Wong†, Vashisth Tiwari†

† Equal contribution

April 13, 2024

This is an explorative study building off the Visual prompt tuning paper [Jia et al., 2022]. Our GitHub for the project is `https://github.com/Vashistht/Project-Visual-Prompt-Tuning`

The results of the ablation studies can be accessed at:

- Varying Prompt Locations
- Varying the Number of Prompt Tokens

Project title: Visual Prompt Tuning
Team Name: MochaMinds
Members: Disha Hegde, Henry Wong, Vashisth Tiwari

1. **What do you expect to have done by this Milestone? (Copy-Paste from your project proposal)**

   Implementation of VPT-Shallow: This includes baseline metrics (accuracy and IOU on various tasks). As elaborated on above, we do not plan on delivering VPT-Deep due to limited compute. After Milestone #3, we will commence ablation studies and potential improvements.

2. **Progress and achievements:**
   Describe the progress made towards the milestone. Include any challenges faced, how they were overcome, and any adjustments made to the original project plan. Provide figures/tables to support your progress.

   - Reproduced results from the paper for ViT-Base, along with results for Vit-Small and ViT-Tiny (not implemented in the original paper). Due to the prompt training time for each model (Table 2) we have decided to use ViT-Tiny as our baseline for further ablations.

Model	Test CUB Accuracy (%)
ViT-Tiny	73.2
ViT-Small	84.69
ViT-Base (paper)	86.7
ViT-Base	86.56

Table 1: ViT Models and Test Accuracy

Recreating the results for the ViT base took us around 20 hours on the Tesla T-4. This did not seem practical for conducting ablation studies. Therefore, we added ViT-Small (115 MiB) and ViT-Tiny (37 MiB) [Dosovitskiy et al., 2021]

Model Size	Training Time (hours)
Base	20
Small	11
Tiny	3

Table 2: Time Taken to Train Different ViT Models on CUB Dataset

   - Reproduced the baselines on CUB dataset
   - Performed comparison of prepending vs adding prompts for ViT-Tiny

     We re-created the ablation on prompt location, comparing adding vs prepending prompts and inserting prompts in latent vs pixel space. The following operations are listed below:

     - Add: add prompts element-wise to embeddings

  - Prepend: prepend prompts to sequence of embeddings
  - Prepend in pixel space: prepend prompts in pixel space

  As shown in the VPT paper, prepending prompts in the latent (embedding) space achieves the highest test and validation accuracies of 72.56% and 72.17% respectively. Going forward, we will be prepending prompts.

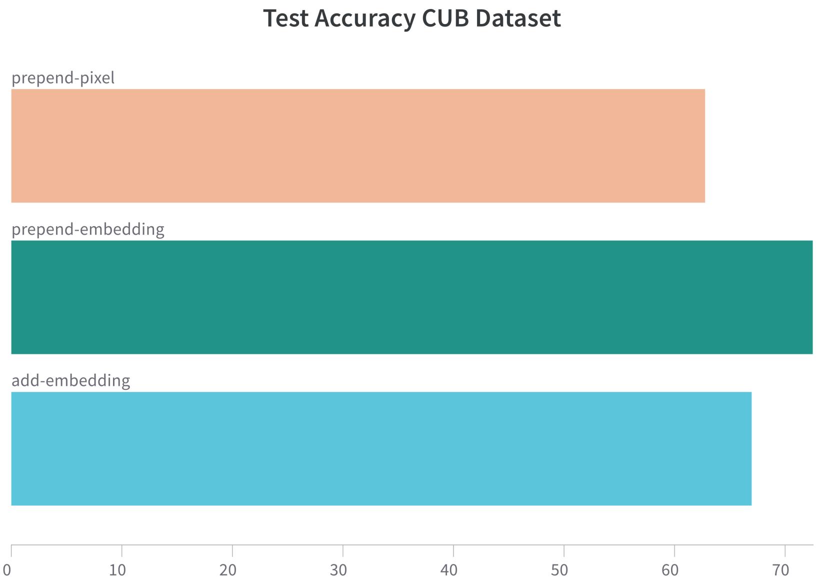

Figure 1: Test Accuracy on CUB Dataset with respect to different prompt locations

- Performed sweep across prompt sizes (token count) for ViT-Tiny

  We also conducted the ablation on prompt sizes varying the number of prompt tokens from 25-150 with the intervals of 25 tokens per prompt. Our findings were that for the CUB dataset, there is **minimal difference across prompt sizes, with 50 prompts and 100 prompts having similar accuracies.** We will be running the prompt size sweep across multiple datasets to confirm this result.

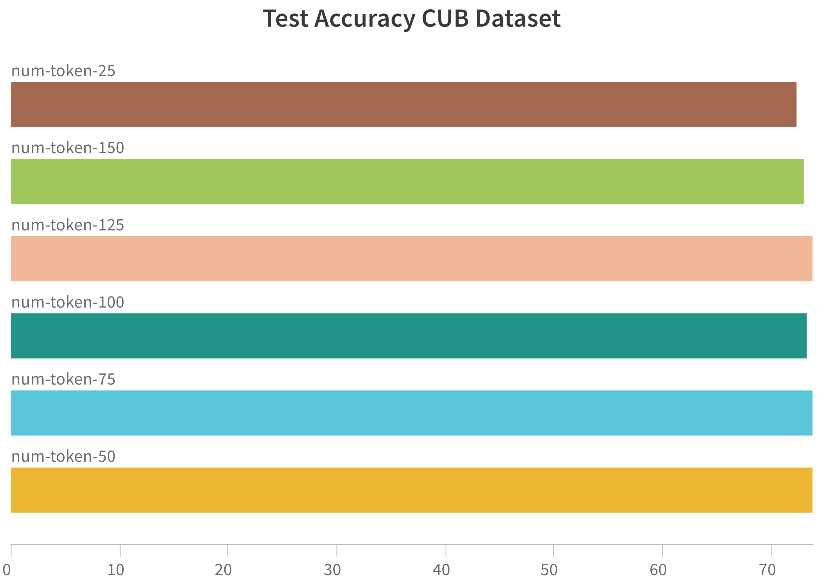

Figure 2: Test Accuracy on CUB Dataset for Different Number of Prompt Tokens

3. **Self-evaluation:**
   Reflect on your current progress and evaluate your own work.

   We have met all proposed goals for this milestone. As a team, we met consistently throughout the two weeks to ensure progress. We also made sure to reach out to the professor and our mentor TA to receive feedback on project scope and expectations; in doing so, we set realistic goals and pruned our project accordingly.

4. **Self-Grade:**
   On a scale from 0 to 5, where 0 is no progress, 1 is minimal progress, 2.5 reflects progress towards half your goals, and 5 meets all proposed goals, rate yourselves.

   4.8/5

5. **Next Step:**
   Outline the next steps in the project plan, including any adjustments or revisions to the original plan that have been made based on progress to date. Include a timeline to complete the following steps and what your project plan is to achieve by the end of the semester.

   - We will not be re-implementing the VPT from scratch, as time and compute restraints would not allow us to perform the more interesting ablation studies if we do so. Additionally, we were advised by the professor, Aswin, and our mentor TA, Anwesa, to not re-implement VPT and instead focus on performing ablations.
   - We will continue using ViT-Tiny, having established a baseline with which to

perform comparison studies, as opposed to comparing against the baseline in the reference work (ViT-Base).

- Apply transfer learning to prompts for new tasks (seeing how using prompts from one task perform over random initializations in the paper) and across datasets (seeing how prompts from one dataset perform on other datasets on the same task)
- Instead of resorting to VPT-Deep, which has been shown to have significantly better performance than VPT-Shallow, investigate how many Transformer Encoder Layers should have a prepended set of learnable parameters before VPT performance saturates

6. **Effort allocation:**
If your team has more than one member, please provide a brief description of responsibilities undertaken by each member.

Our team met in-person on multiple occasions to collaborate on the project. We all contributed to setting up VPT and integrating the new backbones (ViT-Tiny and ViT-Small). Specific tasks undertaken by each member are listed below:

- Disha: Recreated ablations for prompt token sizing.
- Henry: Recreated ablations for prompt location.
- Vashisth: Set up framework for ablation testing and data logging; researched potential improvements past Milestone 3.

References

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning, 2022.