

Supplementary material for Catching the Unknown with Limited Data: Bi-Directional Prompt Tuning in CLIP for Few-Shot Open-Set Adaptation

Moloud Abdar¹
m.abdar1987@gmail.com

Md Mehedi Hasan²
mmhasan@deakin.edu.au

Biplab Banerjee³
getbiplab@gmail.com

Abbas Khosravi²
abbas.khosravi@deakin.edu.au

Pietro Lio⁴
pl219@cam.ac.uk

¹ CHIRP, Child Health Research Centre,
The University of Queensland,
Brisbane, Australia

² Institute for Intelligent Systems,
Deakin University,
Geelong, Australia

³ Centre of Studies in Resources
Engineering,
Indian Institute of Technology Bombay,
Mumbai, India

⁴ Department of Computer Science and
Technology,
University of Cambridge,
Cambridge, UK

1 Preliminaries on CLIP

CLIP [1] is a pre-trained vision-language (V-L) model which has a vision and text encoder. The text encoder is a transformer and the vision encoder is either ResNets [2] or the vision transformer (ViT) [3]. Following existing prompting methods [4, 5, 6, 7], we focus on ViT-based CLIP model. The image encoder decomposes the image I into M non-overlapping fixed-size patches. Each patch is mapped to a patch embedding $P_0 \in \mathbb{R}^{M \times d_p}$, where d_p denotes the dimension of the embeddings. These patch embeddings are sequentially processed through the K transformer layers along with the learnable class (CLS) token. At each layer i , the patch embeddings P_{i-1} are augmented with a learnable class token c_{i-1} , forming the input for the next transformer block T_i :

$$[c_i, P_i] = T_i([c_{i-1}, P_{i-1}]), i = 1, 2, \dots, K. \quad (1)$$

The output of the final transformer block is the class token c_K , which is then mapped to a high-dimensional image feature representation V using a projection function:

$$V = \text{ImageProj}(c_K), \quad (2)$$

where $V \in \mathbb{R}^{d_v}$ and d_v is the dimensionality of the image representation. The text encoder processes the tokenized words by mapping them to embeddings $E_0 = [e_0^1, e_0^2, \dots, e_0^N] \in \mathbb{R}^{N \times d_t}$, where N is the number of tokens and d_t is the embedding dimension. These embeddings undergo K transformer layers, with each layer updating the embeddings through a function S_i ,

$$E_i = S_i(E_{i-1}), i = 1, 2, \dots, K. \quad (3)$$

The final word embeddings corresponding to the last token of the final transformer block are projected into a shared embedding space, generating the final text representation U :

$$U = \text{TextProj}(e_K^N), \quad (4)$$

where $U \in \mathbb{R}^{d_v}$.

For zero-shot classification, each class $y \in \{1, 2, \dots, C\}$ is associated with a natural language prompt (e.g., "a photo of a dog"). For a given image I with its corresponding representation V , the predicted class \hat{y} is obtained by calculating the normalized similarity between the image and all class prompts. The probability of predicting a class \hat{y} given an image I is expressed as:

$$p(\hat{y}|V) = \frac{\exp(\cos(V, U_{\hat{y}})/\tau)}{\sum_{i=1}^C \exp(\cos(V, U_i)/\tau)}, \quad (5)$$

where $\cos(V, U_i)$ represents the cosine similarity between the image representation V and the class text embedding U_i , and τ is a temperature parameter that controls the sharpness of the similarity scores.

2 Experimental Setup

Datasets. We evaluate our proposed method on three benchmark image classification datasets: Office-Home [8], DomainNet [9], MiniImageNet/CUB [9, 10]. Office-Home is a large-scale dataset for visual cross-domain classification with four distinct domains. For our experiments, we only choose Real-World and Clipart domains. DomainNet is a large-scale cross-domain benchmark with 345 classes and six domains. Following [9], we evaluate our method on the following source-target combinations: Clipart to Painting, Real to Clipart, and Real to Painting. MiniImageNet is a benchmark dataset derived from the larger ImageNet dataset and contains a subset of 100 classes. CUB (Caltech-UCSD Birds) consists of fine-grained bird species classification. MiniImageNet/CUB focuses on the generalization capability of a model from a diverse but generic dataset (MiniImageNet) to a specialized, fine-grained dataset (CUB).

Training and Evaluation Protocol. We take inspiration from the evaluation protocol in [9]. We assume that four samples per class are available in the source domain and one sample per class in the target domain during fine-tuning. We fine-tune our model using the support dataset for 25 epochs using Adam [21] with a learning rate of 0.002, and a batch size of 32. All images are resized to 224x224 pixels. Due to the scarcity of data in the target domain, we employ standard data augmentation techniques, including random horizontal flips, random cropping. In all of our experiments, we utilize the CLIP model with ViT-B/32 as its image encoder. To ensure a fair evaluation, we report the average performance across 10 distinct runs for each dataset and each method.

Baselines. We compare our method with state-of-the-art prompt learning baselines such as CoCoOp [12], CoOp [13], KgCoOp [14], MaPLe [8], and PromptSRC [9]. For further

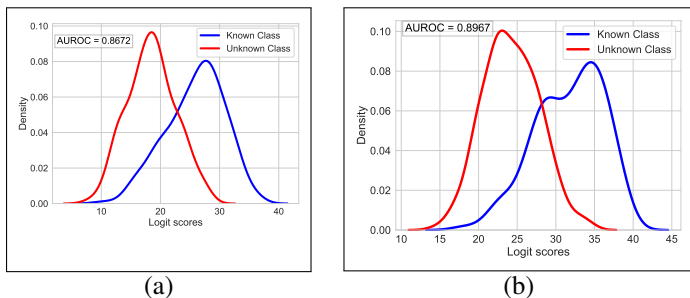


Figure 1: Density plot using the maximum logit scores generated by the models: (a) MaPLE (b) Ours. The figure is generated using Office-Home dataset.

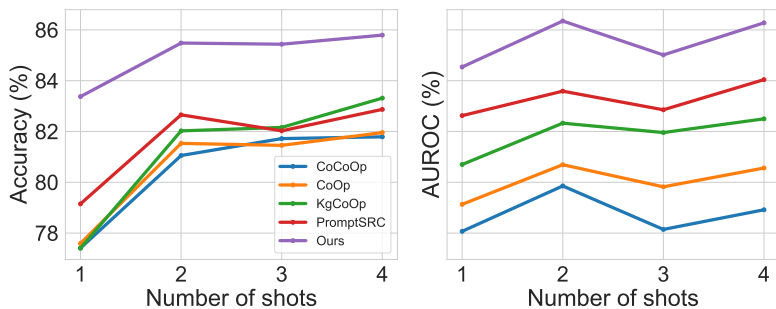


Figure 2: Effect of the number of shots in the target domain. Our method outperforms other state-of-the-art approaches by a significant margin, especially in the lower shot region.

comparison, we also include a recently introduced non-CLIP-based method called DAFOSNET [14].

Metrics. As we focus on distinguishing known samples from unknown samples in both the source and target domains in addition to detection accuracy, we use two metrics to evaluate the performance of our method. Accuracy for recognition capability and area under the receiver operating characteristic curve (AUROC) to identify the open set samples. AUROC measures the model’s ability to distinguish between known and unknown classes under varying decision thresholds. A higher AUROC indicates better open-set recognition performance, reflecting the model’s robustness in correctly identifying samples from novel or unseen categories. To evaluate the model’s performance across domains, we calculate the individual scores for each domain and report the harmonic mean of these scores. The harmonic mean provides a balanced measure, ensuring that the reported performance reflects both domains equally and penalizes large disparities between the two.

3 More Results

Figure 1 demonstrates the effectiveness of the maximum logit scores in detecting the open-set sample. Overall, the results validate the contributions of our method’s components and design choices.

Figure 2 illustrates the performance of various prompt-based methods with varying numbers of examples per class (shots) in the target domain during fine-tuning. All methods show an overall increasing trend in both accuracy and AUROC as the number of shots increases. This is expected as more training data generally leads to better learning opportunities for the model. Our method provides consistent improvements over existing methods, particularly in the lower-shot regimes. The performance gap between methods seems to decrease as the number of shots increases. This suggests that with sufficient training data, the differences between methods may become less pronounced.

4 Attributes

We generated the discriminative attribute using GPT-3.5 models. The prompt used to generate them was giving some discriminative attributes of the following class names'. We present some of the generated attributes below.

```

{
  'alarm_clock': 'circular_or_square_face , large_numbers , clock_hands ',
  'backpack': 'rectangular_or_oval_shape_with_handle , straps_for_shoulders , zippers , pockets , buckles ',
  'batteries': 'cylindrical_shape , polarity_indicators , brand_or_label_markings ',
  'bed': 'large_flat_surface_with_rectangular_frame , headboard_and/or_footboard , elevated_off_the_ground ',
  'bike': 'two_wheels , frame_with_pedals , handlebars , chain_and_gear_system ',
  'bottle': 'cylindrical_shape_or_flask-shaped_or_bell-shaped , with_a_neck_and_cap ',
  'bucket': 'cylindrical_shape_with_an_open_top , handle_for_carrying ',
  'calculator': 'number_pad_with_function_keys , rectangular_shape , display_screen_for_numbers ',
  'calendar': 'grid_layout , often_shows_numbers_as_months_and_days ',
  'candles': 'cylindrical_or_tapered_shape , flame_at_the_top , wax_material ',
  'chair': 'four_legs , rectangular_seat , and_backrest ',
  'clipboards': 'smooth_flat_surface , clip_at_the_top_to_hold_paper ',
  'computer': 'rectangular_screen_and_keyboard , rectangular_box_for_CPU_and_hard-disk ',
  'couch': 'wide_seating_area_with_cushions , armrests_and_backrest ',
  'curtains': 'draping_fabric_attached_to_rod , pleated_or_folded_design ',
  'desk_lamp': 'adjustable_arm_with_light_source , attached_to_a_base ',
  'drill': 'handheld_tool_with_rotating_bit , often_has_a_trigger_button , cordless_or_wired ',
  'eraser': 'small_rectangular_or_cylindrical , soft_rubber_material , often_pink_or_white ',
  'exit_sign': 'rectangular_or_square_with_exit_in_bold_letters , often_red_or_green ',
  'fan': 'blades_enclosed_in_a_grill , base_or_stand , oscillating_or_fixed ',
  'file_cabinet': 'rectangular , often_metallic , multiple_drawers , label_holders_or_handles ',
  'flipflops': 'open-toe_design , thong_strap_over_the_foot , flat_sole ',
  'flowers': 'petals_surrounding_a_central_stem , variety_of_colors , often_with_leaves_or_stems ',
  'folder': 'flat_rectangular_with_pockets_or_tabs , often_made_of_paper_or_plastic , designed_to_hold_documents ',
  'fork': 'handle_with_pronged_end , typically_metallic , used_for_eating ',
  'glasses': 'two_lenses_connected_by_a_frame , two_arms_resting_on_ears , often_transparent_lenses ',
  'hammer': 'handle_with_a_weighted_head , claw_for_removing_nails , metallic_head_with_wooden_or_plastic_handle ',
  'helmet': 'rounded_shell_with_internal_padding , straps_to_fasten_under_the_chin , protective_outer_layer ',
  'kettle': 'rounded_body_with_spout_and_handle , often_metallic_or_plastic , used_to_heat_water ',
  'keyboard': 'rectangular_layout_of_keys , letters , numbers , and_symbols , connected_to_a_computer ',
  'knives': 'sharp_blade_with_handle , vary_in_size (chef's , utility , etc.) , metallic_or_ceramic_blade ',
  'lampshade': 'conical_or_cylindrical_covering , covers_the_bulb_of_a_lamp , diffuses_light ',
  'laptop': 'clamshell_design_with_screen_and_keyboard , hinged_at_the_middle , portable ',
  'marker': 'cylindrical_with_a_felt-tip_end , cap_to_cover_the_inked_tip , variety_of_colors ',
  'monitor': 'rectangular_screen , often_sits_on_a_stand , display_for_computers ',
  'mop': 'long_handle_with_ absorbent_head , used_for_cleaning_floors , often_with_detachable_heads ',
  'mouse': 'palm-sized_device_with_buttons , scroll_wheel_and_sensor_on_the_bottom , used_for_computer_input ',
  'mug': 'cylindrical_with_a_handle , often_ceramic_or_metal , used_for_hot_drinks ',
  'notebook': 'rectangular_with_lined_or_blank_pages , spiral-bound_or_stitched , used_for_writing ',
  'oven': 'large_appliance_with_a_door , heating_elements_inside , often_includes_racks ',
  'pan': 'shallow_with_flat_bottom , handle_for_holding , used_for_cooking ',
  'paper_clip': 'small_bent_wire , U-shaped_for_holding_paper , metallic_or_colored_plastic ',
  'pen': 'cylindrical_with_an_inked_tip , click_or_cap_mechanism , used_for_writing ',
  'pencil': 'thin_cylindrical_wood_or_mechanical_body , graphite_tip_for_writing , eraser_on_one_end ',
  'post_it_notes': 'small_square_or_rectangular_paper , adhesive_on_the_back , brightly_colored_for_quick_notes ',
  'printer': 'rectangular_with_paper_tray , output_slot_for_printed_documents , often_connected_to_a_computer ',
  'push_pin': 'small_head_with_pointed_metal_tip , used_to_attach_things_to_boards , plastic_or_metallic_head ',
  'radio': 'box-like_structure_with_dials_or_buttons , speakers_and_antenna , may_have_a_display_for_tuning ',
  'refrigerator': 'large_box_with_door(s) , cooling_compartment_inside , shelves_and_drawers ',
  'ruler': 'flat , straight_edge , marked_with_measurements , often_plastic_or_metal ',
  'scissors': 'two_blades_joined_at_a_pivot , handles_with_finger_holes , used_for_cutting ',
  'screwdriver': 'handle_with_a_long_shaft , flat_or_cross-shaped_tip , used_to_drive_screws ',
  'shelf': 'flat_surfaces_attached_to_a_wall_or_frame , used_for_storage_or_display , often_wood_or_metal ',
  'sink': 'basin_with_faucet_for_water , often_installed_in_kitchens_or_bathrooms , drains_at_the_bottom ',
  'sneakers': 'closed-toe_with_laces_or_velcro , rubber_sole_for_grip , used_for_casual_wear_or_sports ',
  'soda': 'cylindrical_or_bottle-shaped , label_with_brand_and_ingredients , often_carbonated ',
  'speaker': 'box_with_mesh_covering , emits_sound , often_connected_to_audio_devices ',
  'spoon': 'handle_with_a_bowl-shaped_end , used_for_eating , typically_metallic_or_plastic ',
  'tv': 'large_rectangular_screen , often_with_a_stand_or_wall-mounted , display_for_video_content ',
  'table': 'flat_surface_with_legs , vary_in_size_and_shape , used_for_placing_objects ',
  'telephone': 'handset_with_buttons_or_dial , often_connected_by_a_wire_or_wireless , used_for_communication ',
  'toothbrush': 'handle_with_bristles_at_the_end , used_for_cleaning_teeth , often_made_of_plastic ',
  'toys': 'small_objects_designed_for_play , often_colorful , varied_shapes_and_materials ',
  'trash_can': 'cylindrical_or_rectangular_with_an_open_or_covered_top , used_for_waste_disposal , often_plastic_or_metal ',
  'webcam': 'small_camera_lens_attached_to_a_frame , often_attached_to_computers_or_monitors , used_for_video_calls '
}

```

References

- [1] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [3] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

-
- [4] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023.
- [5] Debabrata Pal, Deeptej More, Sai Bhargav, Dipesh Tamboli, Vaneet Aggarwal, and Biplab Banerjee. Domain adaptive few-shot open-set learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18831–18840, 2023.
- [6] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.
- [7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [8] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.
- [9] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.
- [10] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. -, 2011.
- [11] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023.
- [12] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.
- [13] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.