SVL-Adapter: Self-Supervised Adapter for Vision-Language Pretrained Models

Omiros Pantazis[1]
omiros.pantazis.16@ucl.ac.uk
Gabriel Brostow[1,4]
g.brostow@cs.ucl.ac.uk
Kate E. Jones[1]
kate.e.jones@ucl.ac.uk
Oisin Mac Aodha[2,3]
oisin.macaodha@ed.ac.uk

[1] University College London
[2] University of Edinburgh
[3] Alan Turing Institute
[4] Niantic

Abstract

Vision-language models such as CLIP are pretrained on large volumes of internet sourced image and text pairs, and have been shown to *sometimes* exhibit impressive zero- and low-shot image classification performance. However, due to their size, fine-tuning these models on new datasets can be prohibitively expensive, both in terms of the supervision and compute required. To combat this, a series of light-weight adaptation methods have been proposed to efficiently adapt such models when limited supervision is available. In this work, we show that while effective on internet-style datasets, even those remedies under-deliver on classification tasks with images that differ significantly from those commonly found online. To address this issue, we present a new approach called SVL-Adapter that combines the complementary strengths of both vision-language pretraining and self-supervised representation learning. We report an average classification accuracy improvement of 10% in the low-shot setting when compared to existing methods, on a set of challenging visual classification tasks. Further, we present a fully automatic way of selecting an important blending hyperparameter for our model that does not require any held-out labeled validation data. Code for our project is available here: https://github.com/omipan/svl_adapter.

1 Introduction

Learning transferable representations of visual data is a core problem in computer vision. Until recently, the standard approach for tasks like supervised image classification involved training models to predict the discrete class labels depicted in a collection of images. However, we have begun to see a new set of approaches that make use of large-scale image and natural language text pairs collected from the internet as a source of training signal. Methods such as CLIP [43] and ALIGN [31] pretrain on hundreds of millions of text and image pairs, from which they exhibit impressive transfer in both the zero- and low-shot settings [29].

The challenge of using models of this size is that fine-tuning them to new tasks can be prohibitively computationally expensive and potentially require significant amounts of data.

Inspired by work in natural language processing [32, 47], a series of subsequent methods have been proposed in vision to adapt these large-scale models in a data-efficient manner. Approaches include optimizing for the most effective text prompt [30, 36, 64] or refining the learned representations with compact adapter networks [20, 61]. However, the majority of these methods do not significantly change the underlying visual representations contained in models such as CLIP. As a result, they are fundamentally limited by the expressiveness of the original representations. This can be problematic if one wants to adapt these internet-trained models to new tasks that differ from the types of images commonly found on the internet (see Fig.1), *e.g.* medical image analysis [33], remote sensing [11], biodiversity monitoring [5], *etc.*. Unlike curated image collections such as ImageNet [14], the aforementioned datasets exhibit more real-world challenges such as partial views, occlusion, poor illumination, diverse backgrounds, low image quality, and domain shifts. These properties make it difficult to perform transfer learning from models trained on conventionally curated content.

Our focus is on making CLIP-like models more effective on challenging tasks that potentially fall out of the distribution of the visual or language content that they were originally trained on. First, we show that existing methods that tune prompts or visual features, *e.g.* [20, 36, 61, 64], perform poorly on these types of challenging datasets. Next, we observe that recent advances in Self-Supervised Learning (SSL), *e.g.* [9, 21, 24], provide a complementary approach for learning visual representations in an self-supervised manner that can be combined with the outputs of models such as CLIP. This is especially relevant in practical applications where images are typically available, but supervision is lacking. Building on these observations, and inspired by existing adapter methods [20], we propose a new method for visual classification in the low-shot regime and validate it across multiple challenging visual classification tasks. Our approach also addresses a significant limitation of many existing adapter methods, which is their requirement for held-out labeled validation data for hyperparameter selection. We present an automated hyperparameter selection method that foregoes the need for labeled data by making use of model predictions directly.

Our contributions are summarized as follows: 1) We present a new visual classification approach, SVL-Adapter, that combines the best of both large-scale vision-language pretraining and targeted self-supervised learning. 2) We outline a method for selecting a key hyperparameter for our model that does not require obtaining any expensive held-out labeled validation data as is commonly done in existing related works. 3) Through detailed experimental evaluation on ten conventional and six challenging visual classification tasks, we show that our SVL-Adapter outperforms existing methods on average for both zero- and low-shot learning, and is significantly better on challenging datasets whose visual properties differ from those commonly found online.

2 Related Work

2.1 Adapting Vision-Language Models

Large pretrained *language* models have demonstrated remarkable success across a variety of natural language tasks from question-answering and sentence completion to language translation [7, 15, 42]. These models are powerful, but are consequently expensive to train or fine-tune owing to their large size. Thus, there is a need for efficient adaptation methods to enable transfer to new datasets. Multiple works have sought remedies by devising lightweight adapter modules containing a small number of trainable parameters [27, 28, 50], bias mitigation calibration [63], or by learning task-specific soft text prompts [35]. As a re-

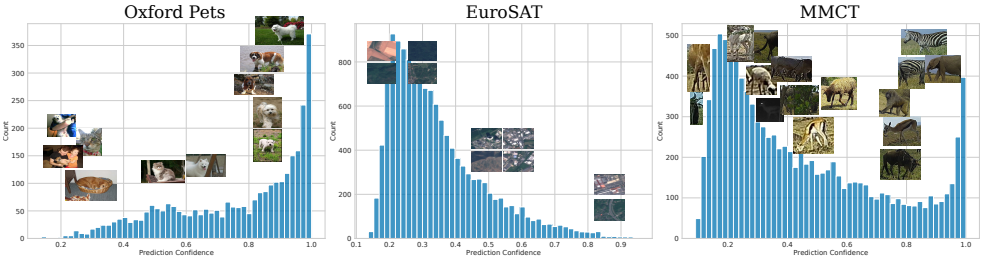

Figure 1: Histograms of confidence scores from zero-shot predictions of CLIP [43]. Representative images for different confidence bins are also displayed. (Left) CLIP is confident for images from datasets representative of the types commonly found online. (Middle) However, when the data distribution is significantly different, *e.g*. satellite images, it tends to have low confidence. (Right) In practice, real-world datasets can contain a mix of 'easy' and 'hard' images, *e.g*. from camera traps.

sult, practitioners can harness the power and generality of large pretrained language models on some downstream tasks of interest.

Recently, a related line of work has been explored in the *vision-language* domain through a series of approaches that take advantage of the large number of images with corresponding text descriptions that can be found online [31, 43, 58, 59]. For example, CLIP [43] utilizes contrastive learning between the embeddings of image and text pairs, which are encoded via separate image and language encoders. Essentially, CLIP learns to represent the two modalities by pulling together the image representations with their paired text ones, while repelling the text embeddings corresponding to different images. The advantage of these approaches is that they can learn from the large unstructured collections of images and text which are readily available on the internet [31].

Vision-language models trained on large datasets can be applied to a wide range of downstream tasks in either the zero- or low-shot setting. In the zero-shot classification setting, the practitioner must select the set of relevant text descriptions corresponding to the classes of interest. [43] showed that how these text prompts are constructed can have a significant impact on the downstream performance of models such as CLIP. Drawing inspiration from language model adaptation, two main families of approaches for adapting CLIP-like models to downstream visual tasks have been proposed: prompt learning and feature adaptation. Both these methods typically rely on the existence of a small number of labeled examples for each class in the given target dataset, *i.e*. the low-shot setting.

Prompt learning seeks to address the major limitation of having to hand-craft the text prompts, by automating how they are constructed. CoOp [64] optimizes the prompt context using a set of learnable vectors in either a unified or class specific way. CoCoOp [65] extended this, addressing the difficult problem of generalizing to unseen classes by learning to generate vectors conditioned on each image. In contrast, feature adaptation approaches directly tune the representations that are extracted from the visual and text encoders of models like CLIP. CLIP-Adapter [20] added a lightweight fully connected neural network *adapter* that is applied to frozen CLIP features and performs fine-tuning of its parameters with limited supervision on the downstream task of interest. Performance improvements are reported across a variety of datasets, with the best performing variant only adapting the visual encoder features. Tip-Adapter [61] obtains even better results by constructing a key-value

cache model from the low-shot samples, and fine-tunes for a smaller number of epochs. A tuning-free version of Tip-Adapter was also proposed, which is faster at adapting at training time, but performs worse.

While relatively efficient, these adapter style approaches are incapable of making large-scale changes to the underlying representations extracted from the backbone visual or text encoders. This poses a significant issue if the images from an evaluation task of interest significantly differ from the distribution seen by the vision-language model at training time. In our proposed approach, we do not restrict feature extraction to only the visual encoders learned by CLIP. Instead, we propose to combine the impressive zero-shot performance of CLIP and features extracted via targeted self-supervised learning.

2.2 Self-Supervised Learning (SSL)

SSL aims to learn high quality visual representations using only the signal contained in unlabeled images, *i.e*. without the need for additional supervision provided by humans. Successive advances in SSL have further shrunk the gap between supervised and unsupervised representations, when evaluated on some downstream tasks [8, 9, 17, 21, 24, 60]. One representative family of approaches is contrastive learning [22, 39, 56], where the goal is to embed augmented views of a given image close in feature space, while pushing away the representations of other images in the same batch [9] or using a memory bank [24]. Methods have also been explored that focus on retrieving more informative positive examples during training that exhibit more natural image variation than can be expressed by simple artificial augmentations [7, 8, 40]. Other variants on the contrastive paradigm also report strong performance without the inclusion of any negative examples during training [10, 21]. There are also further SSL approaches that are not limited to instance discrimination, but instead use information from nearest neighbors [17], prototype clustering [8], and image patch reconstruction [25].

Most relevant to this work is the observation that SSL has been shown to be effective on multiple 'real-world' tasks, *e.g*. remote sensing [7], medical image analysis [4], and biodiversity monitoring [40]. In these application domains, it is sometimes impossible or prohibitive to even obtain large collections of unlabeled images. However, [13] showed that self-supervised training on ImageNet [14] is still highly effective even when only using $< 25\%$ of the unlabeled images at training time. In this work, we leverage these advances in SSL by learning representations on unlabeled data, which we combine with the outputs from large vision-language models, the result of which is comparable or better than either alone.

2.3 Few-Shot Image Classification

Being able to learn from limited examples poses a great challenge for computer vision that many few-shot learning (FSL) approaches have attempted to tackle, *e.g*. [19, 44, 45, 46, 48, 54, 55, 62]. For example, meta-learning, a prominent solution for few-shot recognition [19, 44, 45, 46, 62], utilizes a meta-learner to transfer learned knowledge from a support set of classes to enable it to perform few-shot classification on new classes. Interestingly, [62] showed that training a linear classifier on top of features learnt from SSL can outperform sophisticated meta-learning methods in few-shot image classification. Furthermore, [29] showed that a very simple FSL method trained on vision-language model features [43] easily beats existing FSL methods. Thus, the success of transfer learning from models that are pre-trained on large-scale external data inspires us to focus our research on the efficient adaptation of these powerful models for few-shot, and zero-shot, learning.

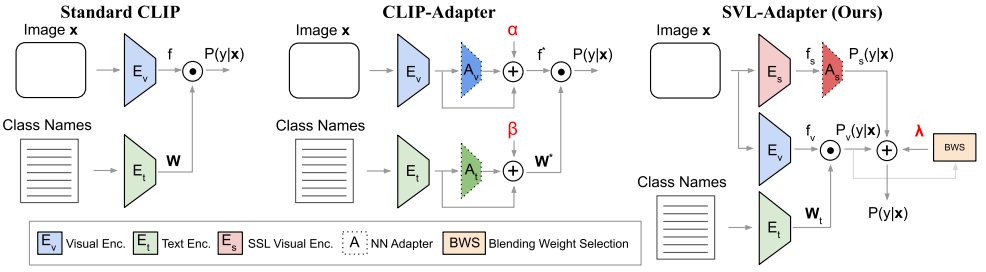

Figure 2: (Left) Zero-shot classification with CLIP [43] makes use of frozen image and text encoders, E_v and E_t, to make predictions at test time. (Middle) CLIP-Adapter [20] introduces two small additional adapter neural networks, A_v and A_t, whose outputs are combined with the visual and text embeddings from CLIP using hyperparameters α and β. (Right) Our SVL-Adapter makes use of an additional self-supervised *encoder* E_s which feeds into an adapter A_s. Unlike CLIP-Adapter, our approach fuses outputs at the class prediction level, where the weighting λ of this combination is automatically controlled by our blending weight selection module.

3 Method

3.1 Framework

Given an image $\mathbf{x}$ as input, our goal is to predict the class label $y \in \{1,...,K\}$ depicted in the image. The conventional approach would be to train an image classifier, parameterized as a deep neural network, to perform this task. However, in many real world settings we typically only have access to a limited amount of training data (*i.e.* low-shot learning), making it difficult to perform the end-to-end training of large models. To address this issue, recent work has shown that pretraining models with paired image and natural language supervision is an effective technique for learning representations that can be later adapted to downstream tasks with minimal additional supervision [31, 43, 58].

CLIP: One such representative approach is CLIP [43] (see Fig. 2 (Left)). CLIP consists of two main parts: a visual encoder E_v that takes an image as input and outputs a feature embedding $\mathbf{f} = E_v(\mathbf{x})$, and a text encoder E_t that encodes the set of class names into the same embedding space, $\mathbf{W} = E_t(\mathcal{C})$. Here, $\mathbf{W} \in \mathbb{R}^{D \times K}$ is a matrix where each of the D dimensional column vectors corresponds to an embedding of the K classes of interest at inference time, *i.e.* $\mathcal{C} = \{C_1,...,C_K\}$. Each entry C_k, is the natural language class name corresponding to the class label y. To make a prediction for an unseen test image $\mathbf{x}$, one simply performs a matrix multiplication of its feature embedding $\mathbf{f}$ and the estimated classifier weight matrix $\mathbf{W}$, *i.e.* $P(y|\mathbf{x}) = \sigma(\mathbf{f}^\intercal \mathbf{W})$, where σ is the softmax function. The visual encoder can be implemented as a ResNet [23] or a visual transformer [16], and the text encoder as a transformer [53].

CLIP-Adapter: As presented, CLIP cannot make use of additional labeled data without performing expensive end-to-end fine-tuning or simply using the visual encoder as a fixed feature extractor. Several subsequent methods have been proposed to efficiently adapt large pretrained models like CLIP to the low-shot learning setting. One such example, is CLIP-Adapter [20] (see Fig. 2 (Middle)). CLIP-Adapter performs fine-tuning on light-weight residual feature adapters to improve classification performance when only modest supervision is available. This is achieved via an additional pair of small fully connected neural

network adapters A_v and A_t, for visual and text feature adaption respectively. Specifically, $A_v(\mathbf{f}) = \text{ReLU}(\mathbf{f}^\intercal \mathbf{W}_v^1)\mathbf{W}_v^2$ and $A_t(\mathbf{W}) = \text{ReLU}(\mathbf{W}^\intercal \mathbf{W}_t^1)\mathbf{W}_t^2$.

The output feature embedding and classifier weight matrix are then computed as

$$\mathbf{f}^* = \alpha A_v(\mathbf{f})^\intercal + (1-\alpha)\mathbf{f} \tag{1}$$

$$\mathbf{W}^* = \beta A_t(\mathbf{W})^\intercal + (1-\beta)\mathbf{W}. \tag{2}$$

The final prediction is then computed as $P(y|\mathbf{x}) = \sigma(\mathbf{f}^{*\intercal}\mathbf{W}^*)$. During low-shot training, they need to estimate the weights, $\{\mathbf{W}_v^1, \mathbf{W}_v^2, \mathbf{W}_t^1, \mathbf{W}_t^2\}$, of the two adapters. This is done by applying a standard cross entropy loss on the low-shot labeled data.

3.2 SVL-Adapter

One major limitation of approaches like CLIP and CLIP-Adapter is they do not make significant changes to the underlying representations encoded by the visual encoder. The reason for this is simple: fine-tuning large models requires lots of supervision, much more than what is available in the low-shot setting. This is not necessarily an issue if the images for the downstream classification task come from the same distribution as those commonly found on the internet. However, if the images differ significantly, *e.g.* medical or biodiversity monitoring images, then the representations that can be extracted from the visual encoder stand a good chance of being unsuitable for the task at hand.

We exploit the fact that while little to no label supervision may be available, we often have access to the unlabeled images at test time. Recent progress in self-supervised learning [9, 21, 24, 25] has resulted in methods that can extract informative visual representations without requiring any supervised labels. To avail of these representations, we propose a new approach, the Self-supervised Vision-Language Adapter (SVL-Adapter) that combines the strengths of both vision-language pretraining and self-supervised learning (see Fig. 2 (Right)).

We introduce a new encoder E_s, that is trained using self-supervision on the target dataset. The output of this encoder is a feature vector $\mathbf{f}_s$ that is fed into an adapter network A_s. Unlike CLIP-Adapter, the output of this adapter is not a transformed feature encoding, but instead a prediction over the classes of interest, $P_s(y|\mathbf{x}) = A_s(\mathbf{f}_s) = \sigma(\text{ReLU}(\mathbf{f}_s^\intercal \mathbf{W}_s^1)\mathbf{W}_s^2)$. We then combine these predictions with the output of the standard Zero-shot CLIP model, $P_v(y|\mathbf{x}) = \sigma(\mathbf{f}_v^\intercal \mathbf{W}_t)$,

$$P(y|\mathbf{x}) = \lambda P_v(y|\mathbf{x}) + (1-\lambda)P_s(y|\mathbf{x}). \tag{3}$$

During training we learn the weights $\{\mathbf{W}_s^1, \mathbf{W}_s^2\}$. We train E_s on the target training dataset using a self-supervised constrastive objective [9], which does not require any labeled data. While training E_s on large datasets could be expensive, in practice we start from an ImageNet initialized model which leads to fast convergence on relatively small downstream datasets. Also this step only has to be performed once as only the adapter A_s needs to be retrained when the amount of supervision available changes.

3.2.1 Blending Weight Selection

The best results for CLIP-Adapter in [20] are obtained when the hyperparameters α and β in Eqns. 1 and 2 are selected using held-out validation data. In the case of low-shot learning, any and all labeled data is precious and would likely be more valuable to use for training and not hyperparameter selection.

To overcome this issue, we propose a conceptually simple and efficient approach for selecting our prediction blending weight λ in Eqn. 3 which does *not* require any labeled validation data. From Fig. 1 we observe that the confidence scores corresponding to the Zero-shot CLIP predictions vary heavily among datasets. Based on this, and under the assumption that when CLIP is *not* confident we should more heavily weigh the knowledge acquired by low-shot learning, we define λ as being analogous to CLIP's average prediction confidence score on the N test images of the given dataset, $\lambda = \frac{1}{N}\sum_{i=1}^{N} \max_k P(y_i = k|\mathbf{x}_i)$. In our experiments we compare our SVL-Adapter method to existing approaches in the setting where we select λ using validation data ('SVL-Adapter') or where we estimate it using the outputs from CLIP as outlined above ('SVL-Adapter*').

4 Experiments

4.1 Implementation Details

Datasets. We evaluate our approach on ten standard image classifications datasets: Caltech101 [18], OxfordPets [41], StanfordCars [34], Flowers102 [38], Food101 [6], FGVCAircraft [37], SUN397 [57], DTD [12], UCF101 [49], and EuroSAT [26], that are typically utilized to test vision-language adaptation. Additionally, we include six challenging tasks that do *not* come from the distribution of images commonly found on the internet. These datasets are: FMoW [11] that contains satellite images of land or buildings, OCT [33] for retina disease identification from OCT images [33], and the camera trap datasets MMCT [40], CCT20 [5], ICCT [1], and Serengeti [51].

Models. To train the self-supervised feature encoder E_s on each dataset for our SVL-Adapter, we use SimCLR [9] for 200 epochs on images of size 112×112 with a standard ResNet50 [23] backbone followed by a two-layer projection head. SSL takes on average two hours for each dataset evaluated. We provide additional results using alternative self-supervised methods in the supplementary material. The trainable adapter module A_s in SVL-Adapter is a two-layer neural network with 256 hidden dimensions and an output size equal to the number of classes, and is trained for 50 epochs. When not using our blending weight selection, λ is tuned with a validation set, by sweeping through 20 values ranging from 0 to 1. For CLIP, we use the ResNet50 visual encoder E_v and transformer text encoder E_t. For each of the standard datasets we use a single prompt template as defined in [43] and for the rest we define prompts that do not rely on any in-domain expertise. For the transformation of CLIP features we follow the pre-processing protocol of CLIP [43]. We provide results for various ablations of our model in the supplementary material.

Baselines. To evaluate our approach, we compare with zero-shot and linear probe CLIP [43] and state-of-the-art vision-language adaptation baselines [20, 61, 64]. Specifically, we include the best performing version of CoOp [64] that places the class token at the end of the learnable sequence without class-specific context, the best variant of CLIP-Adapter [20] that only adapts the visual features, and both the training-free and fine-tuned version of the more recent Tip-Adapter [61]. Across all scenarios, the features of the encoders stay frozen and tuning is only performed on the adapters. For low-shot learning, we construct 1, 2, 4, 8, and 16 examples per class training sets. For Zero-shot CLIP we simply apply CLIP's ResNet50 variant to the test images. For our Zero-shot SVL-Adapter*, training resembles the few-shot task but uses the most confident CLIP pseudolabels per predicted label instead. In our experiments we keep the 16 most confident predictions for each class. Evaluation always takes place on the full test set for each dataset.

4.2 Results

4.2.1 Low-Shot Classification

First we evaluate our SVL-Adapter across the multiple datasets described above in the low-shot regime with different amounts of supervision, *i.e*. number of 'shots' per class. To validate the hypothesis that images that differ from internet-style datasets need a different approach we split the datasets into two distinct sets which we term "Standard" and "Challenging". In Fig. 3 we observe that the top performing existing method, Tip-Adapter-F, reports less than 50% Top-1 accuracy on average in the 16-shot setting across the "Challenging" tasks, while its corresponding performance on the "Standard" datasets is above 75%. This discrepancy motivates the development of approaches that can retain the benefits of CLIP while adapting more efficiently to real-world tasks that do not exhibit internet-like visual properties. A more detailed, per-dataset, breakdown is presented in Fig. 4. We also provide these results in tabular for in the supplementary material.

To this end, we show that SVL-Adapter, which uses self-supervised features as a starting point for visual adaptation, results in significant accuracy gains (10% on average) across the "Challenging" tasks when compared to existing methods, where the largest gains are observed in the 2-shot setting and above. In addition, we observe that our approach still remains competitive when applied to the "Standard" datasets, and thus it constitutes a strong and universal baseline. Moreover, SVL-Adapter*, which uses our automatic blending weight selection method outlined in Sec. 3.2.1, is comparable with SVL-Adapter on the "Challenging" tasks despite not requiring any labeled validation data. However, it is not as strong on the other datasets. This result is important for practitioners who wish to adapt CLIP for their tasks but cannot afford to label an additional held-out validation set for hyperparameter tuning.

4.2.2 Zero-Shot Classification

Here we attempt to further improve the already impressive zero-shot performance of vision-language models such as CLIP. To achieve this, we take the most confident predicted pseudolabels from CLIP for the classes of interest as the training data for SVL-Adapter, thus making it compatible for zero-shot transfer. Essentially, we keep the adaptation pipeline the same and just replace the ground truth labels normally used for low-shot adaptation with pseudolabels. Utilizing SVL-Adapter* enables us to keep the task truly zero-shot as we do not use any labeled data for hyperparameter tuning. The baseline we compare against is the standard zero-shot version of CLIP which also uses a ResNet50 backbone. The results illustrated denoted as 'Zero-shot SVL-Adapter*' in Figs. 3 and 4 show significant improvements across the majority of the datasets when compared with the standard zero-shot CLIP baseline. Specifically, for the "Challenging" and the "Standard" datasets we record an average of 8% and 5% improvement in Top-1 accuracy. Thus, we can infer that vision-language models, combined with adapted self-supervised features, significantly improve zero-shot classification performance.

4.3 Limitations

While we report large improvements for zero- and low-shot classification compared to current state-of-the-art methods on challenging datasets, there are some notable cases where we do not perform as well, *e.g*. StanfordCars and FGVCAircraft in Fig. 4. It is known that

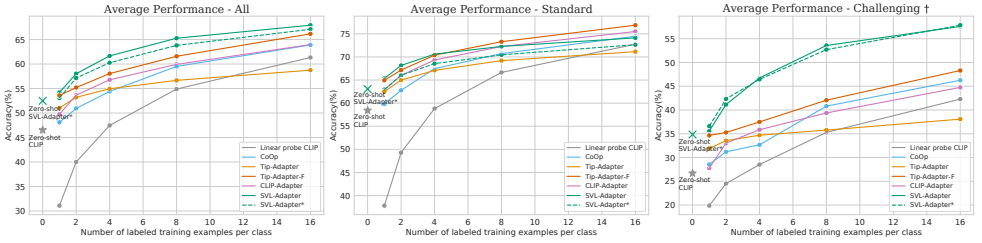

Figure 3: (Left) Average zero- and low-shot test Top-1 accuracy across all 16 datasets. (Middle) Results for the ten "Standard" datasets commonly used in existing work. Here, our SVL-Adapter is competitive with current SoTA methods. (Right) On the "Challenging" datasets, which differ more from CLIP training data, SVL-Adapter significantly outperforms existing methods. SVL-Adapter* refers to our approach with automatic blending weight selection as described in Sec. 3.2.1. The per-dataset results are illustrated in Fig. 4.

fine-grained datasets such as these pose a challenge to current SSL methods [13]. However, our approach is agnostic to the underlying SSL method used, and will benefit from newer, more effective representations, from this highly active research area. We also need to train a self-supervised representation on each dataset of interest. Fortunately, this is not a very time consuming operation due to the moderate size of most datasets of interest, *i.e*. practitioners are more likely to work with 'small' datasets containing thousands of images, not millions.

5 Conclusion

We presented SVL-Adapter, a self-supervised vision-language adapter for zero- and low-shot image classification. We showed that large-scale web-trained models such as CLIP fail to effectively generalize to challenging visual classification tasks that do not come from the distribution of images commonly found online. Furthermore, recent methods for adapting these models also fail to significantly improve performance. By combining the complementary strengths of self-supervised learning and vision-language pretraining, our approach results in large improvements in low-shot classification accuracy on challenging visual classification tasks without requiring any additional supervision at training time. We also showed that SVL-Adapter is applicable in the zero-shot learning setting, where it improves over the conventional baseline despite not requiring any additional supervision.

Acknowledgements: This work was in part supported by the Turing 2.0 'Enabling Advanced Autonomy' project funded by the EPSRC and the Alan Turing Institute. The research is also supported by the Biome Health Project funded by WWF-UK.

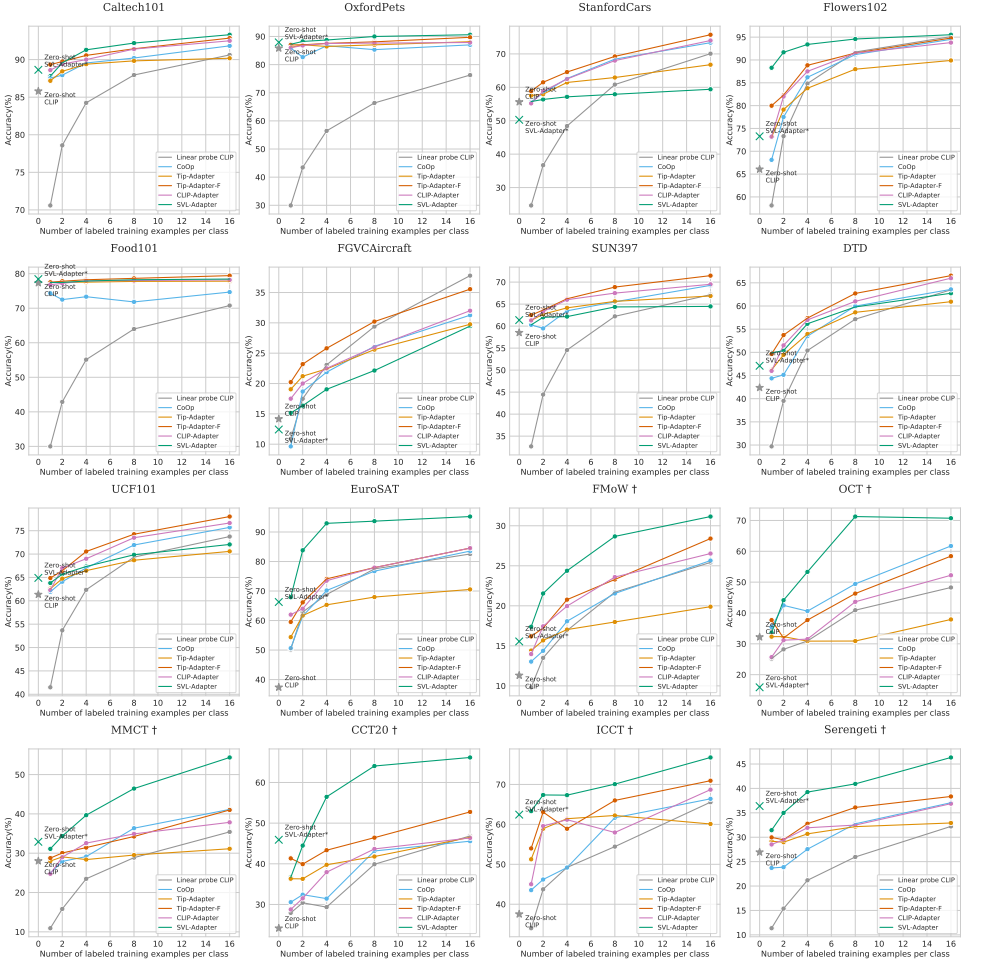

Figure 4: Per-dataset, zero- and low-shot, test Top-1 accuracy results across 16 different datasets. In each case, we report the average of three runs. Datasets marked with † are added by us, and pose a significantly greater challenge to existing methods. A summary of the results is presented in Fig. 3.

References

[1] Island Conservation Camera Traps. http://lila.science/datasets/island-conservation-camera-traps.

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *ICCV*, 2021.

[3] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *ICCV*, 2021.

[4] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Patricia MacWilliams, S Sara Mahdavi, Ellery Wulczyn, et al. Robust and efficient medical imaging with self-supervision. *arXiv preprint arXiv:2205.09723*, 2022.

[5] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018.

[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[8] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

[11] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018.

[12] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014.

[13] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. When does contrastive visual representation learning work? In *CVPR*, pages 14755–14764, 2022.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[17] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *ICCV*, 2021.

[18] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshops*, 2004.

[19] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[20] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

[22] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

[26] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226, 2019.

[27] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[29] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *CVPR*, 2022.

[30] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

[31] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[32] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 2020.

[33] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), 2018.

[34] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[35] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[36] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, 2022.

[37] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[38] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, 2008.

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[40] Omiros Pantazis, Gabriel J. Brostow, Kate E. Jones, and Oisin Mac Aodha. Focus on the positives: Self-supervised learning for biodiversity monitoring. In *ICCV*, 2021.

[41] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012.

[42] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[43] 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 *ICML*, 2021.

[44] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[45] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018.

[46] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850. PMLR, 2016.

[47] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[48] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[50] Asa Cooper Stickland and Iain Murray. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. In *ICML*, 2019.

[51] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1), 2015.

[52] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision*, pages 266–282. Springer, 2020.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017.

[54] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016.

[55] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3): 1–34, 2020.

[56] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018.

[57] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[58] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[59] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[60] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

[61] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *ECCV*, 2022.

[62] Xueting Zhang, Debin Meng, Henry Gouk, and Timothy M Hospedales. Shallow bayesian meta learning for real-world few-shot recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 651–660, 2021.

[63] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *ICML*, 2021.

[64] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022.

[65] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022.