Thesis framework

Barış Deniz Sağlam
Informatics Institute
Middle East Technical University
Ankara, Turkey
e155841@metu.edu.tr

***Abstract*—**

***Index Terms*—unsupervised domain adaptation, transfer learning, zero-shot learning**

I. Framework

A. Background

In the last decade, deep learning has achieved great successes on various computer vision tasks thanks to large-scale labelled datasets. However, it is expensive and time-consuming to label data for each task and domain. This sparked the motivation for self-supervised learning algorithms, and transfer learning.

Transfer learning is to use a pre-trained model for a different task or domain than it is originally trained for, assuming that the model's knowledge is transferrable. For instance, an image classification model trained on ImageNet [1] can be transferred to a different domain such as ImageNet-Sketch [2] or to a different task such as object detection. The performance of transfer learning is determined by generalizability of learned representations across different domains, where domain is defined as a specific distribution of sample space [3]. A direct transfer usually does not perform well due to dataset shift. Two common techniques of transfer learning exist to mitigate this problem; fine-tuning and domain adaptation. When there are labels for the target domain, the model can be fine-tuned on it with no or limited modification on the model's architecture. When there is no label for the target domain, then the problem falls under unsupervised domain adaptation. Domain Adaptation is a particular case of transfer learning that leverages labeled data in one or more related source domains, to learn a model for unseen or unlabeled data in a target domain for the same task [4]. One of the most prevalent domain adaptation technique is to minimize discrepancy between source and target domains so that the model representations become domain-invariant while preserving discriminability for the task.

Self-supervised learning exploits the structure and patterns in the data itself to learn useful representations for downstream tasks [5]. The basic idea is to construct auxiliary tasks and loss functions that do not need any labels and to train the model to perform well on these auxiliary tasks [5]. Constrastive learning, masked prediction, auto-regressive prediction, reconstruction are common tasks used in self-supervised learning. The representations from a pre-trained self-supervised model can be used to train another model where a small-scale labelled dataset exist.

While domain adaptation overcomes domain shift problem by adapting a model to a target domain, few-shot learning aims creating a model that produces generalizable representations for all relevant tasks at once. Low-shot learning stands for building generalizable representations, usually with large-scale data and using them with zero or few labelled samples given for a downstream task. Unlike domain adaptation, low-shot learning can be applied when the target task is different. The success of large language models on few shot learning and domain generalization [6], [7] inspired recent breakthroughs in computer vision models that are more generalizable such as CLIP [8] and ALIGN [9]. These type of models are composed of image encoder and text encoder. They are trained on web-scale dataset consisting of image-text pairs with contrastive loss. After pre-training, natural language is used to reference learned visual concepts enabling zero-shot transfer of the model to downstream tasks. They achieve comparable performance on many vision benchmarks to fully supervised alternatives. It's been also shown that their performances improve further with few-shot learning. Zero-shot image generation model DALL-E [10], which is built upon CLIP [8], has been shown to understand abstract, novel, and imaginary concepts.

B. Problem statement

While self-supervised vision-language models (SSVLM) possess decent generalization capabilities, two main limitations of these models are observed: 1. They perform poorly on very specific domains and tasks, similar to non-expert human labellers. 2. They require finding the best performing prompt for a task, also known as prompt-engineering.

C. Related works

Deep Unsupervised Domain Adaptation (DUDA) can be divided into four categories; Discrepancy-based methods, Adversarial discriminative methods, Adversarial generative methods, Self-supervision-based methods [5].

Discrepancy-based methods aims to minimize the discrepancy among different domains. The idea is to make the model produce task-specific but domain-invariant features. There are various ways to explicity measure discrepancy such as multiple kernel variant of maximum mean discrepancies, second-order statistics of features, and etc. [5]. One of the challenges for these methods is to maintain the task-specific decision boundaries between classes [11] while making domains indiscrimanable. [11] proposes a minimax problem in which two classifiers are trained to maximize the discrepancy on the target domain, and then generate features that minimize this discrepancy. They measure domain discrepancy as L1 distance of class probability distributions predicted by each classifier.

Adversarial Discrimantive methods achieves the same by training a domain discriminator such that the generated features become indiscrimanable by it. Sun et al. [12] accomplishes aligning source and target domain by jointly training the model on primary classification task on source dataset and on auxiliary self-supervised tasks on both source and target datasets. These self-supervised tasks includes rotation prediction, flip prediction, and patch location prediction for images.

D. Proposed method

How the representations from SSVLM can be leveraged to adapt an existing task-specific model to a different domain is an open research question. In this thesis, new techniques will be explored to leverage capabilities of such large multi-modal pretrained models for domain adaptation. We hypothesize that the representations produced by these models contain world knowledge which can be utilized for adapting task-specific models. For instance, an image-segmentation model trained on day-time images can be adapted to night-time images by using the embeddings from SSLVM representing this domain shift. Human generated or learned prompts or combination of them can be used for this manner.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[2] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 506–10 518.

[3] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 3, pp. 766–785, 2019.

[4] G. Csurka, "Domain Adaptation for Visual Applications: A Comprehensive Survey," feb 2017. [Online]. Available: https://arxiv.org/abs/1702.05374v2

[5] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez, A. L. Sangiovanni-Vincentelli, S. A. Seshia, and K. Keutzer, "A Review of Single-Source Deep Unsupervised Visual Domain Adaptation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 473–493, feb 2022.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[9] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916.

[10] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[11] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3723–3732.

[12] Y. Sun, E. Tzeng, T. Darrell, and A. A. Efros, "Unsupervised domain adaptation through self-supervision," *arXiv preprint arXiv:1909.11825*, 2019.