PREDICTING THE GENERALIZATION GAP IN DEEP MODELS USING ANCHORING

Vivek Narayanaswamy[1], *Rushil Anirudh*[2], *Irene Kim*[2], *Yamen Mubarka*[2]
Andreas Spanias[1], *Jayaraman J. Thiagarajan*[2]

[1] Arizona State University, [2] Lawrence Livermore National Laboratory

ABSTRACT

We address the problem of predicting the generalization gap of deep neural networks under large, natural, and synthetic distribution shifts between source and target domains. This is crucial in understanding how models behave in uncontrollable 'in-the-wild' scenarios, but existing techniques fail when target domain becomes very different from the source. Accurately capturing the relationship and distance between the source and target domains is critical for a reliable post-hoc estimation of generalization. In this paper, we propose a novel strategy for directly predicting accuracy on unseen target data with the help of anchoring and pre-text encoding in predictive models. Anchoring has been shown previously to perform effectively in characterizing domain shifts, which we exploit for predicting the generalization gap. Our experiments on the PACS dataset along with synthetic ablations indicate that our approach produces well calibrated accuracy estimates outperforming existing baselines.

***Index Terms*—** Predicting generalization, Deep Neural Networks, Uncertainty Estimation, Calibrated models

1. INTRODUCTION

With tremendous success exhibited by AI methods [1, 2, 3], off-the shelf black-box neural networks are being increasingly deployed to guide decision making even in critical applications such as healthcare. However, these models have been shown to function reliably only when the test distribution overlaps significantly with the training distribution [4, 5] which is seldom the case, making it difficult to safely rely on such model predictions on new test domains. An important step towards promoting the adoption of these models in practice is not only to ensure that they behave predictably on regimes where the training data provides meaningful evidence, but also to provide an accurate estimation of expected generalization accuracy when utilized on completely unlabeled, 'in-the-wild' target distributions.

This work was supported in part by the ASU SenSIP Center, Arizona State University. This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC.

Estimating generalization accuracy on target distributions is an important topic of research [6, 7, 8]. For instance, Jiang *et al*. [6], showed that distances between the training distribution and model decision boundaries can be a strong indicator to predict generalization error for unseen examples. However, this approach makes the assumption that the train and test distributions are similar, and hence the predicted generalization gap need not be accurate under distribution shifts. Recently developed strategies include training a post-hoc predictor based on metrics used commonly in domain adaptation settings [9, 10] that quantify differences in data distributions to predict the generalization performance. For e.g., Deng *et al.* [8] use the Frechét distances between the training and the target set to fit regression models to estimate accuracy. The authors demonstrate the existence of linear relationships between accuracy gap and the distribution distances. On similar lines, Guillory *et al.* [7] utilize the difference of confidence (DoC) metric between the train and the target sets in order to predict change in accuracies on unseen datasets and show that DoC significantly outperforms other distribution difference metrics over a variety of natural and synthetic distribution shifts. Despite the simplicity of these methods, we find that there still exists a significantly wide accuracy gap when applied on real-world domain shifts as the regressor is not guaranteed to be calibrated well enough to reflect the uncertainties between the train and target domains.

Uncertainty estimation in machine learning (ML) [11, 12] is a powerful tool to characterize model behaviour under distribution shifts and identify regimes of improper sampling in data. Recently, Anirudh & Thiagarajan [13] introduced a novel uncertainty estimation technique, $\Delta-$UQ, based on *anchoring* — where the input is transformed into a tuple consisting of an anchor sample (random sample drawn from a prior) and a pretext encoding of the input with the anchor to train predictive models. During inference, an anchor marginalization strategy is used to obtain the prediction for each sample over multiple randomly chosen anchors. When a suitable pretext encoding is chosen, $\Delta-$UQ has been found to effectively distinguish between in- and out-of-distribution samples, and to produce meaningful uncertainties using a single model, in contrast to stochastic approaches such as deep ensembles [11]. Interestingly, we observe that anchoring a predictive model with the pre-text task as done in [13], implicitly

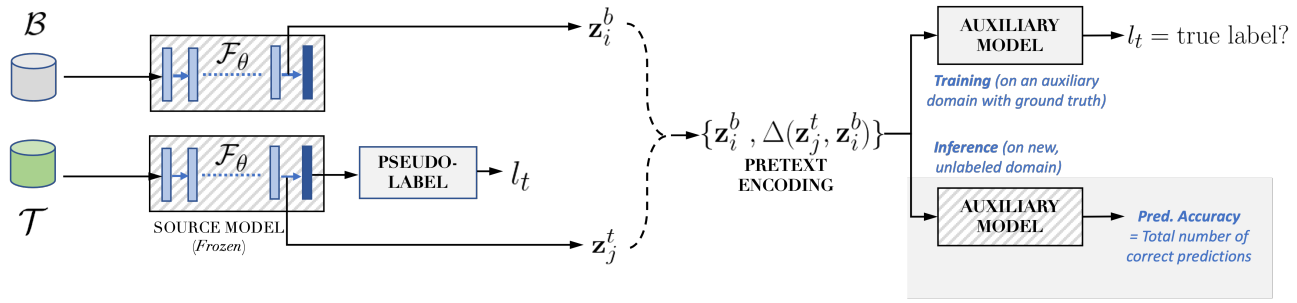

Fig. 1. Overview of our proposed approach to predict accuracy on unseen target distributions. Utilizing the intermediate features $\mathbf{z}_i^b$, $\mathbf{z}_j^t$ extracted from samples of the source $\mathcal{B}$ and target $\mathcal{T}$ distributions respectively from a pre-trained classifier $\mathcal{F}$, we construct pre-text encodings of the form $\{\mathbf{z}_i^b, \Delta(\mathbf{z}_j^t, \mathbf{z}_i^b)\}$ to train auxiliary models that can be used to predict generalization accuracy. This encoding strategy effectively captures the important differences between the source and target distributions which can be leveraged to estimate generalization gaps.

defines a metric between two distributions – the anchor distribution and the source distribution on which the model is trained. As a result, we hypothesize that an such a strategy can in fact provide meaningful information regarding shifts between different data distributions and can be leveraged in order to directly predict the generalization accuracy on unseen target distributions – without relying on summary metrics.

Proposed Work: In this paper, we propose a novel strategy for directly predicting accuracy of deep models on previously unseen, unlabeled target distributions. Specifically, we utilize the $\Delta-$encoding scheme from [13] to train an auxiliary model, which is comprised of two main components – (a) a decoder model that tries to undo the pretext encoding to recover the representation of a target sample obtained from the pre-trained model, implicitly capturing the relationship between the source and target datasets; and (b) a binary classifier that acts on the residual between the decoded sample and the target representation, to predict if the source network correctly classified this sample or not. By training this entire process end-to-end (while keeping the source model frozen), our experiments on synthetic and the multi-domain PACS dataset [14] show that, in addition to providing well-calibrated target accuracy estimates, our proposed approach also outperforms existing baselines.

2. PROPOSED APPROACH

In this section, we describe our approach for predicting accuracy of a trained model on unseen target distributions. An overview of our approach is illustrated in Figure 1.

2.1. Preliminaries and Notations

Let $\mathcal{F}_\theta$ denote a multi-class classifier (source model) parameterized by θ that takes an image $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$ as input to predict the output label $\hat{\mathrm{y}} \in \mathcal{Y} := [1, \cdots, K]$. Here, $\mathbf{x}$ is a C channel image of height H and width W and K represents the total number of classes. Let $\mathcal{D}_\Phi$ denote the decoder network with parameters Φ and $\mathcal{G}_\Psi$ corresponds to

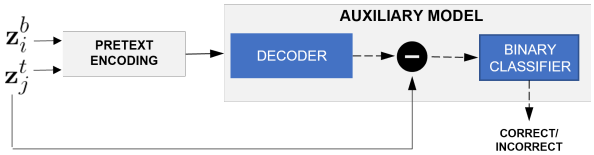

Fig. 2. The auxiliary model block consists of two components (i) a decoder that tries to undo the pre-text encoding to recover a representation of the target sample obtained from the pre-trained model capturing the relationships between the source and targets (ii) a binary classifier to predict whether the target sample has been correctly classified or not by $\mathcal{F}$.

the binary classifier with parameters Ψ. In this paper, we assume that the $\mathcal{F}_\theta$ is pre-trained on the source distribution $\mathcal{B}$ consisting of N samples $\{x_i^b,\ y_i^b\ |i = 1 \ldots N\} \sim \mathcal{B}$. Let $\mathcal{T}$ denote the target distribution consisting of M samples $\{x_j^t,\ y_j^t\ |j = 1 \ldots M\} \sim \mathcal{T}$. The set of labels present in both the base and target distributions are assumed to be identical. Additionally, we define an additional calibration dataset, $\mathcal{T}_c \neq \mathcal{B}$, which is used to train $\mathcal{D}_\Phi$ and $\mathcal{G}_\Psi$. At inference time, samples from the target domain are passed through the trained models to finally obtain an estimate of accuracy of $\mathcal{F}_\theta$. Note, in all of this the pre-trained model $\mathcal{F}_\theta$ is kept frozen.

2.2. Pretext Encoding Scheme

We define pretext encoding as a function $\Delta(Q, R)$, where Q denotes a query sample and R is an anchor with the same dimensions as Q, randomly drawn from a distribution $P(R)$. We reformulate the problem of training the auxiliary models using tuples constructed using anchors and the pre-text encodings $\{R, \Delta(Q, R)\}$. The tuple is realized by concatenating R and $\Delta(Q, R)$ (in the channel dimension). Following [13], we consider the pretext encoding function as $\Delta(Q, R) = Q - R$.

In our approach, we construct the tuples using latent features computed for the base and target distributions using $\mathcal{F}_\theta$, *i.e.*, $R = \mathbf{z}_i^b$ and $Q = \mathbf{z}_j^t$. For training the models, we randomly choose a single anchor R for every input sample Q

in a batch to ensure that each input sample is combined with different random anchors as the training progresses.

2.3. Training the Auxiliary Models

2.3.1. Intuition

The auxiliary models which include the decoder $\mathcal{D}_\Phi(.)$ and the binary classifier $\mathcal{G}_\Psi(.)$ are critical to evaluate generalization performance on unseen target data (Figure 2). The tuples extracted from the pretext encoding stage are used to jointly train $\mathcal{D}_\Phi$ and $\mathcal{G}_\Psi$. The key intuition behind our methodology is based on the 'explicit' approach of [13]. The task of training a predictive model directly using the tuples can be decomposed into two stages (i) Using a decoder to recover Q from the $\Delta-$encoding and (ii) Using a predictor that utilizes the estimated output from the decoder to make predictions. Since the decoder is stochastic (input Q can be associated with any random R during training), the representation $\mathbf{d}$ from the decoder can be interpreted as a 'reconstruction' of the input Q by averaging over different choices of R. When $\mathbf{d}$ exactly matches Q there exists no uncertainty in the input. On the other hand, if there is a mismatch, it denotes the epistemic uncertainty in recovering Q over the distribution of anchors.

By selecting anchors from the base distribution $\mathcal{B}$ and query samples from the target distribution $\mathcal{T}$, wherein the two datasets can have non-overlapping manifolds, recovering the input samples under different anchor choices can lead to larger discrepancies which can be exploited to detect domain shifts.

2.3.2. Decoder

In our work, the decoder $\mathcal{D}_\Phi$ operates on the tuple to produce an intermediate representation $\mathbf{d}$ with the same dimensions as that of $\mathbf{z}_j^t$. We then compute residuals $|\mathbf{d} - \mathbf{z}_j^t|$ (indicative of the uncertainties) to train the binary classifier $\mathcal{G}_\Psi$.

2.3.3. Binary Classifier

In order to train the binary classifier, which can be eventually used to estimate the generalization performance of the underlying pre-trained classifier, we obtain labels from $\mathcal{F}_\theta$ indicative of whether the data sample x_j^t has been correctly classified or not. We first determine the true class likelihoods $P(y_j^t|x_j^t)$ and perform the following pseudo-labeling strategy to prepare the labels for the binary classifier.

$$l_t = \begin{cases} 1, & \text{if } P(y_j^t|x_j^t) >= \tau_2, \\ 0, & \text{if } P(y_j^t|x_j^t) <= \tau_1, \\ \text{ignore}, & \text{otherwise.} \end{cases} \quad (1)$$

Here, τ_1, τ_2 are user-specified thresholds. The binary classifier which then estimates the likelihoods $\hat{p}_j^t$ of the target data samples being correctly classified by $\mathcal{F}_\theta$ is jointly trained with the decoder using the Binary Cross Entropy loss function.

2.4. Predicting Generalization

During inference with the pre-trained classifier $\mathcal{F}$ and auxiliary models $\mathcal{D}$ and $\mathcal{G}$, we estimate the mean likelihood of the unseen target data sample being correctly classified by analyzing the consistency in the prediction with respect to K different anchors drawn from the base distribution $\mathcal{B}$. Therefore for a given target sample, we marginalize the impact of the anchors and compute the prediction as follows:

$$\overline{\hat{p}_j^t} = \frac{1}{K} \sum_K \mathcal{G}(|\mathcal{D}(\mathbf{z}_k^b, \Delta(\mathbf{z}_j^t, \mathbf{z}_k^b)) - \mathbf{z}_j^t|), \quad (2)$$

We finally estimate the overall generalization accuracy on the entire unseen target dataset, by aggregating the individual mean predictions as follows:

$$\text{Acc}(x_j^t | j = 1 \dots M) = \frac{1}{M} \sum_j \mathbb{I}(\overline{\hat{p}_j^t} > \gamma), \quad (3)$$

where γ is the threshold of detection.

3. EXPERIMENTS

Datasets. We evaluate our proposed approach for predicting generalization performance on unseen target domains and synthetic variations in the Photo-Art-Cartoon-Sketch (PACS) [14] dataset. The dataset contains 9991 images across the different domains. All domains share the same label space where each image is associated with one of the following 7 categories namely person, house, horse, guitar, giraffe, elephant and dogs. In all our experiments we utilize the images from the photo domain as the base distribution $\mathcal{B}$ and the images from the sketch domain as the auxiliary domain $\mathcal{T}_c$ to train the auxiliary models. The cartoon and the art domains are considered only for predicting the generalization performance. Further, we perform synthetic augmentations onto the photo, cartoon, and the art domains to create datasets with different distribution shifts such horizontal flips, brightness and hue, random rotations and Gaussian blur of low severity. In total, we evaluate our model, and baselines on a total of 14 test domains containing both natural and synthetic distribution shifts, to predict the generalization gap.

Setup. We use a Resnet-18 model [15] pre-trained on Imagenet [16] as the classifier $\mathcal{F}$ and extract the latent features from the last residual block and perform average pooling to obtain $\mathbf{z}_i$. The decoder $\mathcal{D}$ and the binary classifier $\mathcal{G}$ are fully connected neural networks with 5 hidden layers of 512 neurons and 2 hidden layers of 512 neurons respectively. All models are trained using the ADAM optimizer with a learning rate of $3e^{-4}$ for training. We use predictions obtained with the source model $\mathcal{F}$ as pseudo-labels to train the auxiliary model, and use thresholds $\tau_1 = 0.25$ and $\tau_2 = 0.65$ in an attempt to better guide the auxiliary model training process, as explained in (1). In order to obtain the final prediction from

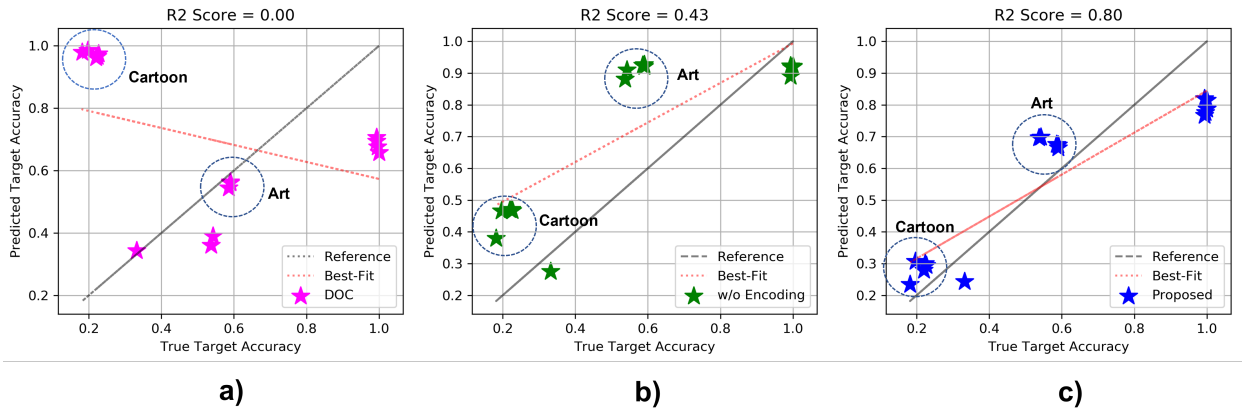

Fig. 3. Comparison of the generalization performance of the proposed approach on different target domains and synthetic shifts over existing baselines. We find that our approach reliably estimates the generalization accuracy with a strong linear relation between the true and predicted target accuracies against the baseline approaches.

Table 1. Means and Standard Deviation of the accuracy gaps over different target distributions.

Methods	**|True - Predicted Accuracy| Mean $\pm$ Std**
DOC	0.3650 ± 0.2903
w/o Δ-Encoding	0.211 ± 0.113
Proposed	$\mathbf{0.12 \pm 0.06}$

the binary classifier, we use a threshold, $\gamma = 0.6$, which we found to be sufficiently conservative in all our experiments.

Baselines. (i) *Difference of Confidence (DoC)* [7]. DoC is a recently proposed approach for predicting generalization gaps on unseen distributions that has been shown to outperform existing baselines that operate on conventional distributional distances such as Maximum Mean Discrepancy (MMD) [9] and Frećhet distances. Following [7], we fit a linear regressor to the DoC scores to predict change in accuracies by constructing random subsets of the source domain (Photo) and the calibration domain (Sketch) with different synthetic augmentations and evaluate the regressor on all unseen distributions. (ii) *Auxiliary model training w/o pretext encoding and anchoring*. As an ablation, we consider a baseline that directly takes the intermediate feature representation for the auxiliary domain, and predicts whether or not the main ResNet model, $\mathcal{F}$, got the prediction correct. We find that this model tends to overfit easily, so we simply use the binary classifier portion of our main model as the auxiliary model.

Results and Discussion In Fig. 3, we show true and predicted accuracies for 16 test domains (14 unseen, and test splits of photo and sketch used as training and auxiliary domains respectively) across the three techniques considered here. We find that our method significantly outperforms the others significantly, with a strong linear relationship between the true and predicted accuracies, resulting in an R2 score of 0.8. This is in comparison to 0.43 for the ablation model without pretext encoding. Interestingly, we also note that the ablation model also suffers from mostly being over confident in its predictions (most of the values are above the diagonal), whereas the anchoring and pretext encoding is less biased in its estimates on average. The DoC [7] predictions shown in Figure 3 (left most) suffer primarily due to its failure on domains that are significantly shifted from the source or auxiliary domains. Corroborating Figure 3, Table 1, provides the mean and standard deviations of the accuracy gaps over different target distributions. It can be observed that our proposed approach outperforms the baselines in producing significantly low accuracy gaps.

4. CONCLUSION

In this paper, we proposed a novel strategy for directly predicting accuracy of a deep neural network on unseen target distributions using the idea of anchoring and pre-text encoding in predictive modeling. We find anchoring to be an effective strategy to train post-hoc accuracy estimators, that are able to better model the distribution shifts as a function of the generalization gap, unlike existing methods that rely on difference metrics. In particular, we adopted the $\Delta-$encoding scheme from [13] to train auxiliary models containing (a) decoder that tries to undo the pretext encoding to recover the representation of a target sample obtained from the pre-trained model by capturing the relationship between the source and target datasets; and (b) a binary classifier that acts on the residual between the decoded sample and the target representation, to predict if the source network correctly classified this sample or not. Through extensive experiments on the PACS dataset along with synthetic variations, we found that our approach provided reliable accuracy estimates outperforming the existing baseline and ablations.

5. REFERENCES

[1] 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," in *International Conference on Learning Representations*, 2021.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*, 2019, pp. 4171–4186.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[4] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 2010.

[5] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar, "Do imagenet classifiers generalize to imagenet?," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5389–5400.

[6] Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio, "Predicting the generalization gap in deep networks with margin distributions," in *International Conference on Learning Representations*, 2019.

[7] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt, "Predicting with confidence on unseen distributions," *arXiv preprint arXiv:2107.03315*, 2021.

[8] Weijian Deng and Liang Zheng, "Are labels always necessary for classifier accuracy evaluation?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15069–15078.

[9] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.

[10] Xavier Glorot, Antoine Bordes, and Yoshua Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in *ICML*, 2011.

[11] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[12] Jayaraman J Thiagarajan, Bindya Venkatesh, Prasanna Sattigeri, and Peer-Timo Bremer, "Building calibrated deep models via uncertainty matching with auxiliary interval predictors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 6005–6012.

[13] Rushil Anirudh and Jayaraman J. Thiagarajan, "Δ-UQ: Accurate uncertainty quantification via anchor marginalization," *arXiv preprint arxiv:2110.02197*, 2021.

[14] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5542–5550.

[15] 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*, 2016, pp. 770–778.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.