Robust fine-tuning of zero-shot models

Introduction

다양한 데이터 분포(i.e. Distribution shifts)에서 모두 잘 적응할 수 있는 모델이 필요

→ 모델에 Robustness가 필요하다. 하지만 Robustness를 가지게 하는 일은 아주 어려움.

하지만 CLIP, ALIGN, BASIC과 같은 큰 사전 학습 모델이 큰 Robustness함을 가짐이 증명됨.

→ 성공 요인은 아마 heterogeneous dataset을 통한 학습으로 보인다.

→ **또한, zero-shot 상황에서 가장 큰 Robustness improvement를 보였다. (target data distribution에 대한 fine-tuning 없이 예측함)**

보통 target data distribution에 대한 fine-tuning은 성능을 올리는데..

→ 이는 모델의 Robustness를 희생하여 나타난 결과라는 것, 따라서 fine-tuned 모델의 성능은 zero-shot 모델에 비해 다양한 distribution shift에 대해서 robustness 성능은 오히려 낮았다.

⇒ Can zero-shot models be fine-tuned without reducing accuracy under distribution shift?

How to Robustly Fine-tuning pre-trained models..

Analysis Results.

1. The robustness of fine-tuned models varies substantially under even small changes in hyperparameters, but the best hyperparameters cannot be inferred from accuracy on the target distribution alone
2. Second, more aggressive fine-tuning (e.g.,using a larger learning rate) yields larger accuracy improvements on the target distribution, but can also reduce accuracy
   under distribution shift by a large amount

   → 공격적인 fine-tuning을 할수록, target에 대한 정확도는 높아졌지만 그만큼 Robust함이 낮아졌다.

Method

Method has two steps:

1. Fine-tune zero-shot model on target distribution
2. Combine the original zero-shot and fine-tuned model by linearly interpolating between their weights → *weight-space ensembling*

- Weight-space ensembles for fine-tuning (WiSE-FT)

! Step1 ..

g: image encoder used by CLIP, g(x, V_enc) is an input image and parameter of g

$\mathbf{V}_{\text{enc}}$ are the parameters of the encoder g. Standard fine-tuning considers the model $f(x, \theta) = g\left(x, \mathbf{V}_{\text{enc}}\right)^{\top} \mathbf{W}_{\text{classifier}}$ where $\mathbf{W}_{\text{classifier}} \in \mathbb{R}^{d \times k}$ is the classification head and $\theta = [\mathbf{V}_{\text{enc}}, \mathbf{W}_{\text{classifier}}]$ are the parameters of f. We then solve $\arg\min_{\theta} \left\{ \sum_{(x_i, y_i) \in \mathcal{S}_{\text{ref}}^{\text{tr}}} \ell(f(x_i, \theta),\ y_i) + \lambda R(\theta) \right\}$ where ℓ is the cross-entropy loss and R is a regularization term (e.g.,

! Step 2..

alpha in [0,1]..

zero-shot model theta_0, standard fine-tuned model theta_1

1. End-to-end : all value of theta are modified
2. Fine-tuning only a linear classifier. (encoder is fixed)

$$\text{wse}(x, \alpha) = f(x, (1 - \alpha) \cdot \theta_0 + \alpha \cdot \theta_1)\ ,$$

element-wise weighted average of the zero-shto and fine-tuned model's paramters

Linear Classifier만 학습할 때는, Zero-shot에는 따로 Classifier가 없기 때문에 따로 Ensembling 하는 게 없음.

Experimental

Experimental Setup

실험의 목적은 zero-shot과 fine-tuned 모델, WiSE-FT 모델의 성능을 비교하기 위함임

D_ref : target data distribution for fine-tuning (ID, 단, 학습은 하는 거니까 ID라는 표현이 정확하진 않음.)

D_shift : D_ref's shifted data distribution. (OOD)

→ 두 데이터셋 모두 각각의 training set, test set을 가지고 있음.

Acc_ref, ACC_shift: classification accuracy on the reference and shifted test sets

on k-way image classification where the outputs of f are k-dimensional vectors of non-normalized class scores.

- About Distibution Shifts

1. synthetic: adversarial example or artificial change in image contrast, brightness..
2. natural: changes in data distributions arise through naturally occurring variations in lighting, geographic location, crowdsourcing process, image styles

본 논문에서는 어떠한 adversary가 없는 natural distribution에 집중.

Based on ImageNet & ImageNet-V2 & ImageNet-R & ImageNet sketch .. etc

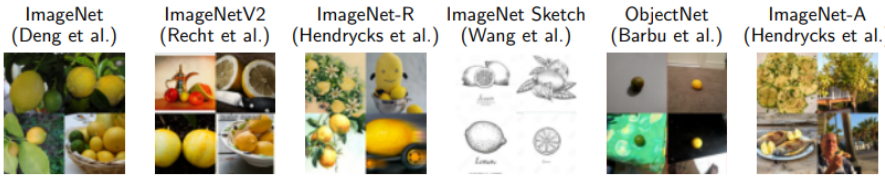

Figure 2. Samples of the class *lemon*, from the reference distribution ImageNet [17] and the derived distribution shifts considered in our main experiments: ImageNet-V2 [81], ImageNet-R [35], ImageNet Sketch [98], ObjectNet [4], and ImageNet-A [38].

- Effective robustness and scatter plots

Robustness를 측정하기 위해서 *effective robustness framework 사용 :*

x축에 대해 ref(ImageNet)에 대한 acc, y축은 shift에 대한 acc를 scatter plots으로 표시해줌.

- Zero-shot model and CLIP

인터넷의 Image-caption pair를 통해 학습된 CLIP

→ Image-encoder g, text encoder h를 통해 서로의 상관성을 높이는 방향으로 학습.

Zero-shot k-way classification **by matching with potential captions**

→ Zero-shot에서 Class category를 모르는 데 어떻게 분류를 할 수 있을까?

→ Text encoder를 통해 획득할 수 있는 caption s_i에 대한 예측 단어에 대한 weight를 g(x)와 곱해서 최종 아웃풋 산출. (f(x) = g(x)^T * W_zero-shot where W_zero-shot is R^d*k with columns h(s_j))

Used model.. ViT-L/14@336px..

Results

2/19