Free Lunch for Few-shot learning : distribution calibraiton

Abstract

- 从有限数量的样本中进学习非常具有挑战性, 因为模型对于来自仅由少量样本组成的biased distribution的数据非常容易overfitting
- 本文提出了一种通过迁移来自有充足样本类别的统计量来calibrate这些小样本的分布, 然后就可以从calibrated的样本分布中采样更多的数据来扩充输入数据集
- 本文假设在feature representation中的每一个dimension都服从Gaussian distribution, 因此分布的mean和variance能够从相似的类别中borrow过来, 因为相似类别的统计量因为有足够的数据, 能够更好的被评估
- 我们的方法可以直接在现有的模型上结合, 无需额外的参数. 实验表明一个简单的logistic regression结合本文方法后在两个数据集上超过了SOTA

Introduction

- 从有限数据进行学习变成重要方向, 主要有几个类别, Meta-Learning, Synthesize data, using pseudo label, etc
- 大多数关注在更强大的模型, 但仅有少数注意力放在数据本身. 自然地, 随着数据的数量慢慢增加, 其真实的分布也能够更加准确的uncovered
- 在训练过程中, 若仅有少量训练数据, 则模型会很容易由于minimizing training loss而倾向于overfitting. 这种在biased distribution上少样本训练数据的学习会损害模型的泛化性能, 因此测试的case都是采样自ground truth的分布中, 然而模型学习到的是一个biased distribution. 综上, 本文考虑对biased distribution做calibration, 使得其能够更加准确的近似与真实的数据

  图示训练在biased distribution的问题

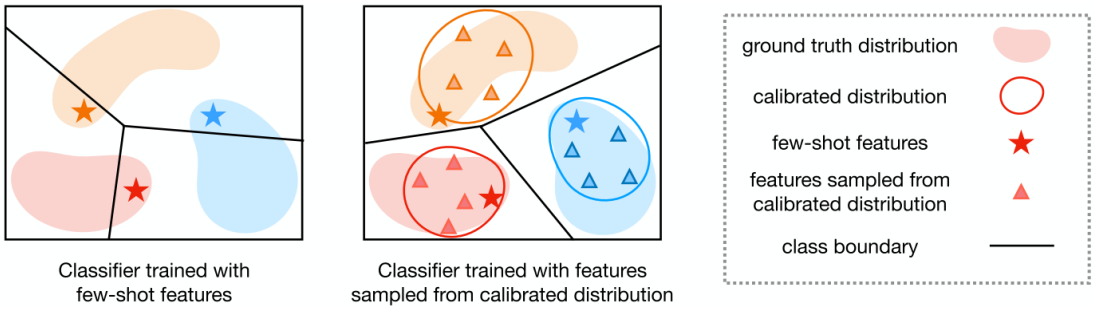

Figure 1: Training a classifier from few-shot features makes the classifier overfit to the few examples (Left). Classifier trained with features sampled from calibrated distribution has better generalization ability (Right).

- 一个在calibrated distribution中采样数据的模型能够更好的在准确分布中泛化. 在本文中, 相比在原数据中进行calibration, 作者在feature space中进行calibration, 因为在feature space中, 数据会有更加小的dimension, 并且能够更容易的calibrating

- 本文中假设在feature space中每个dimension都是服从高斯分布的, 同时观察到相似的类别在feature space中通常拥有相似的mean和variance, **因此这些统计量能够从相似类别的数据中transferred过来**, 同时在样本充足的相似类中估计这些统计量也更加的准确. 如图所示

Table 1: The class mean similarity ("mean sim") and class variance similarity ("var sim") between Arctic fox and different classes.

	Arctic fox	
	mean sim	var sim
white wolf	97%	97%
malamute	85%	78%
lion	81%	70%
meerkat	78%	70%
jellyfish	46%	26%
orange	40%	19%
beer bottle	34%	11%

Related Works

- 小样本学习主要有三条路线. 1) meta-learning的方法 2) generated data 3) data augmentation, 基于模型, 基于synthesis data, 基于data augmentation

Main Approach

- 问题定义

  给出一个有标签数据集 D={(x_i, y_i)}, $x_i \in \mathbb{R}^d$, $y_i \in C$, C是类别的集合. C被分为*base class*和*novel class*, C_b和$\mathbb{C}_n$, 互不相交的完全划分. 任务的目标是在base class上训练一个模型, 并且也能在novel class的数据集上有较好的泛化效果

  N-way-K-shot : N个类别, 每个类别仅有K个数据的few-shot learning. available的有标签数据被称为support set, 测试性能的测试集被称为query set

- Tukey's Ladder of Powers Transformation

  为了让特征分布更加的Gaussian-like, 作者在实现的时候首先把support set和query set使用Tukey's Ladder of Powers transformation进行转换. 该转换可以产生更加skewness (偏度)的分布, 并且使得整个分布更加的像高斯分布

$$\tilde{\mathbf{x}} = \begin{cases} \boldsymbol{x}^{\lambda} & \text{if } \lambda \neq 0 \\ \log(\boldsymbol{x}) & \text{if } \lambda = 0 \end{cases} \tag{3}$$

where λ is a hyper-parameter to adjust how to correct the distribution. The original feature can be recovered by setting λ as 1. Decreasing λ makes the distribution less positively skewed and vice versa.

- Distribution Calibration

  在base class中首先在feature space计算统计量, 如mean和variance. 然后在feature space按照euclidean距离来计算novel class的归属类, 根据距离选择top K个base class :

$$\mathbb{S}_d = \{-\|\boldsymbol{\mu}_i - \tilde{\boldsymbol{x}}\|^2 \mid i \in C_b\},$$
$$\mathbb{S}_N = \{i \mid -\|\boldsymbol{\mu}_i - \tilde{\boldsymbol{x}}\|^2 \in topk(\mathbb{S}_d)\},$$

  然后按照如下公式来做distribution calibration

$$\boldsymbol{\mu}' = \frac{\Sigma_{i \in \mathbb{S}_N} \boldsymbol{\mu}_i + \tilde{\boldsymbol{x}}}{k+1}, \boldsymbol{\Sigma}' = \frac{\Sigma_{i \in \mathbb{S}_N} \boldsymbol{\Sigma}_i}{k} + \alpha,$$

  取top k个base class和采样的样本计算均值, 同时计算covariance加上采样至calibrated distribution数据的离散程度, degree of dispersion

- 对与few-shot leanring, 这个过程需要多次进行, 这可以避免由一个特定的样本带来的bias, 并且能够实现diverse and accurate distribution estimation
- 使用calibrated distribution. 对于一系列类别为Y的calibrated统计量, 可以在calibrated gaussian distribution中采样生成以类别为Y的数据. 同样使用cross-entropy训练模型即可

Experiment

Table 2: 5way1shot and 5way5shot classification accuracy (%) on *mini*ImageNet and CUB with 95% confidence intervals. The numbers in **bold** have intersecting confidence intervals with the most accurate method.

Methods	*mini*ImageNet		*CUB*	
	5way1shot	5way5shot	5way1shot	5way5shot
Optimization-based				
MAML (Finn et al. (2017))	48.70 ± 1.84	63.10 ± 0.92	50.45 ± 0.97	59.60 ± 0.84
Meta-SGD (Li et al. (2017))	50.47 ± 1.87	64.03 ± 0.94	53.34 ± 0.97	67.59 ± 0.82
LEO (Rusu et al. (2019))	61.76 ± 0.08	77.59 ± 0.12	-	-
E3BM (Liu et al. (2020b))	63.80 ± 0.40	80.29 ± 0.25	-	-
Metric-based				
Matching Net (Vinyals et al. (2016))	43.56 ± 0.84	55.31 ± 0.73	56.53 ± 0.99	63.54 ± 0.85
Prototypical Net (Snell et al. (2017))	54.16 ± 0.82	73.68 ± 0.65	72.99 ± 0.88	86.64 ± 0.51
Baseline++ (Chen et al. (2019a))	51.87 ± 0.77	75.68 ± 0.63	67.02 ± 0.90	83.58 ± 0.54
Variational Few-shot(Zhang et al. (2019))	61.23 ± 0.26	77.69 ± 0.17	-	-
Negative-Cosine(Liu et al. (2020a))	62.33 ± 0.82	80.94 ± 0.59	72.66 ± 0.85	89.40 ± 0.43
Generation-based				
MetaGAN (Zhang et al. (2018))	52.71 ± 0.64	68.63 ± 0.67	-	-
Delta-Encoder (Schwartz et al. (2018))	59.9	69.7	69.8	82.6
TriNet (Chen et al. (2019b))	58.12 ± 1.37	76.92 ± 0.69	69.61 ± 0.46	84.10 ± 0.35
Meta Variance Transfer (Park et al. (2020))	-	67.67 ± 0.70	-	80.33 ± 0.61
Maximum Likelihood with DC (Ours)	66.91 ± 0.17	80.74 ± 0.48	77.22 ± 0.14	89.58 ± 0.27
SVM with DC (Ours)	$\mathbf{67.31 \pm 0.83}$	$\mathbf{82.30 \pm 0.34}$	$\mathbf{79.49 \pm 0.33}$	$\mathbf{90.26 \pm 0.98}$
Logistic Regression with DC (Ours)	$\mathbf{68.57 \pm 0.55}$	$\mathbf{82.88 \pm 0.42}$	$\mathbf{79.56 \pm 0.87}$	$\mathbf{90.67 \pm 0.35}$

Table 3: 5way1shot and 5way5shot classification accuracy (%) on *tiered*ImageNet (Ren et al., 2018). The numbers in **bold** have intersecting confidence intervals with the most accurate method.

Methods	*tiered*ImageNet	
	5way1shot	5way5shot
Matching Net (Vinyals et al. (2016))	68.50 ± 0.92	80.60 ± 0.71
Prototypical Net (Snell et al. (2017))	65.65 ± 0.92	83.40 ± 0.65
LEO (Rusu et al. (2019))	66.33 ± 0.05	82.06 ± 0.08
E3BM (Liu et al. (2020b))	71.20 ± 0.40	85.30 ± 0.30
DeepEMD (Zhang et al., 2020)	71.16 ± 0.87	86.03 ± 0.58
Maximum Likelihood with DC (Ours)	75.92 ± 0.60	87.84 ± 0.65
SVM with DC (Ours)	$\mathbf{77.93 \pm 0.12}$	$\mathbf{89.72 \pm 0.37}$
Logistic Regression with DC (Ours)	$\mathbf{78.19 \pm 0.25}$	$\mathbf{89.90 \pm 0.41}$

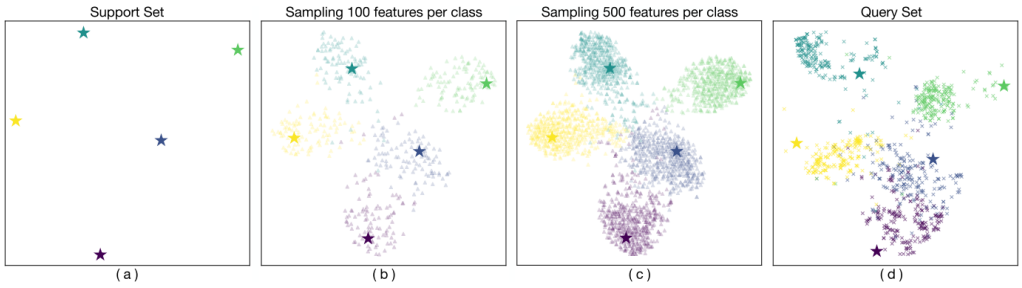

Figure 2: t-SNE visualization of our distribution estimation. Different colors represent different classes. '★' represents support set features, 'x' in figure (d) represents query set features, '▲' in figure (b)(c) represents generated features.

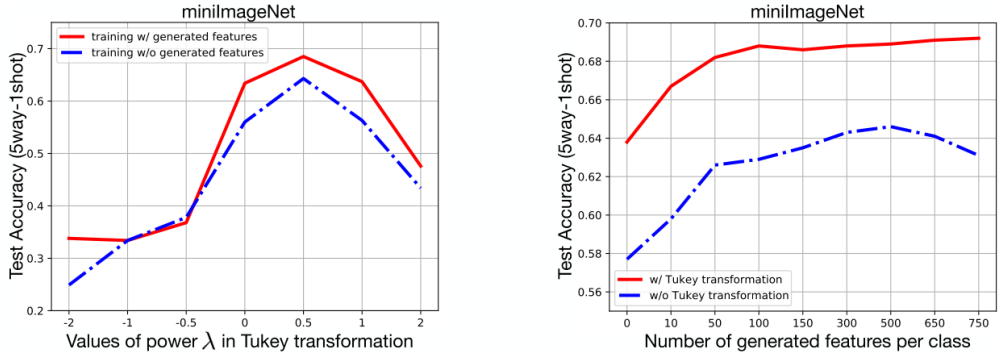

Figure 3: Left: Accuracy when increasing the power in Tukey's transformation when training with (red) or without (blue) the generated features. Right: Accuracy when increasing the number of generated features with the features are transformed by Tukey's transformation (red) and without Tukey's transformation (blue).

小结

这篇文章主要提出了一种distribution calibration的方法. 但有个前提是, 作者都说所有的特征分布都假设是高斯分布, 为此作者还提出了一种Tukey's Ladder Transformation将非高斯分布转换为更像高斯分布. 第二, 作者提出相似的类别中, 数据的统计量也是相似的, 这点无法确定? 如果这个一直成立, 那么很多问题都可以进行. 在本文中作者也没有给出更多的解释, 是observed到的一个规律. 作者将相思类的统计量对few数据进行distribution calibration, 使得能够获取到少样本新类数据的分布, 这样就可以从整个分布中取更多的数据训练

只是这个相似类的统计量相似, 没有一个明确的证明, 还是挺麻烦的... 这篇文章的reviewer居然也没有批判...