论文分享

Adaptive Cross-Modal Few-shot Learning

系统芯片实验室 | 付雯

2021/12/18

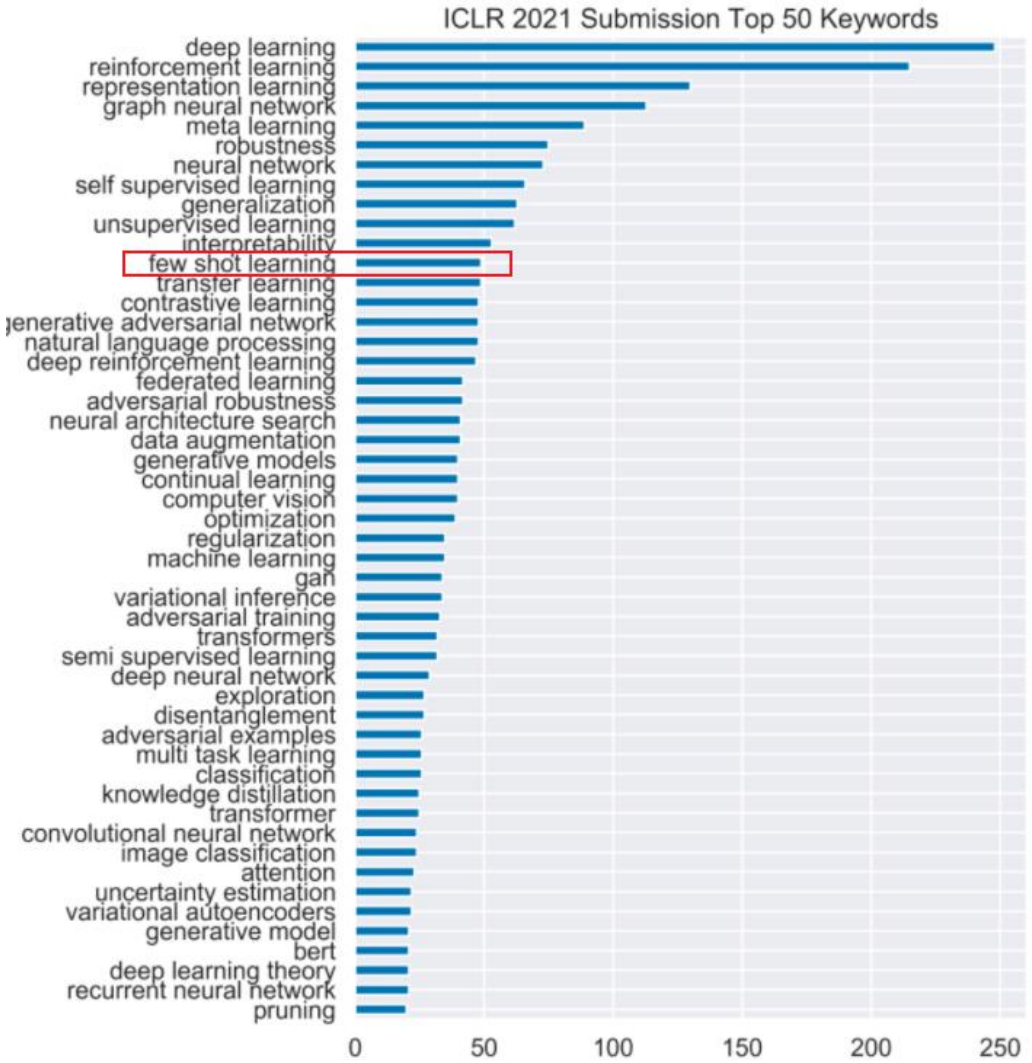
目 录

目　录

研究背景

- 近年来，由于强大的计算设备（如GPU、TPU）、大数据集（如拥有2万多个类的ImageNet）以及先进的模型和算法（如卷积神经网络）的出现，深度学习快速发展，在许多领域击败人类。
  - 残差网络在ImageNet上的分类性能优于人类（2015年）
  - AlphaGo在围棋比赛中击败人类冠军（2016年）
  - ．．．
- 上述成功的人工智能应用都依赖于从大规模数据中进行学习。
  - AlphaGo自我对弈高达3000万盘
  - ISLVRC 2012训练集有128万多张图像
  - ．．．

研究背景



- 在很多场景下，收集大量的有标签数据是**非常昂贵、困难、甚至不可能**的
  - 医疗数据
  - 涉及隐私、伦理以及安全问题的数据
  - 化学数据
  - . . . .
- “是否能仅使用数量有限的样本学习得到一个好的模型？”为了解决这个问题，提出一种新的机器学习范式，称为Few-Shot Learning（FSL）.
- FSL已经成为机器学习的发展中一个十分重要的课题，不论是学术界还是工业界都高度关注。

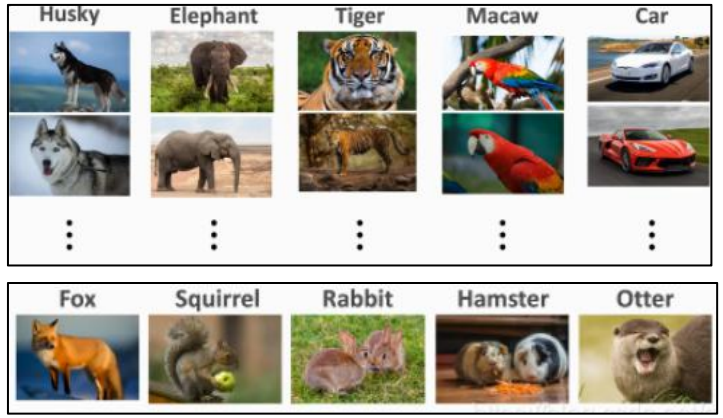
相关定义

- 机器学习
  - 对于一个计算机程序，如果它以P度量的性能在任务T上随经验E改善，那么认为它可以从经验E中学习。
- 小样本学习
  - FSL是一种机器学习问题(由 E, T和P指定)，其中E只包含任务T有限数量的监督信息。
- FSL的三种典型场景
  - 作为模仿人类学习的测试。
  - 当很难或不可能获得足够的有监督信息时。
  - 帮助减轻使用监督信息时要收集大量样本的负担。

相关定义

数据集划分

- **base classes**:用于训练一个深度神经网络，类别多，每类样本也相对较多。
- **novel classes**:用于测试，与base classes中类别没有交集,每个类别样本数目很少(＜5张)。

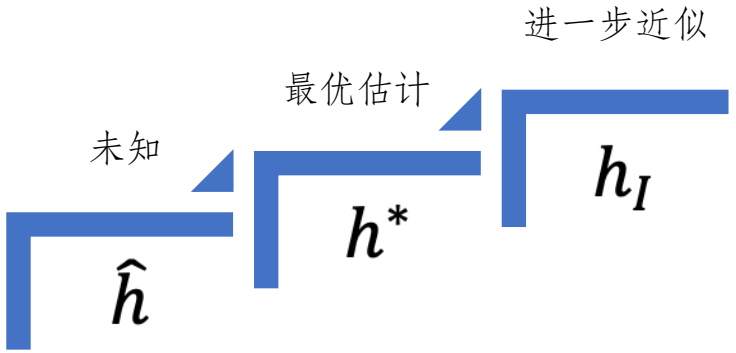


训练模式

- 为了在训练时模仿测试时的情形，采用episode为单位训练。
  - 在训练/测试时，从数据集中随机抽取 C 个类别，每个类别 K(<=5) 个样本（总共 CK 个数据），作为support set输入；
  - 再从这 C 个类中剩余的数据中抽取一批样本作为模型的预测对象构成query set；
  - 模型从 C*K 个数据中学会如何区分这 C 个类别，这样的任务被称为 c-way k-shot 任务；

问题分析

- 符号与术语
  - 考虑图像分类任务，有数据集 $D = \{D_{base}, D_{novel}\}$
  - $p(x,y)$为输入x和输出y的联合概率分布，$\hat{h}$为$x \to y$的最优函数，FSL通过拟合D_{base}并在D_{novel}上进行测试来发现$\hat{h}$。
  - FSL模型确定了函数$h(\cdot;\theta)$的假设空间H，其中θ表示H使用的所有参数
  - FSL算法是一种优化策略，它搜索函数空间H以找到参数化最佳$h^* \in H$的θ
  - FSL性能由损失函数$l(\hat{y},y)$衡量,其中$\hat{y} = h(x;\theta)$
  - 给定一个函数h，希望将其期望风险R最小化: $R(h) = \int l(h(x),y)dp(x,y) = E[l(h(x),y)]$
  - $p(x,y)$未知，使用经验风险近似: $R_I(h) = \frac{1}{I}\sum_{i=1}^{I} l(h(x_i),y_i)$
  - 与样本数相关，I越大，经验风险越接近期望风险
- $\hat{h} = \arg\min_h R(h)$是使期望风险最小的函数;
- $h^* = \arg\min_{h\in H} R(h)$ 是H中使期望风险最小化的函数;
- $h_I = \arg\min_{h\in H} R_I$是$H$中使经验风险最小化的函数;

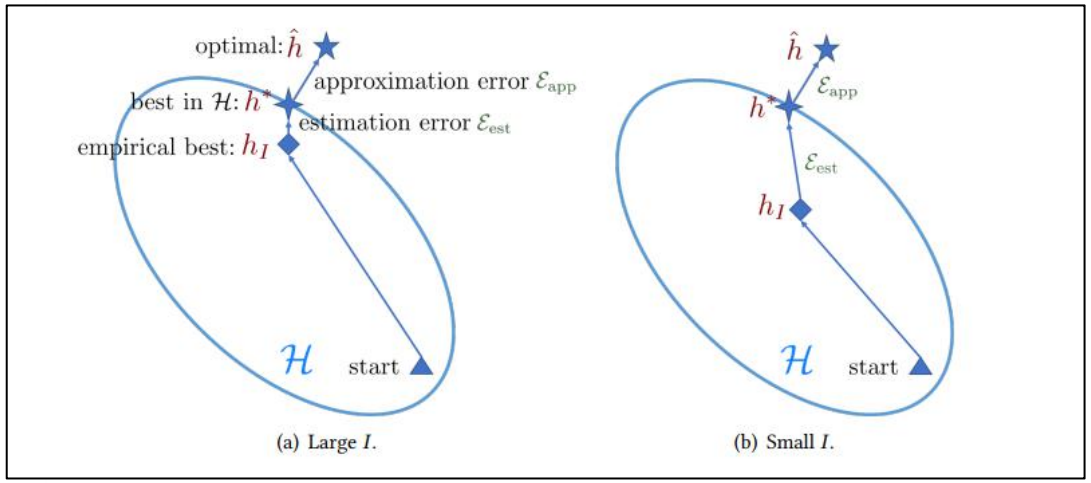


[1]经验风险最小化详细推导见《统计学习方法》

问题分析

损失

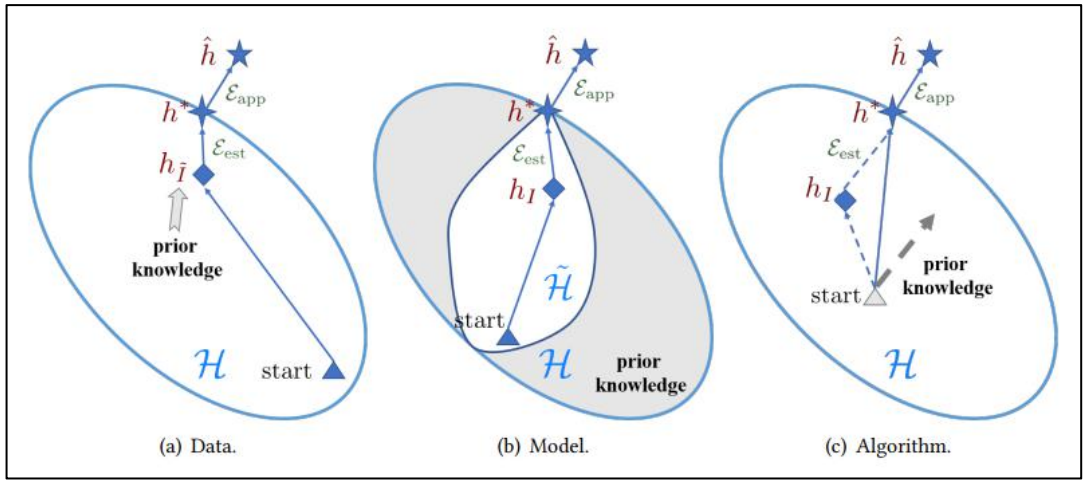
$$\mathbb{E}[R(h_I) - R(\hat{h})] = \underbrace{\mathbb{E}[R(h^*) - R(\hat{h})]}_{\mathcal{E}_{\text{app}}(\mathcal{H})} + \underbrace{\mathbb{E}[R(h_I) - R(h^*)]}_{\mathcal{E}_{\text{est}}(\mathcal{H}, I)},$$



小样本情况下，经验风险不可靠！

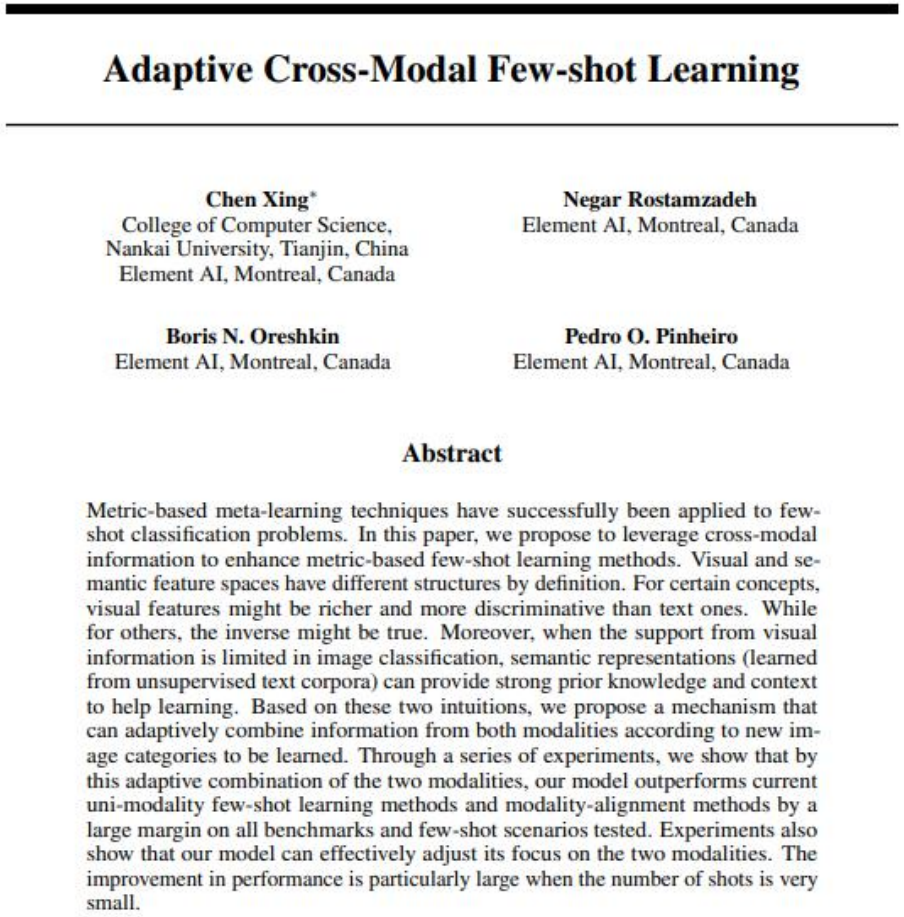
解决办法

- 为了缓解FSL中不可靠的经验风险最小化的问题，使用先验知识，现有的FSL工作可以分为以下几个类型:



目　录

作者与发表概况



Adaptive Cross-Modal Few-shot Learning

Chen Xing*
College of Computer Science,
Nankai University, Tianjin, China
Element AI, Montreal, Canada

Negar Rostamzadeh
Element AI, Montreal, Canada

Boris N. Oreshkin
Element AI, Montreal, Canada

Pedro O. Pinheiro
Element AI, Montreal, Canada

Abstract

Metric-based meta-learning techniques have successfully been applied to few-shot classification problems. In this paper, we propose to leverage cross-modal information to enhance metric-based few-shot learning methods. Visual and semantic feature spaces have different structures by definition. For certain concepts, visual features might be richer and more discriminative than text ones. While for others, the inverse might be true. Moreover, when the support from visual information is limited in image classification, semantic representations (learned from unsupervised text corpora) can provide strong prior knowledge and context to help learning. Based on these two intuitions, we propose a mechanism that can adaptively combine information from both modalities according to new image categories to be learned. Through a series of experiments, we show that by this adaptive combination of the two modalities, our model outperforms current uni-modality few-shot learning methods and modality-alignment methods by a large margin on all benchmarks and few-shot scenarios tested. Experiments also show that our model can effectively adjust its focus on the two modalities. The improvement in performance is particularly large when the number of shots is very small.

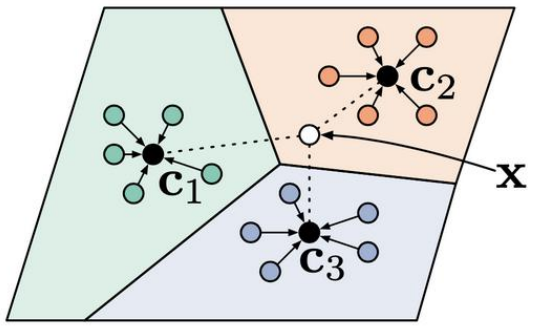
- 发表会议

Neural Information Processing Systems
NeurIPS 2019

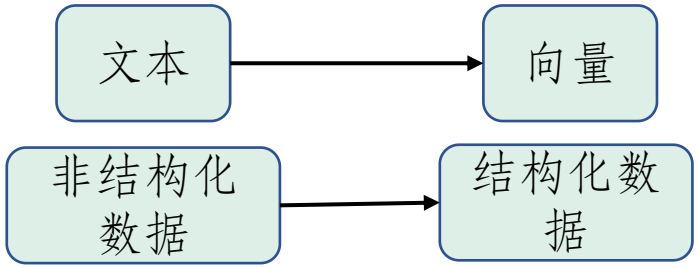
- 作者：

南开大学
Element AI

相关方法



3 way 5 shot



➢ Prototypical network

- 把样本投影到一个低维空间，计算每个样本类别的原型：

$$c_k = \frac{1}{S_k} \sum_{(x_i, y_i) \in S_k}$$

- 在分类的时候，通过对比目标到每个原型的距离，从而得出目标的类别

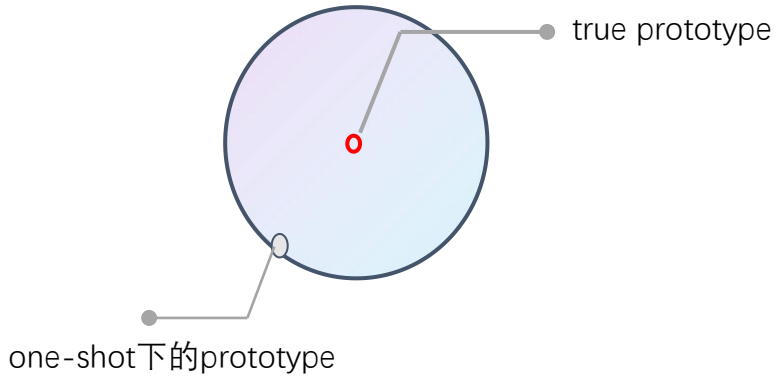
➢ Glove

- 基于全局词频统计的词表征工具，它可以把一个单词表达成一个由实数组成的向量，这些向量捕捉到了单词之间一些语义特性，比如相似性等。
- 通过对向量的运算，比如欧几里得距离或者cosine相似度，可以计算出两个单词之间的语义相似性。

http://www.fanyeong.com/2018/02/19/glove-in-detail/

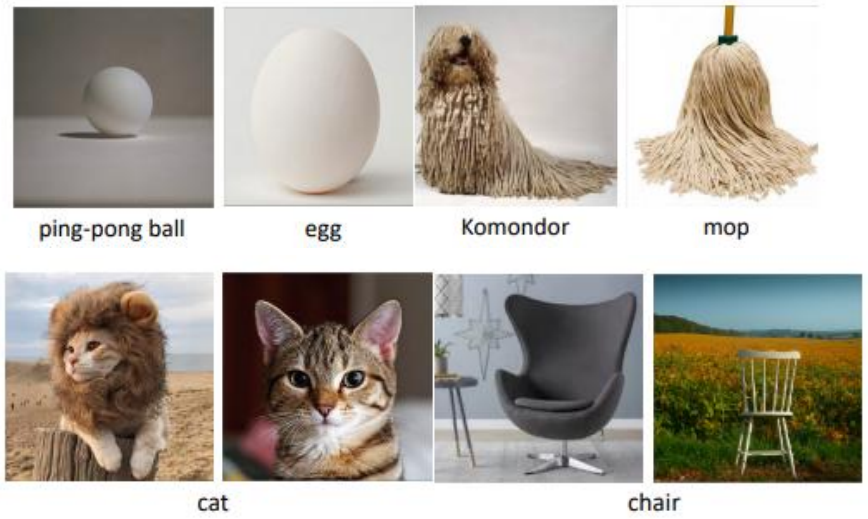
拟解决问题

- 单模态：**当来自视觉方面的支持图像的数量非常小时，图像分类中视觉信息的支持有限，模态提供的信息往往是嘈杂、局部的。**
  - 语义表示（从无监督文本语料库中学习）可以提供强大的先验知识和上下文来帮助学习。 通过引入文本信息（先验知识），帮助原型的学习，使少数样本的特征平均值更好的代表原型
- 多模态：**模态对齐问题，**视觉和语义特征空间具有不同的结构，在训练期间需要对齐两种模态，强制他们具有相同的语义结构。



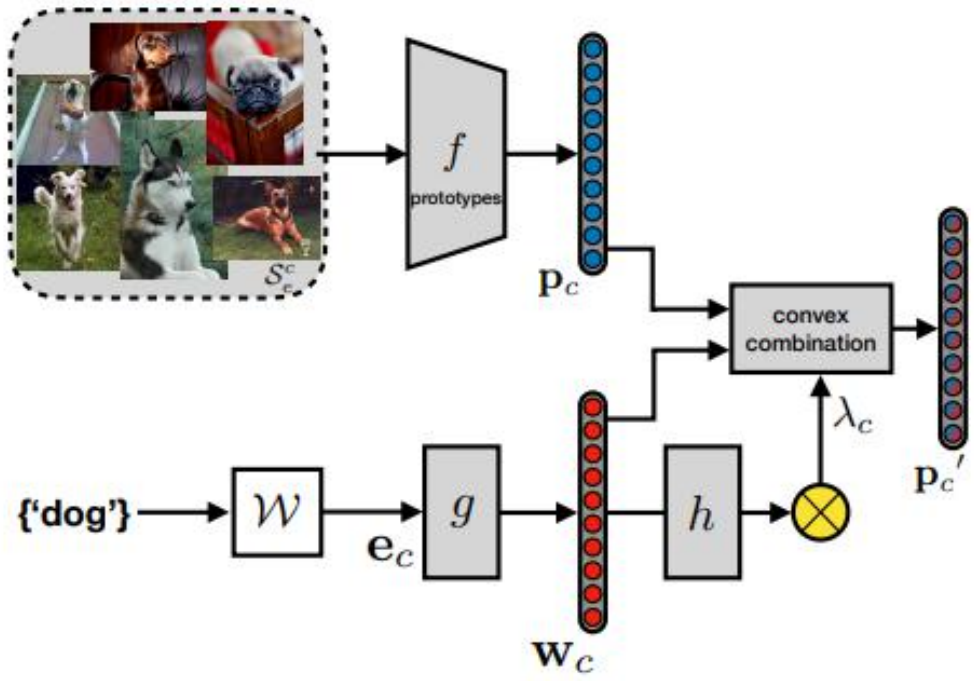
拟解决问题

- 零样本学习比较极端，在测试时没有给出视觉信息，算法需要完全依赖辅助（例如，文本）模态。
- 在另一个极端，当标记图像样本的数量很大时，往往会忽略辅助模态，因为已经能够很好地泛化。
- 小样本学习场景介于这两个极端之间。 因此，本文假设视觉和语义信息都可以用于小样本学习。
- **将它们视为两个独立的知识源，并根据不同的场景自适应地利用这两种模态。**
  - 提出了自适应模态混合机制(Adaptive Modality Mixture Mechanism, AM3)，这是一种自适应、有选择性地结合视觉和语义两种模态信息的方法，用于小样本学习。



创新点探究

➢ 自适应模态混合模型：



- Visiual backbone f: ProtoNet++ / TADAM[2]
- Word embeddings w: Glove
- Semantic transformation g : FC300 - 512
- Transformation h: FC300 − 1

$$\lambda_c = \frac{1}{1+\exp(-h(\mathbf{w}_c))}$$

- 最终的类别原型是视觉和语义特征表示的凸组合。混合系数取决于语义标签嵌入。

$$\mathbf{p}_c' = \lambda_c \cdot \mathbf{p}_c + (1-\lambda_c) \cdot \mathbf{w}_c$$

[2]TADAM: Task dependent adaptive metric for improved few-shot learning

创新点探究

Algorithm 1: Training episode loss computation for adaptive cross-modality few-shot learning. M is the total number of classes in the training set, N is the number of classes in every episode, K is the number of supports for each class, K_Q is the number of queries for each class, $\mathcal{W}$ is the pretrained label embedding dictionary.

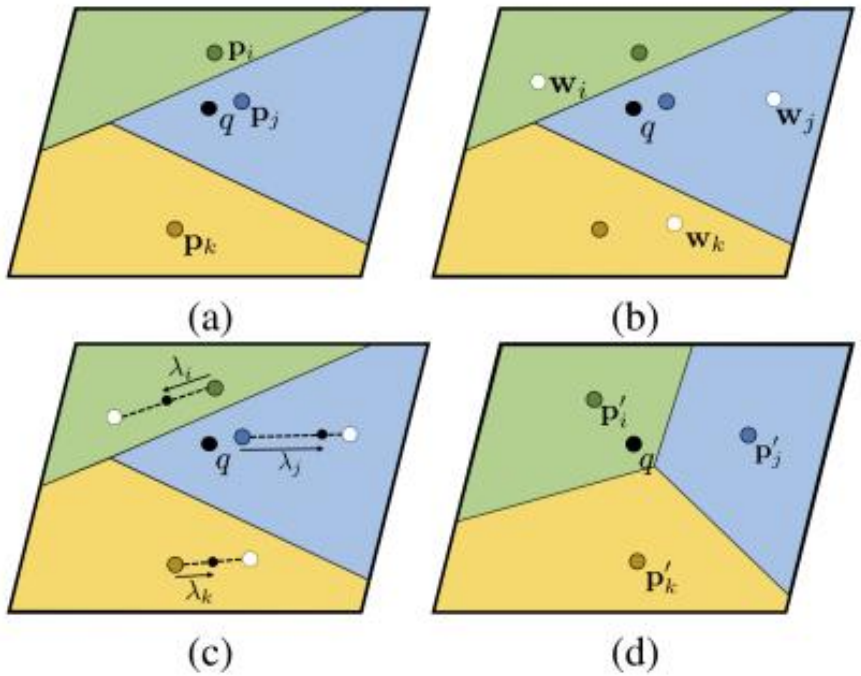
Input: Training set $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_i, y_i \in \{1, ..., M\}$. $\mathcal{D}^c_{\text{train}} = \{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{train}} \mid y_i = c\}$.
Output: Episodic loss $\mathcal{L}(\theta)$ for sampled episode e.

{Select N classes for episode e}
$C \leftarrow RandomSample(\{1, ..., M\}, N)$
{Compute cross-modal prototypes}
for c in C **do**
  $\mathcal{S}^c_e \leftarrow RandomSample(\mathcal{D}^c_{\text{train}}, K)$
  $\mathcal{Q}^c_e \leftarrow RandomSample(\mathcal{D}^c_{\text{train}} \setminus \mathcal{S}^c_e, K_Q)$
  $\mathbf{p}_c \leftarrow \frac{1}{|\mathcal{S}^c_e|} \sum_{(s_i, y_i) \in \mathcal{S}^c_e} f(s_i)$
  $\mathbf{e}_c \leftarrow LookUp(c, \mathcal{W})$
  $\mathbf{w}_c \leftarrow g(\mathbf{e}_c)$
  $\lambda_c \leftarrow \frac{1}{1+\exp(-h(\mathbf{w}_c))}$
  $\mathbf{p}'_c \leftarrow \lambda_c \cdot \mathbf{p}_c + (1 - \lambda_c) \cdot \mathbf{w}_c$
end for
{Compute loss}
$\mathcal{L}(\theta) \leftarrow 0$
for c in C **do**
  for (q_t, y_t) in Q^c_e **do**
    $\mathcal{L}(\theta) \leftarrow \mathcal{L}(\theta) + \frac{1}{N \cdot K}[d(f(q_t), \mathbf{p}'_c)) + \log \sum_k \exp(-d(f(q_t), \mathbf{p}'_k))]$
  end for
end for

- 对比学习
  - 第一项：同类样本更紧凑
  - 第二项：不同类样本互相远离

创新点探究

自适应模态混合模型：



- (a)与查询样本q最接近的视觉原型是p_j。
- (b)语义原型。
- (c)混合模型在语义嵌入的情况下修改了原型的位置。
- (d)更新后，与查询样本最接近的原型现在是类别 i 中的一个，修正了分类。

结果分析

- miniImageNet
  - 从ImageNet数据集中选择了60,000张图像构成的，共100个类别，每个类别有600张图像，每幅图像的尺寸为84*84。
  - 通常选择其中80个类别的图像作为训练集，剩余的20个类别的图像作为验证集。
- tieredImageNet
  - 从ImageNet数据集中选取，包含34个大类，每个大类有包含10-30个小类，共计608个类别，779,165张图像（平均每个类别包含1281张图片）。

结果分析

与单模态的对比结果

miniImageNet

Model	Test Accuracy		
	5-way 1-shot	5-way 5-shot	5-way 10-shot
Uni-modality few-shot learning baselines			
Matching Network [53]	43.56 ± 0.84%	55.31 ± 0.73%	-
Prototypical Network [47]	49.42 ± 0.78%	68.20 ± 0.66%	74.30 ± 0.52%
Discriminative k-shot [2]	56.30 ± 0.40%	73.90 ± 0.30%	78.50 ± 0.00%
Meta-Learner LSTM [38]	43.44 ± 0.77%	60.60 ± 0.71%	-
MAML [7]	48.70 ± 1.84%	63.11 ± 0.92%	-
ProtoNets w Soft k-Means [39]	50.41 ± 0.31%	69.88 ± 0.20%	-
SNAIL [32]	55.71 ± 0.99%	68.80 ± 0.92%	-
CAML [16]	59.23 ± 0.99%	72.35 ± 0.71%	-
LEO [41]	61.76 ± 0.08%	77.59 ± 0.12%	-
AM3 and its backbones			
ProtoNets++	56.52 ± 0.45%	74.28 ± 0.20%	78.31 ± 0.44%
AM3-ProtoNets++	65.21 ± 0.30%	75.20 ± 0.27%	78.52 ± 0.28%
TADAM [35]	58.56 ± 0.39%	76.65 ± 0.38%	80.83 ± 0.37%
AM3-TADAM	**65.30 ± 0.49%**	**78.10 ± 0.36%**	**81.57 ± 0.47 %**

tieredImageNet

Model	Test Accuracy	
	5-way 1-shot	5-way 5-shot
Uni-modality few-shot learning baselines		
MAML† [7]	51.67 ± 1.81%	70.30 ± 0.08%
Proto. Nets with Soft k-Means [39]	53.31 ± 0.89%	72.69 ± 0.74%
Relation Net† [50]	54.48 ± 0.93%	71.32 ± 0.78%
Transductive Prop. Nets [28]	54.48 ± 0.93%	71.32 ± 0.78%
LEO [41]	66.33 ± 0.05%	81.44 ± 0.09%
AM3 and its backbones		
ProtoNets++	58.47 ± 0.64%	78.41 ± 0.41%
AM3-ProtoNets++	67.23 ± 0.34%	78.95 ± 0.22%
TADAM [35]	62.13 ± 0.31%	81.92 ± 0.30%
AM3-TADAM	**69.08 ± 0.47%**	**82.58 ± 0.31%**

方法的增益随shot数的减少而增大。这说明视觉内容越低，语义信息对分类越重要

结果分析

与模态对齐的对比结果

miniImageNet

Model	Test Accuracy		
	5-way 1-shot	5-way 5-shot	5-way 10-shot
Modality alignment baselines extended to metric-based FSL framework			
DeViSE-FSL	56.99 ± 1.33%	72.63 ± 0.72%	76.70 ± 0.53%
ReViSE-FSL	57.23 ± 0.76%	73.85 ± 0.63%	77.21 ± 0.31%
f-CLSWGAN-FSL	58.47 ± 0.71%	72.23 ± 0.45%	76.90 ± 0.38%
CADA-VAE-FSL	61.59 ± 0.84%	75.63 ± 0.52%	79.57 ± 0.28%
AM3 and its backbones			
ProtoNets++	56.52 ± 0.45%	74.28 ± 0.20%	78.31 ± 0.44%
AM3-ProtoNets++	65.21 ± 0.30%	75.20 ± 0.27%	78.52 ± 0.28%
TADAM [35]	58.56 ± 0.39%	76.65 ± 0.38%	80.83 ± 0.37%
AM3-TADAM	**65.30 ± 0.49%**	**78.10 ± 0.36%**	**81.57 ± 0.47 %**

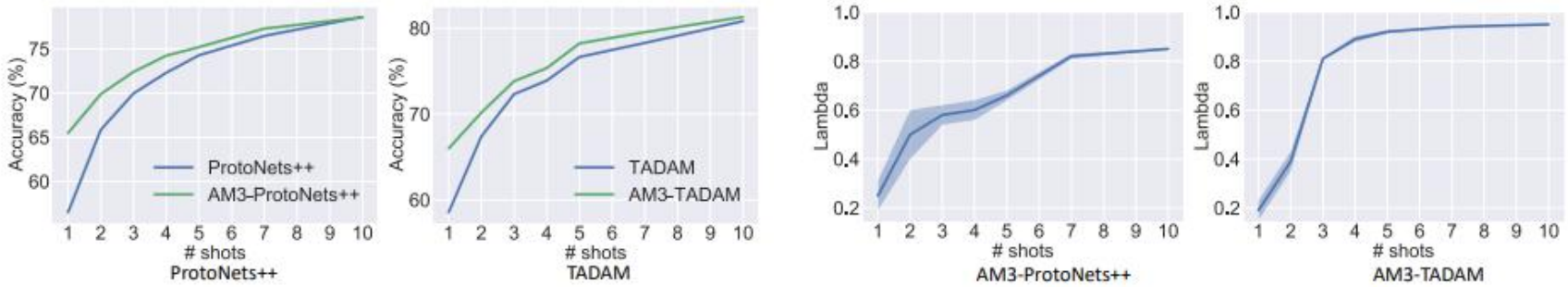
tieredImageNet

Model	Test Accuracy	
	5-way 1-shot	5-way 5-shot
Modality alignment baselines extended to metric-based FSL framework		
DeViSE-FSL	61.78 ± 0.43%	77.17 ± 0.81%
ReViSE-FSL	62.77 ± 0.31%	77.27 ± 0.42%
CADA-VAE-FSL	63.16 ± 0.93%	78.86 ± 0.31%
AM3 and its backbones		
ProtoNets++	58.47 ± 0.64%	78.41 ± 0.41%
AM3-ProtoNets++	67.23 ± 0.34%	78.95 ± 0.22%
TADAM [35]	62.13 ± 0.31%	81.92 ± 0.30%
AM3-TADAM	**69.08 ± 0.47%**	**82.58 ± 0.31%**

- 将现有的模态对齐方法扩展到小样本学习框架，大多数模态对齐方法，表现不如本文方法。
- 这表明，虽然模态对齐方法在ZSL中对交叉模态是有效的，但它不太适合few-shot场景。**一个可能的原因是，当两种模式对齐时，因为两个不同的结构被迫对齐，来自双方的一些信息可能会丢失。**

适应性分析

- 适应性机制是性能提高的主要原因。
  - miniImageNet上测试ProtoNets++和TADAM在1-10shot场景下的准确性。



(a) Accuracy vs. # shots　　(b) λ vs. # shots

思考与讨论

谢谢聆听

请老师同学批评指正