Paper-Reading-Notes $(universal)$

```
### MainTitle

- [x] (conf) [[paperswithcode]()]
    - Author et al. "Title"


| 核心在哪? | 精读? 代码? | 关键词? | 亮点? | 笔记时间? |
| --------- | ----------- | ------- | ----- | --------- |
|           |             |         |       |           |

---



+ **背景? 提出了什么问题?**
+ **为了解决此问题提出了什么具体的idea?**
+ **如何从该idea形式化地对问题建模、简化并解决的?**
+ **理论方面证明的定理与推导过程?**
+ **这个任务/解决方法有什么意义?**
+ **对论文的讨论/感想?**
```

Adaptive Task Sampling (class-pair based)

- ☑ **Adaptive Task Sampling (class-pair based)** (ECCV 2020) [paperswithcode]
    - Liu et al. "Adaptive Task Sampling for Meta-Learning"

核心在哪?	精读? 代码?	关键词?	亮点?	笔记时间?

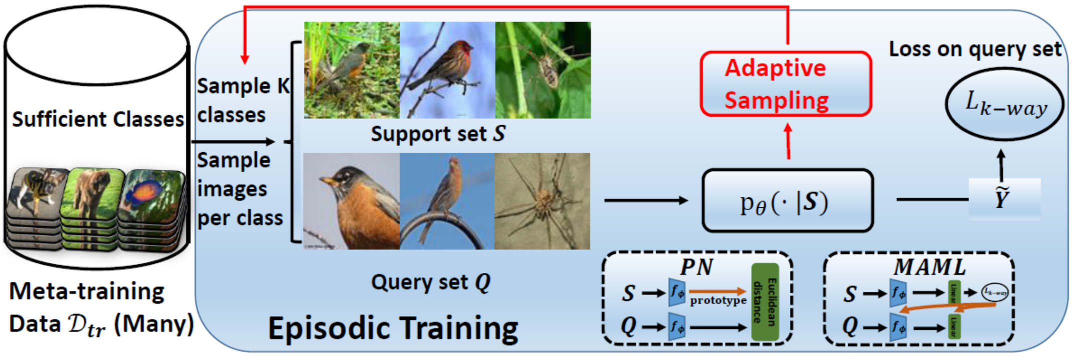


Fig. 1: The episodic training paradigm for meta-learning few-shot classification.

While a rich line of work focuses **solely on how to extract meta-knowledge across tasks**, we exploit the complementary problem on **how to generate informative tasks**.

We argue that the randomly sampled tasks could be **sub-optimal and uninformative** (e.g., the task of classifying “dog” from “laptop” is often trivial) to the meta-learner. In this paper, we propose **an adaptive task sampling method** to improve the generalization performance.

In summary, our work makes the following contributions. **(1)** We propose a class-pair based adaptive task sampling approach for meta-learning methods, to improve the generalization performance on unseen tasks. **(2)** We further develop a greedy class-pair based approach that not only significantly reduces the complexity of task distribution computation, but also guarantees the generation of an identical distribution as that in the non-greedy approach. **(3)** We study the impact of the adaptive task sampling method by integrating it with various meta-learning approaches and performing comprehensive experiments on the miniImageNet and CIFAR-FS few-shot datasets, which quantitatively demonstrates the superior performance of our method. **(4)** We also conduct an extensive investigation of different sampling strategies, including class-based method, easy class-pair based method and uncertain class-pair based method. The results show that hard class-pair based sampling consistently leads to more accurate results.

- 背景? 提出了什么问题?

  an episodic training paradigm.

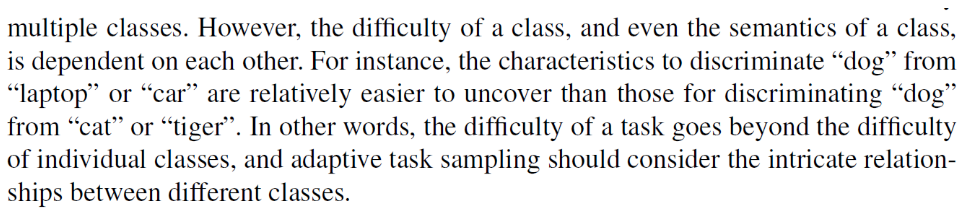
  A series of few-shot tasks are sampled from meta-training data for **the extraction of transferable knowledge across tasks**, which is then applied to **new** few-shot classification tasks consisting of unseen classes during the meta-testing phase.

  - 问题的提出:

    Despite their noticeable improvements, these meta-learning approaches leverage **uniform sampling** over classes to generate few-shot tasks, which ignores the intrinsic relationships between classes when forming episodes.

    上述方法是 uniform sampling, 这忽略了forming episodes时候类之间的内在联系. 在一些领域中 比如集成学习Adaboost对challenging training examples优先训练后续分类器.

    - 很自然的提出问题: Can we perform adaptive task sampling and create more difficult tasks for meta-learning?
    - 难点: one key challenge in task sampling is to define the difficulty of a task.



multiple classes. However, the difficulty of a class, and even the semantics of a class, is dependent on each other. For instance, the characteristics to discriminate “dog” from “laptop” or “car” are relatively easier to uncover than those for discriminating “dog” from “cat” or “tiger”. In other words, the difficulty of a task goes beyond the difficulty of individual classes, and adaptive task sampling should consider the intricate relationships between different classes.

- **Review for Episodic Training:**

  1. In each episode of meta-training, we first **sample K classes** $\mathbb{L}^K \sim \mathbb{C}_{tr}$.
  2. Then, we **sample M and N labelled images per class** in $\mathbb{L}^K$ to **construct** the support set $\mathbb{S}=\{(s_m, y_m)_m\}$ and query set $\mathbb{Q}=\{(q_n, y_n)_n\}$, respectively.

     从之前sample的类里面sample出 $\mathbb{S}, \mathbb{Q}$.
  3. The episodic training for few-shot learning 是在**query set**上最优, The model is parameterized by θ and the loss is the negative loglikelihood of the true class of each query sample, 即优化:

     $$\ell(\theta)=\underset{(S,Q)}{\mathbb{E}}\left[-\sum_{(q_n,y_n)\in Q}\log p_\theta\left(y_n \mid q_n, S\right)\right]$$

     $p_\theta\left(y_n \mid q_n, S\right)$ 是在support set上的分类概率.

     注意啊上面的损失是在 **query set** 上测的, 但是训练(后验)是在**support**上的.

     梯度下降 $\Delta\ell(\theta)$.

- **Review for Instance-base Adaptive Sampling for SGD:**

  Select Sample 的概率:

  - 第一次:

    $$p_0(i \mid \mathbb{D})=\frac{1}{|\mathbb{D}|}$$

  - 之后:

    instance i at iteration $t+1$ **according to the current prediction probability** $p\left(y_i \mid x_i\right)$ and the selection probability at previous iteration $p^t(i)$

    $$p^{t+1}(i) \propto \left(p^t(i)\right)^\tau e^{\alpha\left(1-p\left(y_i \mid x_i\right)\right)}$$

    where the hyperparameters τ is a discounting parameter and α scales the influence of current prediction.

    This multiplicative update method has a close relation to maximum loss minimization [47] and AdaBoost [16].

- 为了解决此问题提出了什么具体的**idea?**

  a straightforward extension of the instance-based sampling.

- 如何从该**idea**形式化地对问题建模、简化并解决的**?**

  - **Class-based Sampling**:

    We propose a class-based sampling (c-sampling) approach that **updates the class selection probability $p_C^{t+1}(c)$ in each episode.**

    具体选择类概率的更新公式如下:

Given $\mathbb{S}^t$ and $\mathbb{Q}^t$ at episode t, we could update the class selection probability for each class in current episode $c \in \mathbb{L}_K^t$ in the following way,

$$p_C^{t+1}(c) \propto (p^t(c))^\tau e^{\alpha \frac{\sum_{(q_n,y_n)\in\mathbb{Q}^t} \mathbb{I}[c\neq y_n]p(c|q_n,\mathbb{S}^t)+\mathbb{I}[c=y_n](1-p(c|q_n,\mathbb{S}^t))}{NK}}. \tag{3}$$

Note that we average the prediction probability of classifying each query sample n into incorrect classes in $\mathbb{L}_K^t$. Then we can sample K classes without replacement to construct the category set $\mathbb{L}_K^{t+1}$ for the next episode.

每个类的难度不是独立的.

取出类别二元组，无向概率图模型 马尔可夫随机场，这里不是最大团.

更新$C(i,j)$，该类别对在上一次就混淆了，接下来就要挑这个.

不能接受的计算复杂度，则使用贪心算法.

- **理论方面证明的定理与推导过程?**
- **这个任务/解决方法有什么意义?**
- **对论文的讨论/感想?**

Stochastic optimization with importance sampling for regularized loss minimization. (ICML 2015)

> Diversity with Cooperation: Ensemble Methods for Few-Shot Classification
>
> Transferable Meta Learning Across Domains
>
> A Closer Look at Few-shot Classification
>
> Boosting Few-shot Image Recognition via Domain Alignment Prototypical Networks
>
> Cross-domain Few-shot Classification via Learned Feature-wise Transformation

CloserLookFewShot

- ☑ (ICLR 2019) [paperswithcode]
    - Chen et al. "A Closer Look at Few-shot Classification"

核心在哪?	精读? 代码?	关键词?	亮点?	笔记时间?

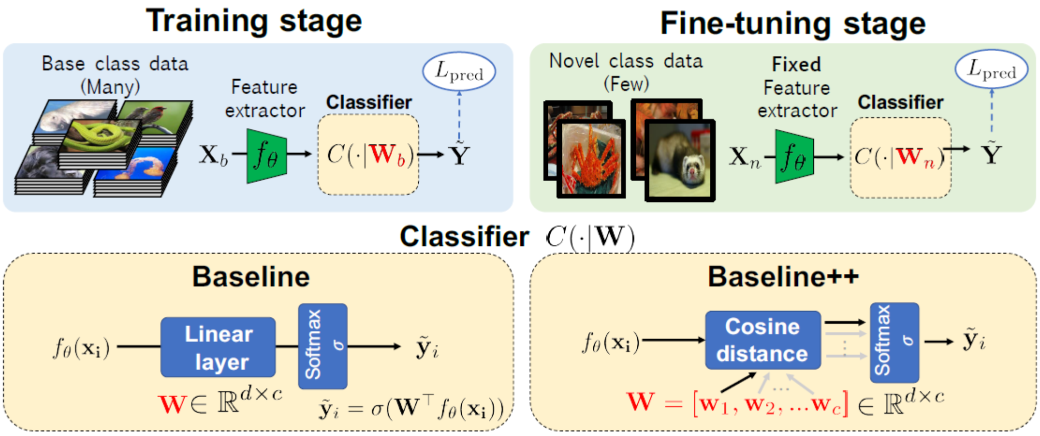
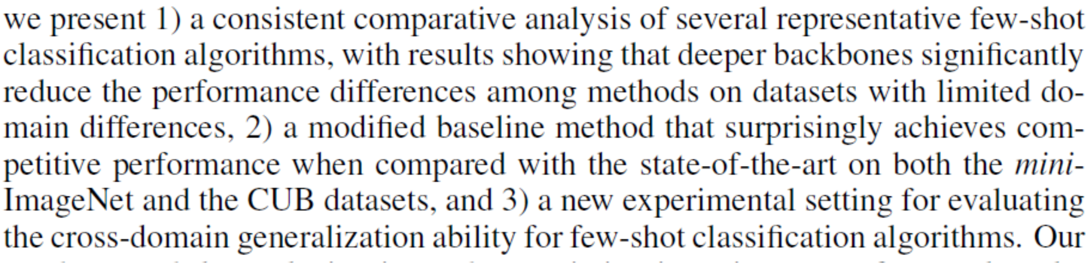
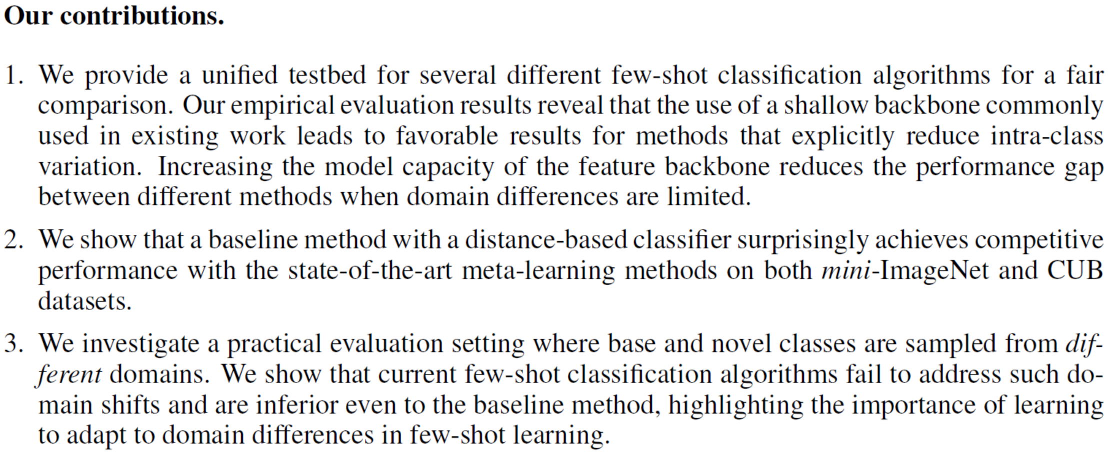


Figure 1: **Baseline and Baseline++ few-shot classification methods.** Both the baseline and baseline++ method train a feature extractor f_θ and classifier $C(.|\mathbf{W}_b)$ with base class data in the training stage In the fine-tuning stage, we fix the network parameters θ in the feature extractor f_θ and train a new classifier $C(.|\mathbf{W}_n)$ with the given labeled examples in novel classes. The baseline++ method differs from the baseline model in the use of cosine distances between the input feature and the weight vector for each class that aims to reduce intra-class variations.



we present 1) a consistent comparative analysis of several representative few-shot classification algorithms, with results showing that deeper backbones significantly reduce the performance differences among methods on datasets with limited domain differences, 2) a modified baseline method that surprisingly achieves competitive performance when compared with the state-of-the-art on both the *mini*-ImageNet and the CUB datasets, and 3) a new experimental setting for evaluating the cross-domain generalization ability for few-shot classification algorithms. Our



Our contributions.

1. We provide a unified testbed for several different few-shot classification algorithms for a fair comparison. Our empirical evaluation results reveal that the use of a shallow backbone commonly used in existing work leads to favorable results for methods that explicitly reduce intra-class variation. Increasing the model capacity of the feature backbone reduces the performance gap between different methods when domain differences are limited.
2. We show that a baseline method with a distance-based classifier surprisingly achieves competitive performance with the state-of-the-art meta-learning methods on both *mini*-ImageNet and CUB datasets.
3. We investigate a practical evaluation setting where base and novel classes are sampled from *different* domains. We show that current few-shot classification algorithms fail to address such domain shifts and are inferior even to the baseline method, highlighting the importance of learning to adapt to domain differences in few-shot learning.

reducing intra-class variation 是重要的: when the feature backbone is shallow.

- **背景? 提出了什么问题?**

  在一个更加公平的基础上来比较各个Few-Shot Learning算法之间的优劣.

  在研究 复现和实验现有算法过程中, 总结出的一份实验报告.

- **为了解决此问题提出了什么具体的idea?**
  - 提出 consistent **comparative experiments** to compare several representative few-shot classification methods on common ground.

    在设置基类和新类的有限域差异时, 使用一个deep backbone可以缩小不同方法之间的性能差距.
  - 在distance-based classifier中替换线性分类器, 可以达到SotA.
  - 提出了一个practical evaluation setting. 其中存在基类和新类之间的领域转换.

- **如何从该idea形式化地对问题建模、简化并解决的?**

  较多的base class的标记数据: $\mathbf{X}_b$, 较少的novel class标记数据 $\mathbf{X}_n$.

在训练过程中 $\mathbf{X}_n$ 是unseen的.

- 训练过程: 交叉熵损失 分类器 $C(.\mid\mathbf{W}_b)$ 由 $\mathbf{W}_b^\top f_\theta(\mathbf{x}_i)$ 的线性层, 然后是softmax函数.
- Fine-tuning 过程: adapt 模型去识别新类, 固定feature extractor f_θ 的参数, 使用交叉熵损失训练一个新的分类器 $C(.\mid\mathbf{W}_n)$.

基于baseline的提升:

在训练时 explicitly reduces intra-class variation among features.

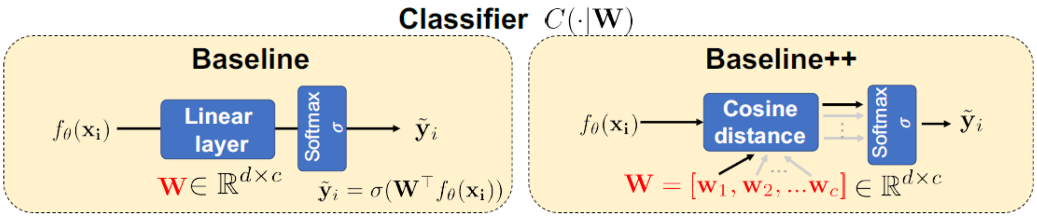

Classifier $C(\cdot|\mathbf{W})$
Baseline
$f_\theta(\mathbf{x}_i)$
Linear layer
Softmax
σ
$\tilde{\mathbf{y}}_i$
$\mathbf{W} \in \mathbb{R}^{d\times c}$
$\tilde{\mathbf{y}}_i = \sigma(\mathbf{W}^\top f_\theta(\mathbf{x}_i))$
Baseline++
$f_\theta(\mathbf{x}_i)$
Cosine distance
Softmax
σ
$\tilde{\mathbf{y}}_i$
$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ...\mathbf{w}_c] \in \mathbb{R}^{d\times c}$

Figure 1: **Baseline and Baseline++ few-shot classification methods.** Both the baseline and baseline++ method train a feature extractor f_θ and classifier $C(.|\mathbf{W}_b)$ with base class data in the training stage In the fine-tuning stage, we fix the network parameters θ in the feature extractor f_θ and train a new classifier $C(.|\mathbf{W}_n)$ with the given labeled examples in novel classes. The baseline++ method differs from the baseline model in the use of cosine distances between the input feature and the weight vector for each class that aims to reduce intra-class variations.

- 理论方面证明的定理与推导过程?
- 这个任务/解决方法有什么意义?
- 对论文的讨论/感想?

MainTitle

- ☐ (ICLR 2020) [paperswithcode]
  - Tseng et al. "Cross-Domain Few-Shot Classification via Learned Feature-Wise Transformation"

核心在哪?	精读? 代码?	关键词?	亮点?	笔记时间?

- 背景? 提出了什么问题?
- 为了解决此问题提出了什么具体的**idea**?
- 如何从该**idea**形式化地对问题建模、简化并解决的?
- 理论方面证明的定理与推导过程?
- 这个任务/解决方法有什么意义?
- 对论文的讨论/感想?

About Attention/GNN

- 近期:

  Attentive Graph Neural Networks for Few-Shot Learning

Few-Shot Learning With Attention-Weighted Graph Convolutional Networks For Hyperspectral Image Classification

- 相关:

  A Dual Attention Network with Semantic Embedding for Few-shot Learning

  Multi-attention Meta Learning for Few-shot Fine-grained Image Recognition

  Attentive Weights Generation for Few Shot Learning via Information Maximization

- 高引:

  Dynamic Few-Shot Visual Learning Without Forgetting

  Few-Shot Learning with Graph Neural Networks

  TADAM

  Cross Attention Network for Few-shot Classification

  Few-Shot Semantic Segmentation with Prototype Learning

FewShotWithoutForgetting

☑ (CVPR 2018) [paperswithcode]

  - Gidaris et al. "Dynamic Few-Shot Visual Learning without Forgetting"

核心在哪?	精读? 代码?	关键词?	亮点?	笔记时间?

基于外部记忆的. base类和novel类之间是没有相同的. weight generator还把base weight作为输入. 先训练出feature extractor固定, 再训练后面的weight generator.

注意力在注意啥? 代码中如何拆解为一般形式并理解?

- **背景? 提出了什么问题?**

  需要做到:

  (a) the learning of the **novel categories needs to be fast**, 在新类上的学习要快.

  (b) to not **sacrifice** any recognition accuracy on **the initial categories** that the ConvNet was trained on, i.e., to not “forget” (from now on we will refer to those initial categories by calling them base categories). 原来的类上的性能不能降.

  not only is able to **recognize these base categories**, but also learns to **dynamically** recognize novel categories from only a few training examples (provided only at test time) while also not forgetting the base ones or requiring to be re-trained on them (dynamic few-shot learning without forgetting).

  不会忘记base class, 也不需要re-trained.

- 为了解决此问题提出了什么具体的**idea?**
    - **Few-shot classification-weight generator based on attention.**

      the first technical novelty of our work is that we enhance a typical object recognition system with an extra component, called **few-shot classification weight generator**:

        1. input: **a few training examples** of a novel category e.g., no more than five examples) and,
        2. based on them, generates **a classification weight vector** for that novel category.
            - acquired **past knowledge** about the visual world:

              通过 incorporating an attention mechanism **over the classification weight vectors** of the base categories.

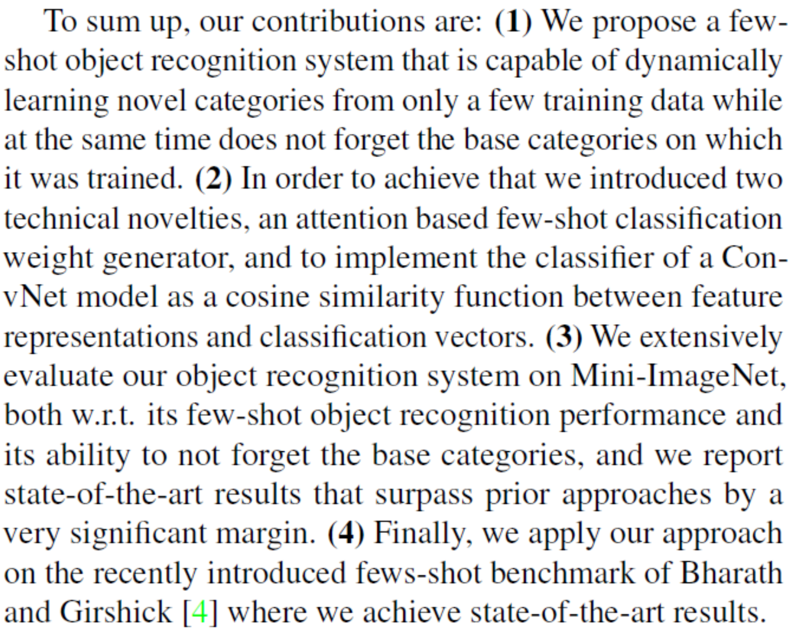
              注意力机制作用在base categories上.

    - **Cosine-similarity based ConvNet recognition model.**

      ConvNet 需要可以处理基类和新类的classification weight vectors.

      问题是: However, as we will explain in the methodology, this is **not** feasible with the **typical dot-product based classifier** (i.e., the last linear layer of a classification neural network).

      implement the classifier as a **cosine similarity function** between **the feature representations and the classification weight vectors.**



To sum up, our contributions are: **(1)** We propose a few-shot object recognition system that is capable of dynamically learning novel categories from only a few training data while at the same time does not forget the base categories on which it was trained. **(2)** In order to achieve that we introduced two technical novelties, an attention based few-shot classification weight generator, and to implement the classifier of a ConvNet model as a cosine similarity function between feature representations and classification vectors. **(3)** We extensively evaluate our object recognition system on Mini-ImageNet, both w.r.t. its few-shot object recognition performance and its ability to not forget the base categories, and we report state-of-the-art results that surpass prior approaches by a very significant margin. **(4)** Finally, we apply our approach on the recently introduced fews-shot benchmark of Bharath and Girshick [4] where we achieve state-of-the-art results.

- 如何从该**idea**形式化地对问题建模、简化并解决的?
    - **3.Methodology**

      base categories: K_{base}

      第 b 类的训练样本数量: N_b, 第 b 类的第 i 个: $x_{b,i}$.

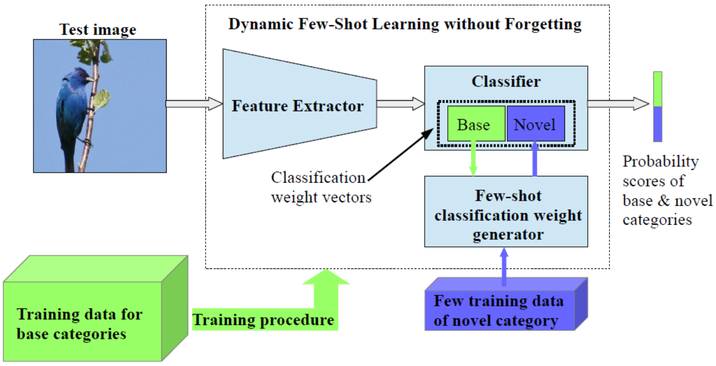

Test image
Dynamic Few-Shot Learning without Forgetting
Feature Extractor
Classifier
Base
Novel
Classification weight vectors
Few-shot classification weight generator
Probability scores of base & novel categories
Training data for base categories
Training procedure
Few training data of novel category

Figure 1: Overview of our system. It consists of: (a) a *ConvNet based recognition model* (that includes a feature extractor and a classifier) and (b) a *few-shot classification weight generator*. Both are trained on a set of base categories for which we have available a large set of training data. During test time, the weight generator gets as input a few training data of a novel category and the classification weight vectors of base categories (green rectangle inside the classifier box) and generates a classification weight vector for this novel category (blue rectangle inside the classifier box). This allows the ConvNet to recognize both base and novel categories.

接下来，如上两个部分：a ConvNet-based recognition model that is able to recognize both base and novel categories, a few-shot classification weight generator: dynamically generates **classification weight vectors** for the novel categories at test time.

- **a ConvNet-based recognition model**:

  feature extractor: $F(.\mid\theta), z=F(x\mid\theta)\in\mathbb{R}^{d}$.

  classifier: $C(.\mid W^{*}), p=C\left(z\mid W^{*}\right)$, **the probability classification scores p.**

  这个阶段学到了：

  - classification weight vectors: W^{*}
  - θ

- **Few-shot classification weight generator**:

$$w_{n}^{\prime}=G\left(Z_{n}^{\prime}, W_{base}\mid\phi\right)$$

- **3.1.Cosine-similarity based recognition model**

  计算第k类的score：

$$s_{k}=z^{\top}w_{k}^{*}$$

  请注意：However, **the mechanisms involved during learning those classification weights w_{k}^{*} are very different.**

  - 对于base类的权值，学习过程是缓慢的.
  - 对于novel类的权值，学习是快速的，因为基于输入特征.

  Due to those differences, the weight values in those two cases (i.e., base and novel classification weights) **can be completely different,** and so the same applies to the raw classification scores computed with the dot-product operation, which can **thus have totally different magnitudes** depending on whether they come **from the base or the novel categories.**

  base 类权重一样训练得到，novel 类的权重是weight generator的输出. 这样分类器就同时具备对基础类别和新类别进行分类的能力.

  因为这两种方式得到的权重有较大差异(**量级问题**)，传统的点积方法计算相似度会造成较大误差：

  所以修改 分类器$C(.\mid W^{*})$为计算cos相似度：

$$s_{k}=\tau\cdot\cos\left(z, w_{k}^{*}\right)=\tau\cdot\bar{z}^{\mathrm{T}}\bar{w}_{k}^{*}$$

where $\bar{z} = \frac{z}{\|z\|}$ and $\bar{w}_k^* = \frac{w_k^*}{\|w_k^*\|}$ are the l_2 -normalized vectors (from now on we will use the overline symbol $\bar{z}$ to indicate that a vector z is l_2 normalized , and τ is a learnable scalar value. 消除了基础类别权重和新类别权重在数值上的量级差异.

Advantages of cosine-similarity based classifier, 请再看原文，作者为此还做了一个可视化.

- **3.2.Few-shot classification weight generator**

  $w' = G(Z', W_{base} \mid \phi)$ classification weight generator 的构建方式:

  1 Feature averaging based ($\bar{z}_i'$ **是特征向量)**

  $$w' = \phi_{avg} \odot w'_{avg} = \phi_{avg} \odot \left(\frac{1}{N'} \sum_{i=1}^{N'} \bar{z}_i'\right)$$

  2 Attention-based

  上述平均的方法不好，特别是新类别只包含一个样本(one-shot情形)时不能充分利用特征提取网络的信息.

  $$w'_{att} = \frac{1}{N'} \sum_{i=1}^{N'} \sum_{b=1}^{K_{base}} Att\left(\phi_q \bar{z}_i', k_b\right) \cdot \bar{w}_b$$

  $$w' = \phi_{avg} \odot w'_{avg} + \phi_{att} \odot w'_{att}$$

- 理论方面证明的定理与推导过程?
- 这个任务/解决方法有什么意义?
- 对论文的讨论/感想?

CAN

☑ (NeurIPS 2019) [paperswithcode]

- Hou et al. "Cross Attention Network for Few-shot Classification"

核心在哪?	精读? 代码?	关键词?	亮点?	笔记时间?

---

we propose a novel Cross Attention Network to address the challenging problems in few-shot classification.

Firstly, Cross Attention Module is introduced to deal with the problem **of unseen classes.** The module generates **cross attention maps** for each pair of class feature and query sample feature so as to highlight the target object regions, making **the extracted feature more discriminative.**

Secondly, a transductive inference algorithm is proposed to **alleviate the low-data problem**, which iteratively utilizes the unlabeled query set to augment the support set, thereby making the class features more representative.

- 背景？提出了什么问题？

  难点(通用的)：

  the test images in the support/query set are **from unseen classes, thus their features can hardly** attend to the target objects.

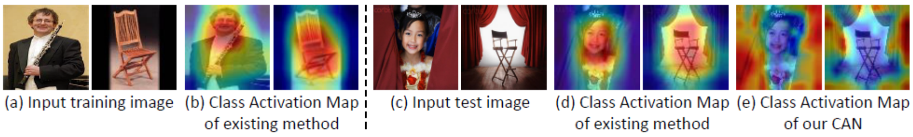
  
  (a) Input training image
  (b) Class Activation Map of existing method
  (c) Input test image
  (d) Class Activation Map of existing method
  (e) Class Activation Map of our CAN

  Figure 1. An example of the class activation maps [44] of training and test images of existing method [32] and our method CAN. Warmer color with higher value.

  the extracted features only capture the information of the objects that are **related to the training classes**.

  For another, the low-data problem makes the feature of each test class **not representative** for the true class distribution, as it is obtained from very few labeled support samples.

  enhance the feature discriminability for few-shot classification.

  To recognize a sample from unseen class given a few labeled samples, human tends to **firstly locate the most relevant** regions in the pair of labeled and unlabeled samples.

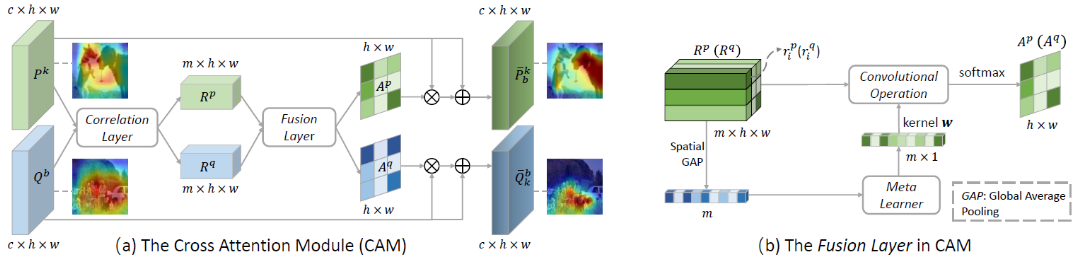
  Transductive Algorithm 未看。

- 为了解决此问题提出了什么具体的**idea**?

- 如何从该**idea**形式化地对问题建模、简化并解决的？

  metric-learning to obtain proper feature representations

  we propose Cross Attention Module (CAM): model **the semantic relevance between the class feature and query feature**, thus draw attention to the target objects and benefit the subsequent matching.

  (a) The Cross Attention Module (CAM)

  (b) The *Fusion Layer* in CAM

  Figure 2. (a) Cross Attention Module (CAM). (b) the *Fusion Layer* in CAM. In the figure, R^p $(R^q) \in \mathbb{R}^{m \times m}$ is reshaped to $\mathbb{R}^{m \times h \times w}$ for a better visualization. As seen, CAM can generate the feature maps that attend to the regions of target object (*coated retriever* in the figure).

  P^k 代表第k类的特征图，Q^b 代表一个query样本的特征图。

  Correlation Layer 计算P, Q之间的correlation map。对应点利用cos距离计算相关性。

- 理论方面证明的定理与推导过程？
- 这个任务/解决方法有什么意义？
- 对论文的讨论/感想？

the episode training mechanism: The episodes used in training **simulate the settings in test.** the settings in test. Each episode is formed by **randomly sampling C classes and K labeled samples per class** as the support set $\mathcal{S}=\{(x_a^s,y_a^s)\}_{a=1}^{n_s}\,(n_s=C\times K)$, and a fraction of the rest samples from the C classes as the query set $\mathcal{Q}=\left\{\left(x_b^q,y_b^q\right)\right\}_{b=1}^{n_q}$. And we denote $\mathcal{S}^k$ as the support subset of the k^{th} class. **How to represent each support class $\mathcal{S}^k$ and query sample x_b^q and measure the similarity between them** is a key issue for few-shot classification.