

Prototypical networks for few-shot learning 论文复现

王云舒

摘要

小样本学习是一类测试集与训练集分布不同或类别不同，只提供很少的与测试集数据分布相同的数据并进行分类的问题。传统的基于梯度优化的深度学习方法难以处理这类小样本问题，即使用微调方法调整网络的分类层也很难达到较高的准确率。原型网络是一个重要的小样本方法，通过用神经网络学习输入图像到嵌入空间的非线性映射，并将类的“原型”作为嵌入空间中支撑集的均值，查找最近的类原型并对嵌入空间的查询点进行分类，其思想直观并且有较好的分类表现。本次复现工作选择该文章，完成了代码重构和实现并补充了 t-SNE 的可视化展示。技术改进部分分别改变了特征提取器的骨架、训练方式和用注意力获取原型方式；应用创新部分对目前自己手势识别工作的跨环境和跨标签问题进行了具体的小样本方法应用尝试。结果表明，复现和改进的方法适用于图像小样本的跨分布域工作，有很大的应用潜力。

关键词：小样本学习；原型网络；元训练；微调；交叉注意力机制

1 引言

本次论文复现选择的题目为用于小样本学习(Few-shot learning)的原型网络，是一篇来自 NIPS2017 的顶级文章，目前引用量已高达 5357 次。小样本问题是机器学习以及深度学习目前研究热门的问题，相比传统的数据集划分为训练集、测试集，数据分布相同的情况，小样本问题的测试集与训练集的分布不同，可能是不同分布域或不同的识别任务标签，且测试集分布域的数据只有很少量的数据可以预先知道并辅助训练或测试。这样导致模型训练时可能准确率较高，测试时准确率很低或者完全无法测试。

从理论上来说，选题工作针对小样本问题下由于新类样本很少，用新数据重新训练模型会过拟合，即经验风险最小化不可靠^[1]的问题提出了一种解决方案。Ravi 等人^[2]研究了在少量数据下，基于梯度的优化算法失败的原因，因为梯度优化算法如 momentum, adagrad, ADAM 等无法在几步内完成优化，多种超参的选取无法保证收敛的速度。其次，不同任务分别随机初始化会影响任务收敛好的解上。虽然 Finetune^[3]这种迁移学习能够缓解这个问题，但当新数据相对原始数据偏差比较大时，迁移学习的性能会大大下降。关于小样本问题的问题定义：

目的：在每个类别仅仅只有少数训练样本的时候学习出一个分类器。

符号定义：

- T : 小样本分类任务，其包含支撑集 S 和查询集 Q ;
- S : 支撑集——带标签的数据集合;
- Q : 查询集——需要学习到的分类器进行评估的无标签数据;
- K : 带有标签的数据个数;
- $N-way K-shot$: 有 N 个类别，每个类别有 K 个样本，总共 $N \times K$ 个样本，根据这些样本得到一个分类器，该分类器能够对剩余的类别甚至类别进行精准识别。

这里的“模型”至少应该包含两个部分：特征提取部分负责将输入的内容转换为一个向量化的嵌入，例如对输入的 84×84 图像，通过特征提取器最后会得到一个 1×512 的向量表示；分类部分负责将嵌入向量按照一定的相似度关系，如欧氏距离、线性分类器等，将向量按照一种距离度量找到最近的有标签向量，并为输入内容分类。本次复现的原型网络即由这两部分组成，特征提取部分由一个 4 层 block 的卷积神经网络（CNN）组成，每一个 block 有一个卷积层、一个归一化层和一个池化层。分类部分直接对嵌入向量进行欧式距离的匹配，没有引入参数。相关工作及优化方案会从这两个方向分别入手进行改进。

从现实应用上来说，以目前我们物联网工作的手势识别为例，基于声波感知的手势识别主要依靠智能设备的扬声器发出 19kHz 的超声波，麦克风接收到的超声波主要分为三个部分：扬声器直接传入的声波、扬声器发出声波后，被手部反射而传入麦克风的声波以及周围环境如墙体反射传入麦克风的声波。由于扬声器波源频率和手指运动速度之间的关系，多普勒效应可以求得 Δf 约为 $20\text{--}300\text{Hz}$ 。之后先通过一个带通滤波器，再通过一个带阻滤波器，经过短时傅里叶变换和 spectrogram 函数绘制光谱图就可以得到表征手势的特征图，如图 1 所示。

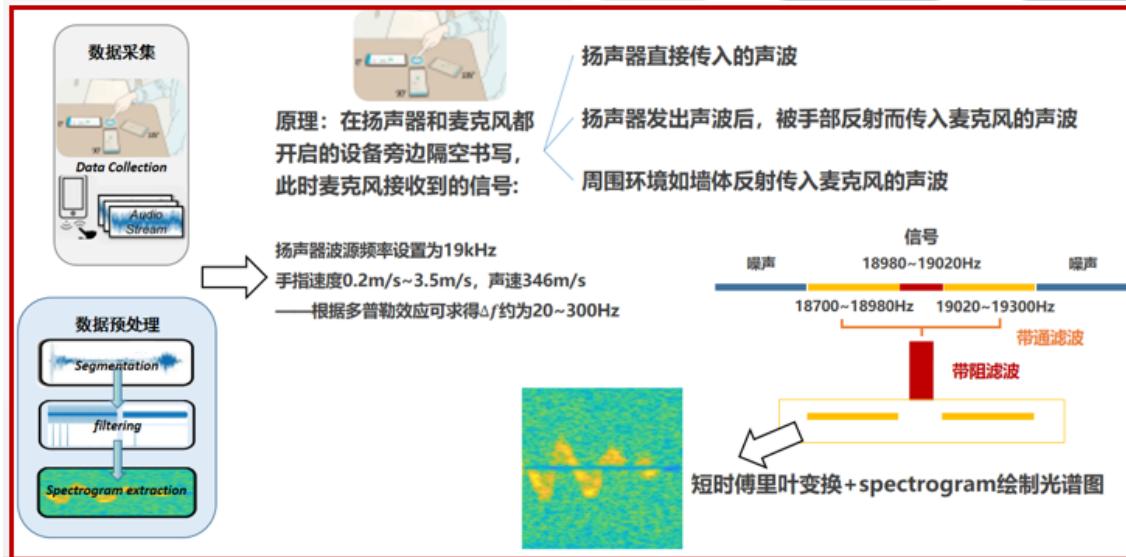


图 1: 基于声波感知的手势识别原理

小样本学习在这个应用中有非常大的两个应用价值：跨类别识别和跨分布识别。跨类别是小样本学习的基础目的。传统深度学习假设我们采集了 0-9 字符每个字符 10000 条，我们可以用每个字符 8000 条数据抽取做训练集，剩下的 2000 条数据做测试集，训练一个神经网络；模型最终能够对一个输入的字符图像 6，输出它的标签——6。跨类别小样本问题是指：如果我们想要识别的是 A-Z 呢？我们仍有 0-9 每个字符 10000 条，但是 A-Z 的字符只有 5 条数据。而最终我们想输入 A-Z 的字符图像，输出对应的 A-Z 标签。为了适应识别内容的改变，我们可以微调调整全连接层等靠近分类的层。但是这种训练方法有点类似“事后补救”，因为前面特征提取部分传入的每个 batch 梯度是随机的，这种训练方式会让模型最后只专注于原来已有的类别，如果新类别和已有类别差别大一些就会导致微调效果较差。如图 2 所示，从 h_1 到 h^* 的优化过程，数据充足且分布相同时这一步可以保证梯度收敛，如果数据不充足且分布不相同就会导致经验化风险最小化不可靠。上述的字符识别例子通过这种简单的微调方式，本文后续也给出了实验数据，效果的确不够理想。

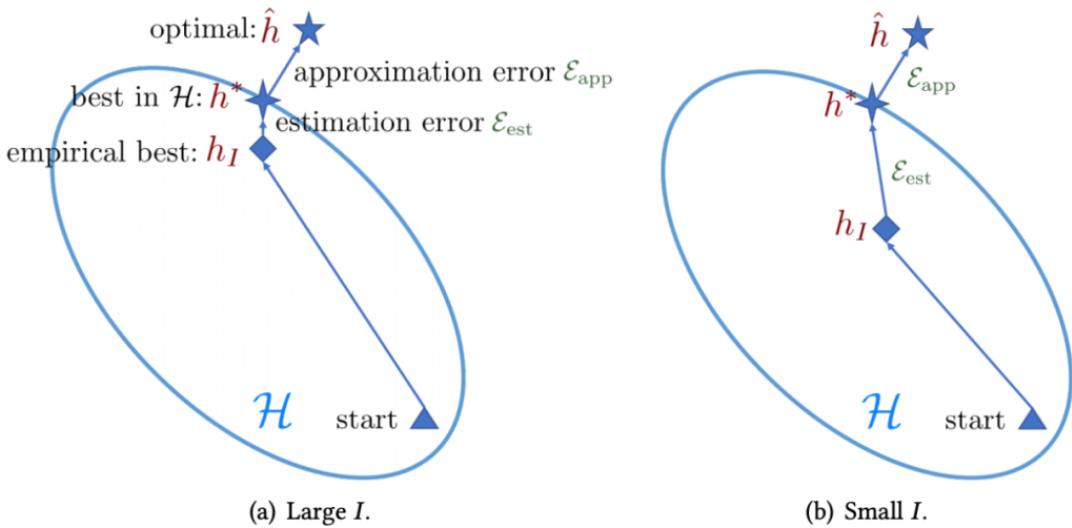


图 2: 数据充足时模型训练和数据不足时模型训练的区别^[4]

如果应用小样本方法，它的核心思想是能训练出一个对所有数据有效的特征提取器，而非只关注原有的固定类别；将数据提取特征后得到特征向量，直接比较特征向量之间的相似度也会比直接进行 softmax 更不容易过拟合。通过训练和测试过程中引入支撑集、查询集的概念，每次不是像传统 batch 那样随机抽取，而是根据这一批类别抽取对应数目的 episode，每个 episode 是先选取一部分类别再每个类别抽取固定的训练样本数目。这种训练方式可以让梯度下降的状态符合我们预期的要求，又能“推开”类间距离，又能“拉进”类内距离。更多具体的分析可以参考 3.1.2 原型网络方法分析章节。

小样本学习不仅可以解决跨识别类别的问题，对于物联网领域的跨环境问题也有非常好的表现。跨识别类别和跨环境本质上都是跨分布问题，我们手势识别工作面对的一个非常大的问题就是数据训练时的环境和用户使用时的环境不相同，例如训练时使用三星平板，在实验室环境训练，测试时使用小米手机，在室外环境测试。这种跨域问题如果使用传统深度学习方法，准确率非常低，0-9 数字 4000 条数据训练仅有 0.41 准确率；而使用小样本方法后，2-shot 情况下可以达到 0.85 的准确率。

选择小样本学习中最具有代表性的原型网络工作，复现这篇文章，可以帮助我们更好的了解 few-shot 的基本思想，有助于理解其他在此基础上改进的工作并将其应用到物联网的工作中；应用了小样本学习方法可以帮助我们采更少的数据，适应多种域和不同的书写任务。事实上，经过复现工作，我发现使用 SimpleCNAPs^[5]这一改进的小样本方法，A-Z 的识别准确率几乎能达到不跨类别的识别准确率；甚至对其他的类如书写星形、三角等其他类，只要提供少量样本都可以达到很高的准确率。复现这篇文章可以把它的工作、模型与架构应用到我们的数据集上，以及做出提升和改进，助力物联网智能感知与移动计算的发展。

2 相关工作

2.1 原型网络部分的相关工作

原型网络工作为 2017 年提出，该方法于当时 miniImagenet 数据集^[6]下在 5way-1shot 与 5way-5shot 均达到 SOTA (State-of-the-art)，后来又有许多小样本方法对原型网络结构进行了非常多的优化与改进。复现工作中另一种应用到手势识别中的网络 SimpleCNAPs^[5]是 2020 年 CVPR 达到 SOTA 的方法，它主要针对特征提取器部分进行了“预训练 + 元训练”两个过程的训练。过去十几年的大部分学习工

作中，小样本学习可以沿着两个主要的方向进行区分：

- 1) 如何将图像转换为向量化的嵌入（embedding）；
- 2) 如何计算向量之间的距离（distance）以分配标签。

与原型网络相关的小样本工作主要可以按照图 3 进行分类：

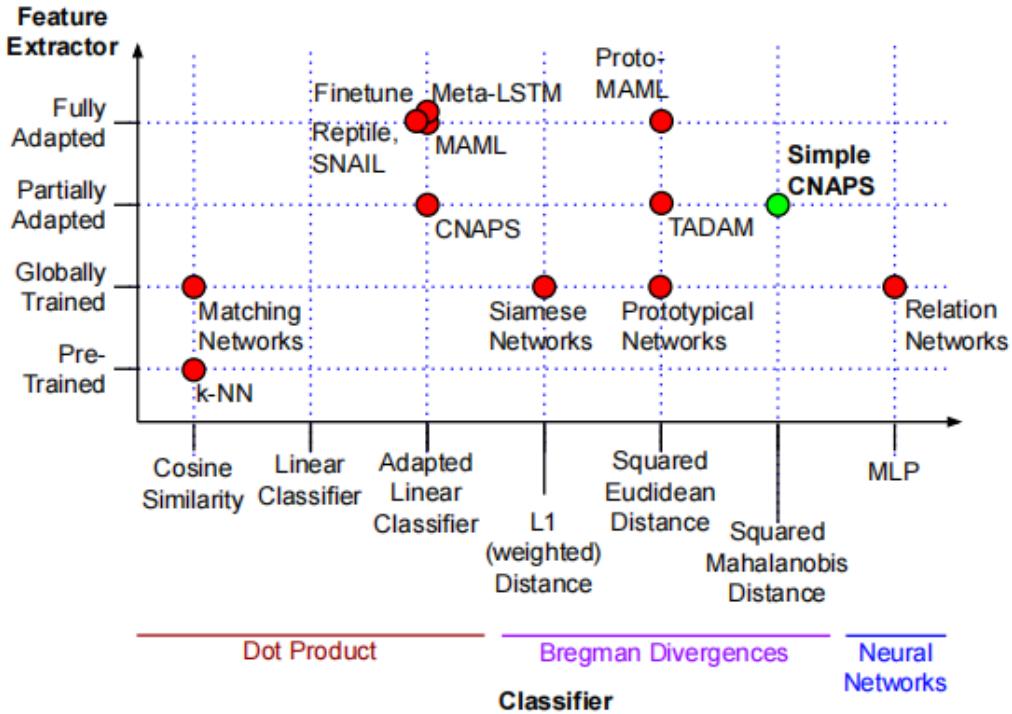


图 3：与原型网络相关的小样本工作^[5]

图 3 中，纵轴代表特征提取器的工作模式，Pre-Trained 代表特征提取器已经预先用其他数据训练好或完全不用训练，可以直接使用，典型代表为 k-NN 方法^[7]；Globally Trained 代表用于训练的数据全部直接进行小样本支撑集——查询集的训练，对特征提取网络（也称嵌入网络）进行全局的训练，原型网络就是这种方式；Partially Adapted 代表特征提取器的训练有预训练和元训练两个部分，可能会将用于训练的数据划分成两部分，一部分预训练一个特征提取器，另一部分元训练来改变这个特征提取器的部分参数，如 CNAPs 的 FiLM 层^[8]，它只去线性改变预训练特征提取器得到的每个特征图。也可能用一个用其他数据预训练好的特征提取器，用于训练的该数据集数据全部去用在元训练上；Fully Adapted 相比 Partially Adapted 方式，会改变预训练后的所有网络参数，希望网络能够快速适应其他各种任务，典型代表为 MAML^[9]方法。横轴代表最后分类器的分类方式，Dot Product 指通过点积寻找最后的特征向量相似度，包括余弦相似度、线性分类等；Bregman Divergences^[10]指 Bregman 散度下，L1 距离度量、平方欧式距离度量、平方马氏距离度量等；也可以用一个神经网络去度量相似性，如使用 MLP 的 Relation Network 工作^[11]。

其他的一些重点的方法和工作，Siamese networks^[12]是一种早期的小样本学习方法，它使用共享特征提取器为支撑图像和查询图像生成嵌入，然后通过选择支撑集和查询集嵌入之间最小的加权 L1 距离进行分类。Matching networks^[13]学习不同的特征提取器来适应支撑集和查询集图像，然后用余弦相似度进行分类。MAML^[9]及其扩展^[2,14]通过学习元参数，使特征提取器能快速适应新任务。CNAPS^[8]是一种基于条件神经过程 CNP^[15]的小样本自适应分类器，对预先训练好的特征提取器用最早由 CV 领

域设计的 FiLM 层^[16]对特征图进行线性调制，参数通过元训练能够更好的适应新任务。它的分类器使用线性分类层；相比之下，SimpleCNAPs^[5]使用了平方马氏距离，它的分类部分思想与原型网络一致，只是原型网络使用的是平方欧式距离，两种距离度量在不同的数据集类型下表现略有不同。

事实上，这样阐述相关工作只是重在了解小样本方法当前的进展。原型网络提出的时间较早，在当时甚至还没有 Partially Adapted 的较好的方法，原型网络在当时就已经是非常先进的网络架构了。后面的诸多工作基本思想和原型网络的“Prototypical”，即原型的思想都是一致的，了解并实现原型网络的架构才能更好的了解改进和优化的方法，做到融会贯通。复现工作的一种改进交叉注意力技术改进利用了当前非常热门的注意力机制，与其相关的还有 Transformer 架构^[17]和各种结合了其思想的各种网络，这些工作不是直接与小样本学习相关的方法，其思想会在复现部分的技术改进相应处予以介绍。

2.2 涉及到手势识别应用的相关工作

人类的手势和手部动作是用于交互的最主要方式，基于声波感知的识别技术中利用多普勒效应带来的频移是最普遍和最直接的方法。以人在智能设备如手机旁边隔空书写为例，手机扬声器发出调制过的超声波，由于人书写带来的多普勒频移效应，手机麦克风接收到的声波会发生频率改变，据此可以生成各手势书写的特征。这类系统通常包括数据采集和预处理、短时傅里叶变换、特征提取、分类识别等步骤。此外，除了多普勒频偏以外，还有基于 FMCW 和基于 CIR 以及结合目标与声源之间距离等其他方式。早期 Kalgaonkar 与 Raj 的工作^[18]（图 4(a)）设计了一种由三个麦克风和一个扬声器来处理信号的设备，基于多普勒频移原理实现了手掌级别的动作识别；之后的工作 Soundwave^[19]（图 4(b)）探索了使用商业的现成设备中的音频组件来识别手掌的隔空书写手势。Qifan 等^[20]（图 4(c)）设计的 Dolphin 系统，利用内置的扬声器和麦克风发射和接收连续的 21 kHz 超声波信号，提取多普勒效应相关的频域特征，采用机器学习模型实现了多达 17 种凌空手势的识别；SoundWrite^[21]利用振幅谱密度和一些其他的声学特征如 MFCC 描述手写特征，使用 KNN 将捕获的特征与数据库中的标记特征进行匹配。Wang^[22]的工作提出了一种动态速度扭曲（Dynamic speed warping, DSW）算法，基于观察手势类型是由手部组件的轨迹而非运动速度，通过动态缩放速度分布并跟踪轨迹的移动距离，其可以匹配来自不同域的具有十倍速度差异的手势信号。

关于最近的一些新研究进展，手势识别追求更细粒度、更高的准确率、更小的所需训练集规模以及更多的目标。随着迁移学习^[23]、小样本学习^[4]以及生成对抗网络^[24]等技术的发展，这些技术也已经应用到手势识别领域中。工作^[24]将基于迁移学习的卷积神经网络用于手势识别，在 sign language digits and Thomas Moeslund’s gesture recognition datasets 上测试其准确率比现有工作要好；工作^[25]通过肌电图进行小样本学习手势识别，虽然与基于声波感知有所不同，但其方法很容易复用到声波感知得到的频谱图上；工作^[26]应用生成对抗网络 GAN，开发了一个场景转移网络，不仅利用来自当前场景的真实样本，还利用来自另一个可用场景的真实样本来生成虚拟样本，对一个基于毫米波的数据与测试平台进行小样本数据集的训练与测试。虽然载体和方法不同，上面的这些方法对基于声波感知的手势识别也存在一些启发。此外，Ultragesture^[27]（图 4(d)）基于信道脉冲响应（CIR），CIR 测量可以提供 7mm 的分辨率，足以识别轻微的手指运动。其将 CIR 测量值封装到图像中，准确率优于基于多普勒效应的方案，并且其可以在大多数移动设备上已经存在的商用扬声器和麦克风上运行，无需任何硬件修改；

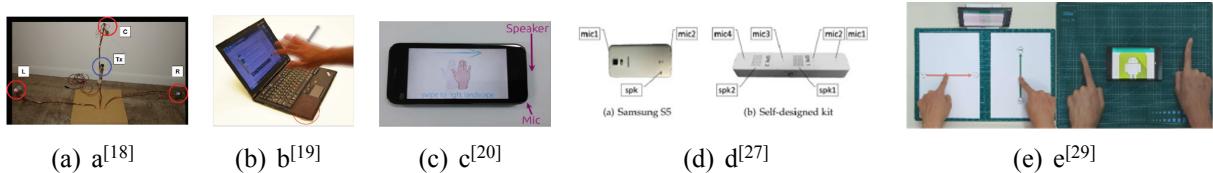


图 4: 基于声波感知的手势 & 手部动作识别应用

RobuCIR^[28]采用跳频机制减轻频率选择性衰落避免信号干扰，这一高精度 CIR 工作可以识别 15 种手势；AMT^[29]（图 4(e)）定义了初级回波的新概念以更好的代表目标运动，通过使用多个扬声器-麦克风对执行动作的多点定位，检测主回波并滤除次要回波，消除目标凸起多径效应而非将其草草假设为粒子，提高了跟踪精度。

3 本文方法

3.1 原文方法概述

3.1.1 问题描述

关于小样本问题的描述在引言部分已经详细给出，如目的、符号定义、模型具体概念等，这里再补充介绍小样本方法：

- 原理：训练得到一个分类器，再只使用任务的支撑集就可以为每个查询样本分配标签；
- 传统微调方法的缺点：分布不同的数据集无法训练出一个能够完全反映类间和类内关系的模型，以致最终的分类效果不明显；
- 解决方法：在数据集上通过提取可迁移的知识，该知识能够在支撑集执行更好的小样本学习，从而成功地对查询集进行分类；
- 情景训练：
 - $N-way K-shot$ 场景： T 为待查询样本
 - $S = \{(x_i, y_j)\}_{i=1}^{N \times K}$
 - $Q = \{(x_i, y_i)\}_{i=N \times K+1}^{N \times K+T}$
 - $x_i, y_i \in \{C_1, \dots, C_N\} = C_T \subset C$ 分别是第 i 个输入数据及其标签
 - C 是训练数据集或测试数据集的所有类别集合
 - 在每一个 *episode* 中的支撑集用作带标签的数据集，并且模型根据此数据集进行训练。损失的优化是最小化在查询域上的预测损失，逐步执行这个训练过程，直到结果收敛。
- 说明：如果上述 $N \times K$ 个支撑集样本有些是无标签的，那么此任务成为半监督小样本分类。

以本次复现内容的主要数据集 miniImagenet 为例，图 5 展示了支撑集中有 3 个类别狮子、大象和狗，每个类别有两个支撑样例，这个场景即为 3-way 2-shot 场景；这些图像先经过原型网络的特征提取器（即卷积部分）得到各自的向量，相同类别的向量直接加和求平均作为该类别的原型；待查询的图像同样经过原型网络的卷积部分得到它的特征向量，比较该向量和各原型的欧式距离，可以看出该向量与表示狮子的原型向量距离最小，该查询图像的类别就应是狮子。对于更一般化的 $N-way K-shot$ 问题，实际上 N 和 K 都针对的是测试阶段的类别数和提供的样本数，但训练时一个 *episode* 有多少类别、每个类别的支撑集能提供多少样本也可称作 way 和 shot。

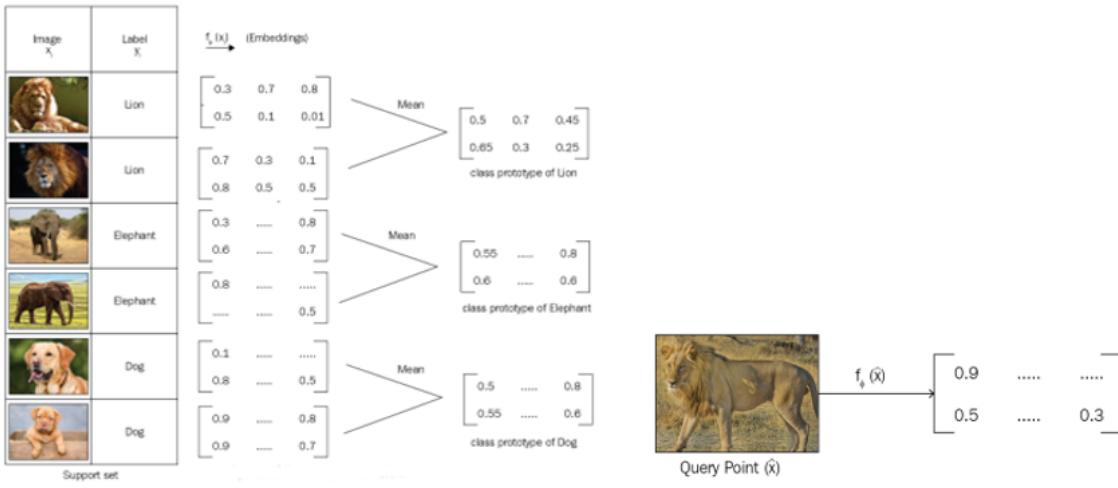


图 5: 原型网络的一个具体示例及分类过程

3.1.2 方法分析

从小样本学习到具体的原型网络方法：调整分类器以适应训练中未见的新任务。原型网络（Prototypical Network）的基本思想：

- 每个类（包括新类）的样本点都围绕单一原型表示聚集在一起；
- 用神经网络学习输入图像到嵌入空间的非线性映射，并将类的原型作为嵌入空间中支撑集的均值。通过查找最近的类原型，对嵌入空间的查询点进行分类；
- 训练过程与测试过程一致，mini-batches (episodes)、support sets、query sets 这样的训练策略很适合小样本问题。

从训练过程上分析，如引言部分的介绍和图 2 的优化过程展示，通过训练和测试过程中引入支撑集、查询集的概念，每次不是像传统 batch 那样随机抽取，而是根据这一批类别抽取对应数目的 episode，每个 episode 是先选取一部分类别再每个类别抽取固定的训练样本数目。这种训练方式可以让梯度下降的状态符合我们预期的要求，又能“推开”类间距离，又能“拉进”类内距离。

从模型结构上分析，相比微调方法，原型网络的一个显著区别是没有传统神经网络方法的全连接层这一分类层；比较典型和简单的微调方法在训练结束，要迁移到其他数据时，通常冻结除全连接层以外的层，即特征提取部分的卷积层不更新参数，只用新的、规模较少的数据重新训练全连接层（本次复现工作中的微调也是按照此思路），这样会导致前面特征提取部分不一定完全适合，且用于分类的全连接层也没有被充分训练。而原型网络的分类结构相当于只是根据欧式距离来衡量从特征提取器得到的特征向量/嵌入的距离，并给出分类结果，这种优化损失的方式可以在特征提取部分更好的泛化，更多的去关注类和类的不同，并且避免了全连接层可能带来的过拟合问题。

3.1.3 整体架构

原型网络算法的核心伪代码及解释如算法 1 所示：

Procedure 1 Training episode loss computation for Prototypical Network.

Input: 训练集 $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$, 其中 $y_i \in \{1, \dots, K\}$. D_k 是 D 中包含所有 $y_i = k$ 时的 (x_i, y_i) 对的子集

Output: 随机训练步长训练得到的损失 J

$$V \leftarrow \text{RANDOMSAMPLE}(1, \dots, K, N_c) \quad \triangleright \text{为每个 episode 抽取类别}$$

for k in $\{1, \dots, N_c\}$ **do**

- $S_k \leftarrow \text{RANDOMSAMPLE}(D_{v_k}, N_s) \quad \triangleright \text{根据 label 抽取支撑集数据}$
- $Q_k \leftarrow \text{RANDOMSAMPLE}(D_{v_k} \setminus S_k, N_Q) \quad \triangleright \text{根据 label 抽取查询集数据}$
- $c_k \leftarrow \frac{1}{N_c} \sum_{(x_i, y_j) \in S_k} f_\phi(x_i) \quad \triangleright \text{从支撑集中计算出原型}$

end

$J \leftarrow 0 \quad \triangleright \text{初始化损失值}$

for k in $\{1, \dots, N_c\}$ **do**

- for** (x, y) in Q_k **do**

 - $J \leftarrow J + \frac{1}{N_c N_Q} [d(f_\phi(x), c_k) + \log \sum_k \exp(-d(f_\phi(x), c_k))] \quad \triangleright \text{更新损失}$

- end**

end

学习过程通过 SGD 最优化方法，最小化真实标签 k 的负对数概率： $J(\phi) = -\log p_\phi(y = k|x)$ 。训练时的 episodes 是通过从训练集中随机选择一个类的子集来形成的，然后在每个类中选择一个示例子集作为支持集，其余的子集作为查询集。算法 1 提供了计算训练集损失 $J(\phi)$ 的伪代码及注解。

图 6 中 (a) 展示了原型网络将数据抽象成一个嵌入空间的点后如何执行分类：支撑集的每个点求空间上的平均，这个中心将作为原型，之后需要查询的数据映射到嵌入空间的点以后，比较该点和每个原型的距离，选择最近的原型的标签作为自己的标签；(b) 展示了原型网络直观上的全部流程。

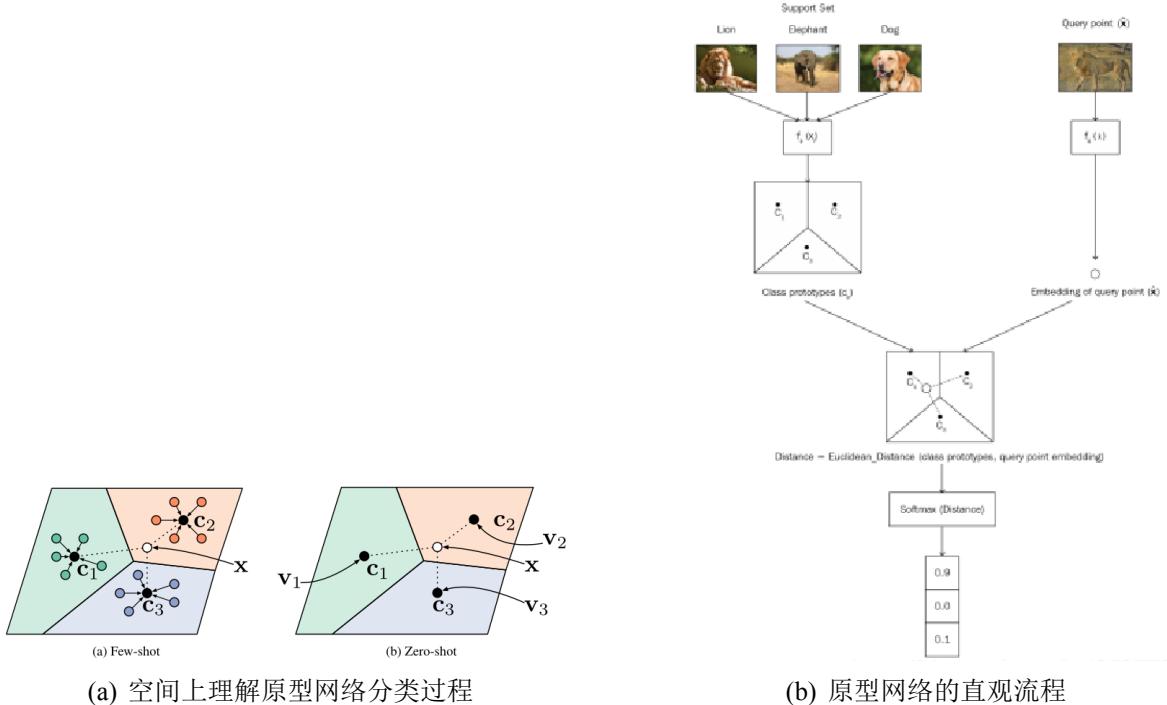


图 6: 原型网络整体架构

3.2 特征提取模块

原型网络的特征提取器是一个结构非常简单的 Conv4Net 结构，即一个四层 Block 的卷积，每个 Block 结构都相同。其实现代码和结构示意图如下及图 7。这个比较简单的结构实际上有非常大的可优

化空间，例如简单一些的将其替换为目前常用的 ResNet 结构，以及复杂一些的设计预训练-元训练两次训练调整特征提取器，本次复现工作还尝试了加入当前广泛研究的 attention 机制，在 miniImagenet 数据集上取得了很好的效果。

```

1 def EmbeddingBlock(input_channels):
2     return nn.Sequential(
3         nn.Conv2d(input_channels, 64, kernel_size=3, padding=1),
4         nn.BatchNorm2d(64),
5         nn.ReLU(),
6         nn.MaxPool2d(2, ceil_mod = False)
7     )
8 def embedding_module(input_channels=3):
9     return nn.Sequential(
10        embeddingBlock(input_channels),
11        EmbeddingBlock(64),
12        EmbeddingBlock(64),
13        EmbeddingBlock(64),
14        nn.Flatten(start_dim=1)
15    )

```

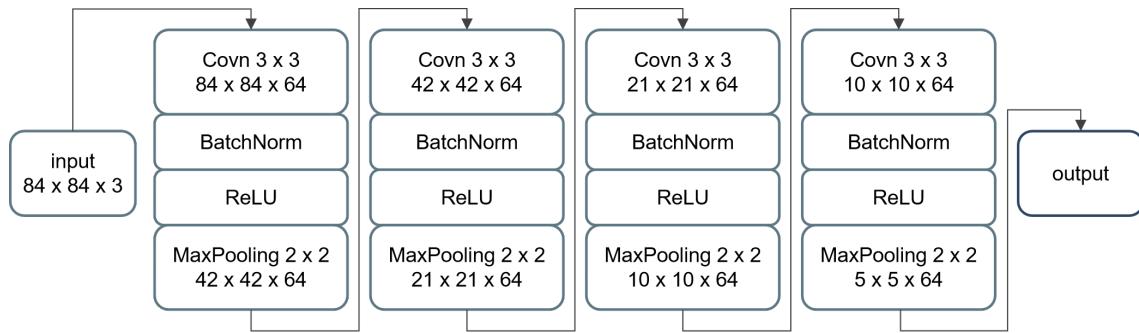


图 7: 特征提取模块的四层卷积结构

3.3 嵌入分类模块

公式中的变量含义同算法 1，原型的计算公式如式 (1):

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

简单来说，每个原型的计算是将支撑集中，相同的类的各 shot 样本计算完特征向量后直接求取每个维度数值的平均值。作者在原文章节 2.5 和 2.6 中给出了数学证明：在分类部分使用欧式距离时，这样的平均计算方法是最优的，可以使平均类内间距最小并使平均类间间距最大。如果使用的是余弦距离等非 Bregman 散度，则这样求原型不再保证最优。

估算类别方法如式 (2):

$$p_\phi(y = k|x) = \frac{\exp(-d(f_\phi(x), c_k))}{\sum_{k'} \exp(-d(f_\phi(x), c_{k'}))} \quad (2)$$

其中 d 采用的即为欧式距离。SimpleCNAPs 等工作也有针对欧式距离的比较，如更换为马氏距离或 L1 距离等。嵌入分类模块没有参数，直接根据已经计算出的特征向量，比较向量的相似度即完成分类任务。这一过程即计算距离并对嵌入向量进行分类的示意图如图 8 所示。

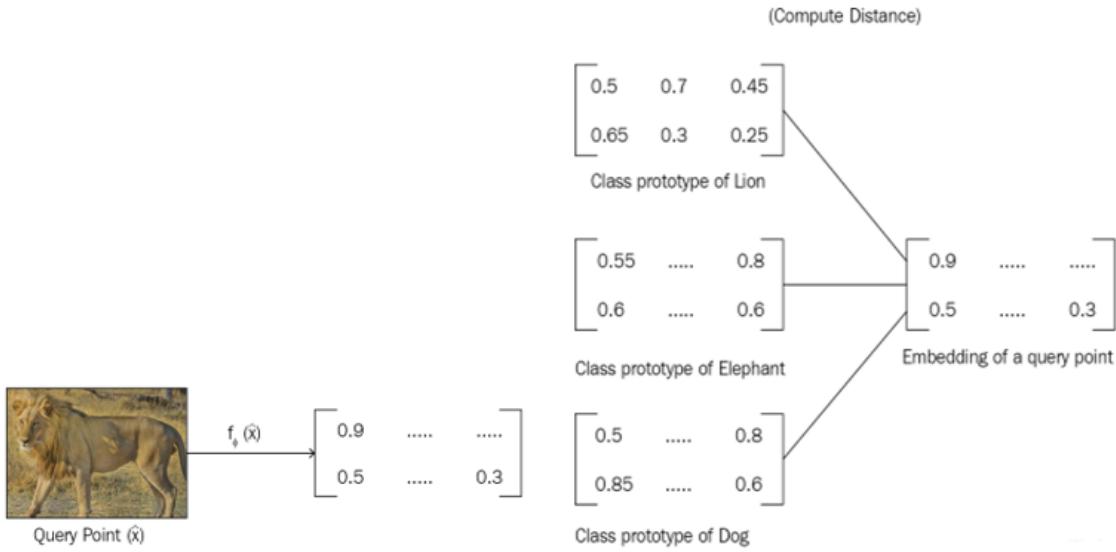


图 8: 计算距离并对嵌入向量分类的示意图

4 复现内容与细节

4.1 与已有开源代码对比

4.1.1 原型网络基础工作复现情况

与已有开源代码对比，原型网络有开源代码工作，但官方版本代码实际上并不易读，并且源代码没有给出另一个主要数据集 miniImagenet 是如何执行的，只给了报告结果，复现工作需要重现报告的准确率。

对于原型网络本身，本次复现工作的复现内容主要为以下三点：

<1> 代码完全重构：以自己的理解完全重写了原型网络代码，官方版本所用的钩子函数等可读性一般，实际上原型网络代码实现思路非常简单，重构后的整体结构非常简洁，并且准确率与报告基本一致，验证了复现的正确性，复现的代码架构如图 X 所示。

<2> 结果复现：源代码没有给出另一个关键数据集 miniImagenet 是如何执行的，复现工作通过调取网上的开源数据集和调整参数、设置等复现了该实验结果。全部的实验结果以.log 文件形式附在了提交附件中。

<3> 结果可视化：给出了自己实现的 t-SNE 代码将模型得到的向量可视化结果（t-Distributed Stochastic Neighbor Embedding，依赖 KL 散度，能够结合概率分布让降维后低维上数据的分布与原始特征空间的分布相似性高），可以直观了解到网络最后得到的原型是什么意义。

上面的内容为原型网络本身的内容，对于创新增量和显著改进，我在自己的工作中主要通过 miniImagenet 数据集体现出技术改进的效果；在实验室自采的声波手势数据集上进行应用的创新。这两个部分代码使用情况会在 4.1.2 和 4.1.3 两个部分再进行补充说明。关于复现过程所用的各部分代码使用情况列表如表 1 所示。

表 1: 开源代码使用情况

内容	引用代码来源	使用情况	备注说明
原型网络 (Conv4Net) 全部基础内容	https://github.com/jakesnell/prototypical-networks	完全重构，是重新实现的代码	增加了 miniImagenet 数据集的使用、调整以及结果 t-SNE 可视化
原型网络 (ResNet)	https://github.com/learnables/learn2learn/blob/master/learn2learn/vision/models/resnet12.py	改进 backbone 的 ResNet12 网络结构来自 l2l 库	—
微调/Finetune	—	完全由个人实现，除基础库外无引用	—
SimpleCNAPs	https://github.com/peymanbateni/simple-cnaps/tree/master/simple-cnaps-src	miniImagenet 部分基本使用源代码作为实验对比，gesture 部分有所改进	miniImagenet 代码所用的数据集是自己在 miniImagenet 对半划分的，gesture 为自己在实验室的工作，有根据 user 编号调整类别抽取策略、修改最后使用的度量函数等
Co-attention	https://github.com/blue-blue272/fewshot-CAN https://github.com/ignacio-rocco/ncnet https://github.com/juhongm999/chm	实现主要参考 Neighbourhood Consensus Networks（是做视觉匹配的）以及 Hough 卷积文章，代码结合以及部分参考，变化比较大，应用也不同	attention 思路主要来自 transformer、SENet、Cross Attention Network（它用的是 Fusion Layer）文章

原型网络基础部分实现的代码结构如图 9 所示，重构后的版本简洁易懂，可以实现各种不一样数据集的执行，通过给出 csv 文件，即训练集、测试集划分就可以使用。test 中还包括了 t-SNE 的可视化。



图 9: 原型网络基础部分实现的代码框架图

4.1.2 技术改进部分代码情况

对于 backbone 使用 ResNet 的改进，代码通过使用 learn2learn 库的 resnet12 模块实现。finetune 部分完全由自己实现，除了调用基础的深度学习库及网络结构的部分。这一部分代码分为预训练 main 函数和微调 Finetune 函数，微调时冻结其它卷积层，只重新训练最后的全连接层。SimpleCNAPs 部分在本次复现的 miniImagenet 数据集上未做技术改动，主要用于尝试效果和对比实验结果，miniImagenet 代码所用的数据集是自己在 miniImagenet 对半划分的；gesture 为自己在实验室的工作，有根据 user 编号调整类别抽取策略、修改最后使用的度量函数。co-attention 思路主要来自 transformer、SENet、Cross Attention Network（CAN 的技术用的是 Fusion Layer）文章；代码实现主要参考 Neighbourhood Consensus Networks（是一个做视觉匹配的工作）以及 Hough 卷积文章，使用情况以及引用代码来源可以参见上方表 1。

4.1.3 应用创新部分代码情况

应用创新部分即应用小样本方法到手势识别项目上。提供的 shot 数根据实际应用场景控制为 2, 这一部分之后对比的一个 baseline 将微调方法更换为最直接的 ResNet18 训练-测试。除了 SimpleCNAPs 方法，其他的技术开源代码使用情况与 miniImagenet 相同。SimpleCNAPs 方法在这里做了一个优化，结合实际情况，在元训练过程（meta-learning）抽取 episodes 的类时，根据数据集的用户编号 user，每个 episode 会随机抽取一个用户并让每个 task 数据都仅来自该用户，用以模拟真实场景，这样做也可以略提高准确率。

4.2 实验环境搭建

硬件上，自己运行的实验环境为实验室服务器，Ubuntu20.04，GPU 为 NVIDIA RTX A6000；软件上，主要需要的包有深度学习的 pytorch1.7.1 以及元学习的 learn2learn 库。所需要的库都已写入 environment.yml 文件中，在 Python 通过 conda env create –name envname –file environment.yml 即可得到 envname 虚拟环境。

4.3 复现内容与技术改进部分

4.3.1 与 Finetune 的对比

实现的微调方法分为预训练和微调两个函数，预训练部分用 miniImagenet 划分出的训练集（Ravi 版本，最通用的划分方式，以下全部相同）进行训练，在 5shot 时将每个样本的所有 shot 数据用于微调，调整预训练模型的全连接层。由于 1shot 时数据过少，该微调方法仅适用于 miniImagenet 的 5shot，微调方法准确率在 5shot 时为 51.07%，低于复现出的原型网络 5shot 时的 68.43%。并且原型网络还可以对 1shot 情况进行分类，表现出明显的优势。

4.3.2 backbone 网络优化

对于 miniImagenet (84*84)，四层卷积层的结构还是有些简单，可以用更深一些，性能普遍来说更好的 Resnet 系列网络：Resnet12 是对于输入 84*84 的一个常用的 Resnet 结构，将图 10(a) 中红色的特征提取部分替换为图 10(b) 结构的 ResNet12，1shot 准确率由复现的 48.87% 提升到 54.11%，5shot 准确率由复现的 68.43% 提升到 70.64%。相比之下有一定的提升。

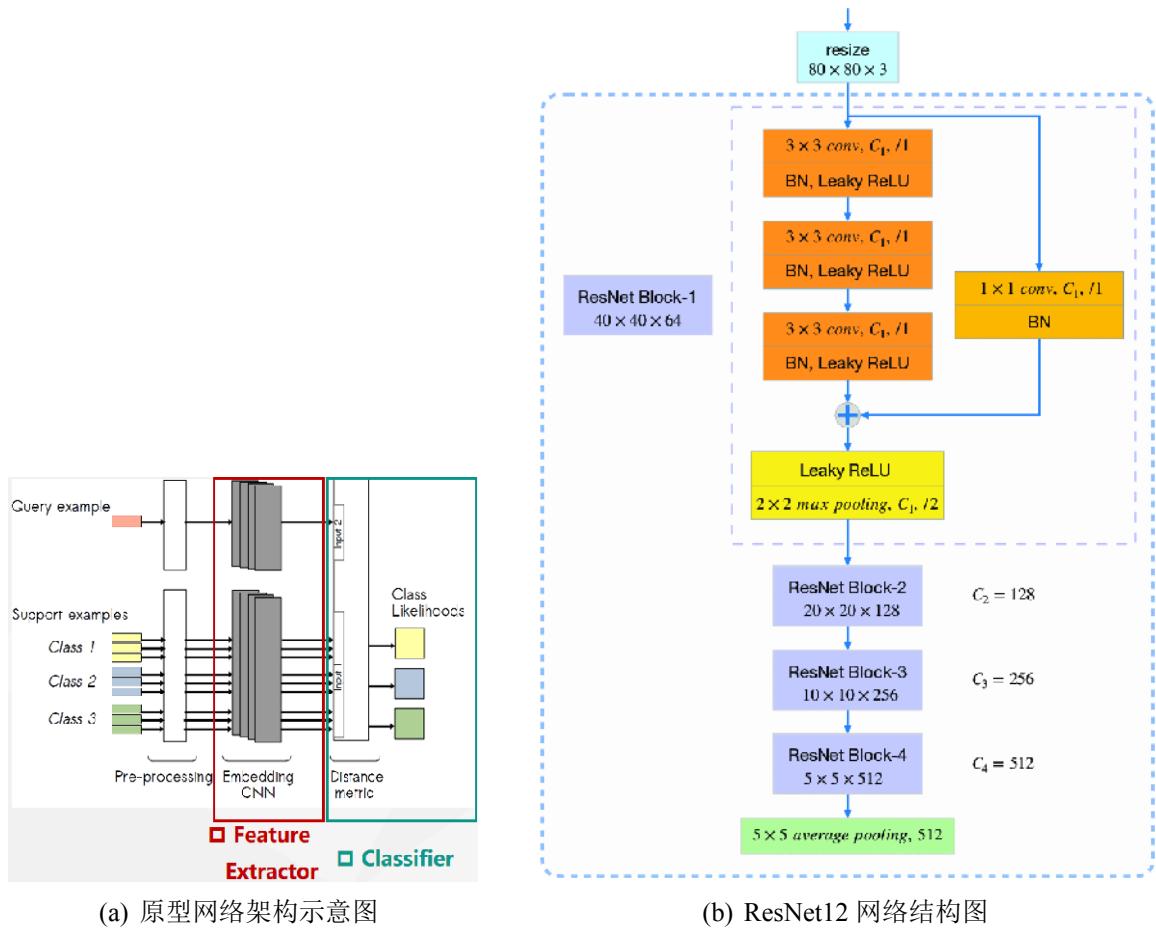


图 10: backbone 网络优化方式

4.3.3 SimpleCNAPs 使用

对于图 10(a) 的原型网络结构示意图，除了简单的优化特征提取器的骨架外，还可以改变特征提取器的架构。原型网络相当于所有数据一次性的训练，是一种全局的训练方式；CNAPs、SimpleCNAPs 等方法会将训练集数据再划分，划分为不相交的预训练数据集和元训练数据集，元训练过程用于调整预训练得到的模型参数。CNAPs 和 SimpleCNAPs 通过引入 FiLM 层（Feature-wise Linear Modulation，最早来自论文《FiLM: Visual Reasoning with a General Conditioning Layer》^[16]，针对的问题是 NLP 生成任务加 CV 分类器），通过生成 γ 与 β 参数，对每个特征图线性调制，可以更好的适应域的变化，其结构如图 11 所示。

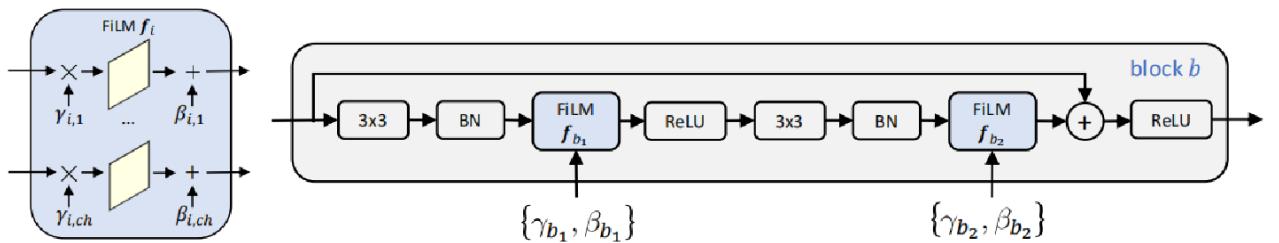


图 11: FiLM 结构示意图

个人测试的实验结果，只用 miniImagenet 的数据量较少，训练集的一半（32 类，19200 条数据）预训练，另一半元训练的准确率 5shot 仅有 45.96%。如果数据足够多时（用 ImageNet 里面不在 miniImagenet 的部分预训练），原论文报告 5shot 准确率可以达到 90%。

4.3.4 co-attention 机制

对于图 10(a) 的原型网络结构示意图，还可以在绿色的部分进行改进，即获取原型的方式上。最后的 cross-attention 即引入交叉注意力机制的改进，核心思想是交叉 attention 能通过图像位置间相关性，告诉我们的查询集和支撑集间“需要关注哪里”。思路主要来自于 Cross Attention Network^[30]，但与其 Fusion Layer 的方法不同，本次实现仍是直接计算相似度然后卷积，再计算回注意力分数。实现主要参考 Neighbourhood Consensus Networks^[31]（做视觉匹配的工作），如图 12(a) 所示；和 Hough 卷积^[32]，如图 12(b) 所示；将其 4D 卷积模块拿到原型网络得到特征向量后的部分来计算交叉 attention。

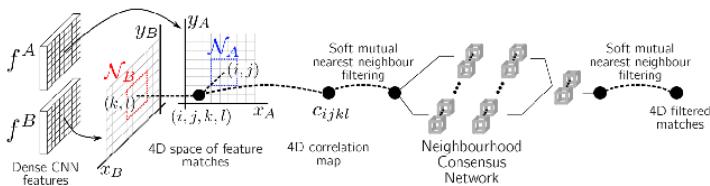
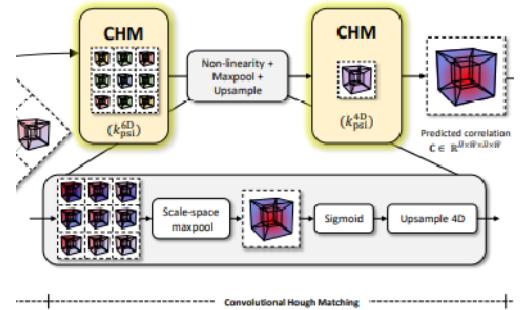


Figure 1: Overview of the proposed method. A fully convolutional neural network is used to extract dense image descriptors f^A and f^B for images I_A and I_B , respectively. All pairs of individual feature matches f_{ij}^A and f_{kl}^B are represented in the 4-D space of matches (i, j, k, l) (here shown as a 3-D perspective for illustration), and their matching scores stored in the 4-D correlation tensor c . These matches are further processed by the proposed soft-nearest neighbour filtering and neighbourhood consensus network (see Figure 2) to produce the final set of output correspondences.

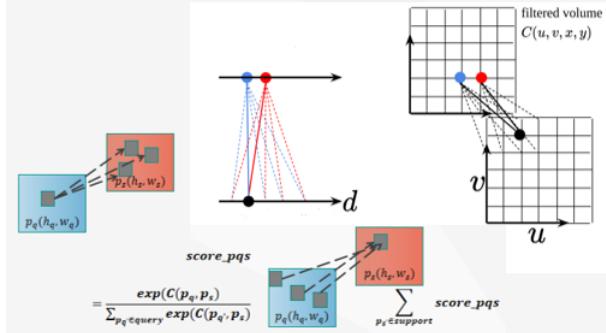
(a) Neighbourhood Consensus Networks



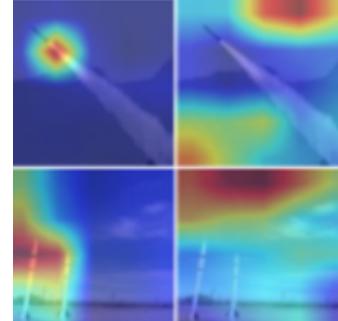
(b) Convolutional Hough Matching

图 12: 复现工作所参考的论文

这种改变原型方式（注意力）的方法可以让 query 找到更关注的部分，并且让 support 能结合和它相似的 query 去调整自己的原型，图 13(a) 展示了 2 维的 attention 是如何计算的。如图 13(b) 所示，引入 attention 机制可以让 query 更去关注哪个原型有和自己主要内容相匹配的；此外还要对 attention 张量卷积是因为设计的结构没有参数，交叉注意力可能会有一些“不相关的相关”，图 13(b) 如果不加限制，网络会认为天空这一占整幅图像比重最大的内容才是最需要关注的内容。加入参数就是来进行限制，例如通过这幅图与其他火箭相关的图的匹配情况，调整网络参数，消除那些不可靠的相关。



(a) 2 维的 attention map



(b) 关注的位置举例说明

图 13: 对引入 attention 机制的一些解释

设计的总体架构如图 14 所示，attention 具体的计算如图 15 所示。这里的创新体现在相比于直接进行 2D 卷积，这种 $H \times W \times H \times W$ 根据 Neighbourhood Consensus Networks^[31]的文章，这样实际上是对原来卷积核中心位置扩展到一个 $3 \times 3 = 9$ 个卷积 map(大小为 3×3)，扩大范围并增加对中心附近的卷积密度，可以更好发挥“邻里一致”的作用。对 attention map 进行卷积的实际操作用两个卷积层，分别将 $HWHW$ 的通道从 1 映射到 L 再映射回 1。此外，处理时还会多引入一个 loss：support 的原型不应因为 query 不一样而差别过大。

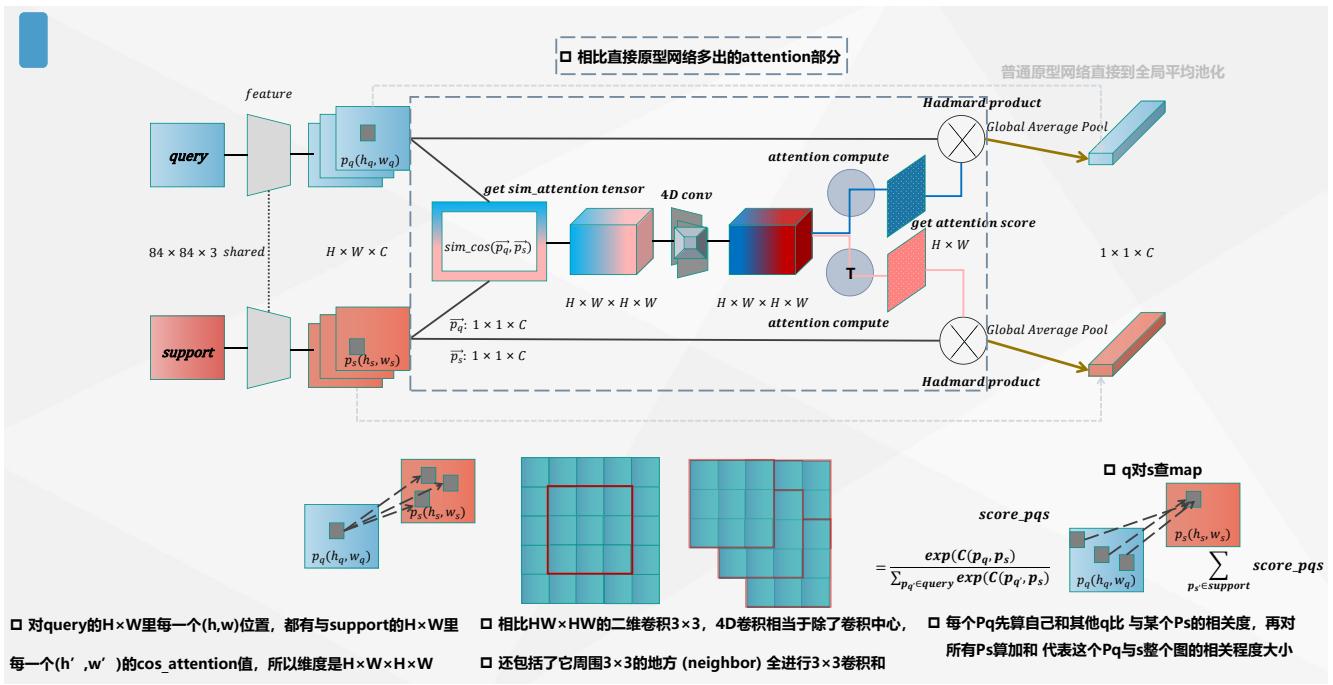


图 14: co-attention 的总体架构，在得到原型后加入注意力（矢量形式，可以放大查看该图）

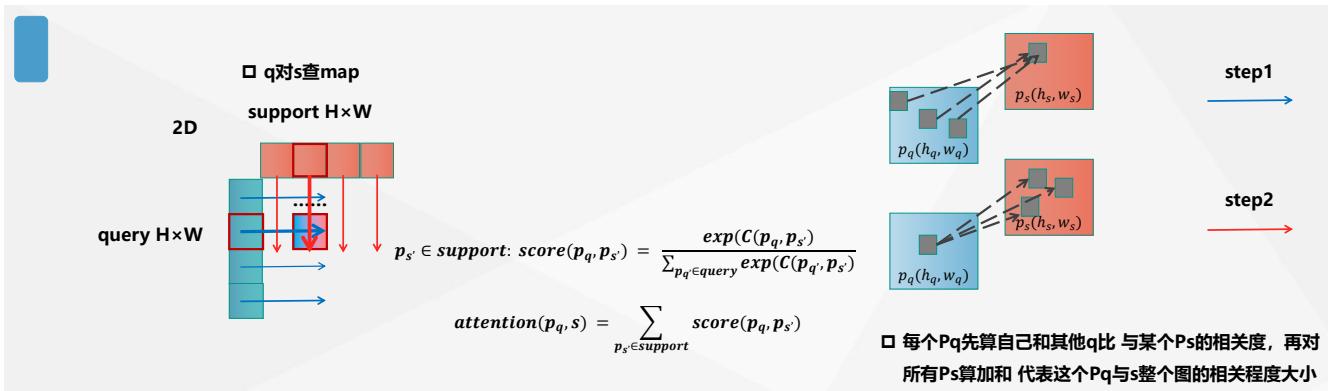


图 15: co-attention 具体计算注意力和加权的方式（矢量形式，可以放大查看该图）

其他一些想法和细节：

- **为什么还要对 attention tensor 卷积？**：相比 transformer 的 QKV，这个 sim 交叉注意力的计算实际上并没有引进参数，完全依赖前面 $1 \times 1 \times C$ 的向量，这样会导致交叉相关张量 C 可能包含不可靠的相关；为了消除那些“不相关的相关”，可以用卷积核再过滤一次匹配内容^[30]，张量上的四维卷积通过分析空间中相邻匹配的一致性来发挥几何匹配的作用；
- **为什么要这种 4D $3 \times 3 \times 3 \times 3$ 卷积？相比 $H \times W \times H \times W$ 2D 的优点？**：相比二维的 $H \times W$ ， 3×3 范围内的卷积，这样相当于卷积了一个 5×5 范围，对中心 3×3 的卷积更重^[31]（中心 1—— 3×3 ）；
- **最后的 attention 图和哈达玛积是什么？**：attention 图是 $H \times W$ 的，对 query 就是反应 query 上某点与整张 support 图的相关程度，所以先求该点相比其他 query 点的相关度，再求对整张 support 点的相关度和；最后的哈达玛积就是点乘，类似向量的点乘，直接点乘为 query 的特征图加权。

4.4 应用创新部分

将小样本学习应用到手势识别上的最大意义：采更少的数据并适应多种域和不同的书写任务。

4.4.1 手势 cross-envir

cross-envir 指训练时所采的数据和测试时（用户实际使用时）的数据并非相同环境，可能体现在设备、噪声、姿势的不同，以设备为例，训练时数据可能在平板上采集，而用户实际用手机使用，由于麦克风、扬声器位置不同导致的物理变化，导致频谱图的变化很大，如图 16 所示。模型在不作调整的情况下很难直接应用。本次复现展示了训练时用十位实验者 p1-p10、平板、实验室环境等变量下的字符 0-9 的 4000 条数据训练，测试时仍是十位实验者 p1-p10、手机、实验室环境等除设备以外相同变量下的字符 0-9 的 2000 条数据测试。直接用 ResNet18 训练测试的准确率仅为 0.41（图 16 右侧部分，不调整模型），作为对比，使用 SimpleCNAPs 一半数据预训练，另一半数据元训练准确率高达 0.74，普通原型网络 ResNet12 作为 backbone 也有 0.57 的准确率，体现了小样本方法的优势。详细的实验结果参考第 5 部分的实验结果分析。

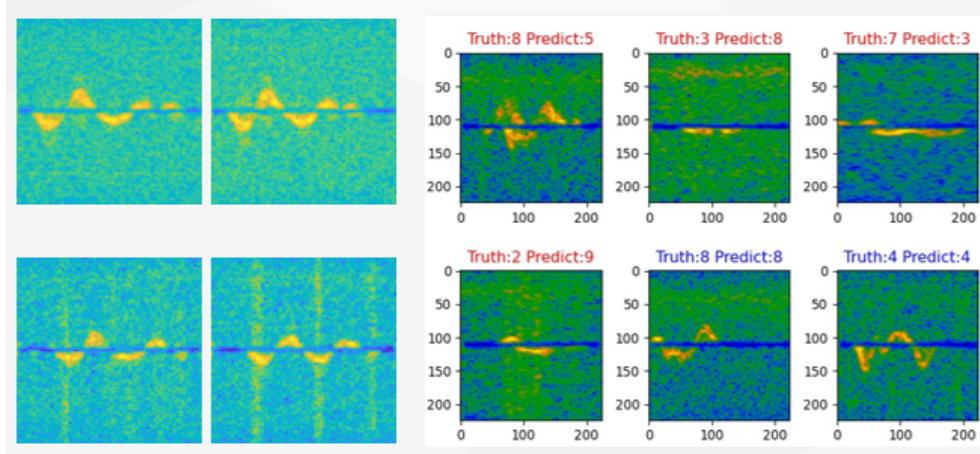


图 16: 跨设备的运行结果，左半部分上方为平板数据，下方为相同字符的手机数据

4.4.2 手势 cross-label

cross-label 指训练时用一些字符如 0-9 数据训练，测试时需要识别其他的一些字符如 A-Z，但只给出了 A-Z 很少量的一些数据，这种跨标签的识别也是我们工作的重点，如果小样本方法可以成功应用，就可以说明采集的数据可以不同于实际测试时的标签，也为用户自定义手势提供了很大的可能。这种情况下用户只需要提供很少量的自定义手势样本就可以获得较高的自定义手势识别准确率。实际实验中，对于 0-9 的 4000 条训练数据，测试时从 A-Z 随机抽取十个类别，每个类别提供 2 个样本，基础原型网络准确率都可以达到 0.61，用 SimpleCNAPs 准确率更可以达到 90%。详细的实验结果参考第 5 部分的实验结果分析。图 17 是工作的一些实物场景以及开发的 APP，这些也说明我们的研究工作大有可为。



图 17: 目前实现的实物场景及 APP

5 实验结果分析

5.1 原型网络本身工作的复现情况

原型网络基础工作的复现情况如表 2 所示，复现准确率与报告准确率基本一致，体现了复现工作的正确性。

表 2: 基础内容的复现结果

Result	omniglot 5-way Acc.		miniImagenet 5-way Acc.	
	1-shot	5-shot	1-shot	5-shot
PROTOTYPICAL NETWORKS (Reported)	98.8%	99.7%	49.42%	68.20%
PROTOTYPICAL NETWORKS (Implemented)	98.8%	99.7%	48.87%	68.43%

5.2 miniImagenet 上的技术改进

miniImagenet 上实现的各方法准确率如表 3 和图 18 所示，可以看到原型网络基础方法准确率在 5shot 时达到 68.43%，准确率明显超过微调方法，体现了小样本方法的优势。

表 3: miniImagenet 实现的准确率

Implemented Methods	5-way Acc.	
	1-shot	5-shot
FINETUNE	-	51.07%
PROTOTYPICAL NETWORKS (CONV4NET)*	48.87%	68.43%
PROTOTYPICAL NETWORKS (RESNET12)	54.11%	70.64%
SIMPLE CNAPs	31.00%	45.96%
CO-ATTENTION PROTOTYPE	63.85%	80.22%

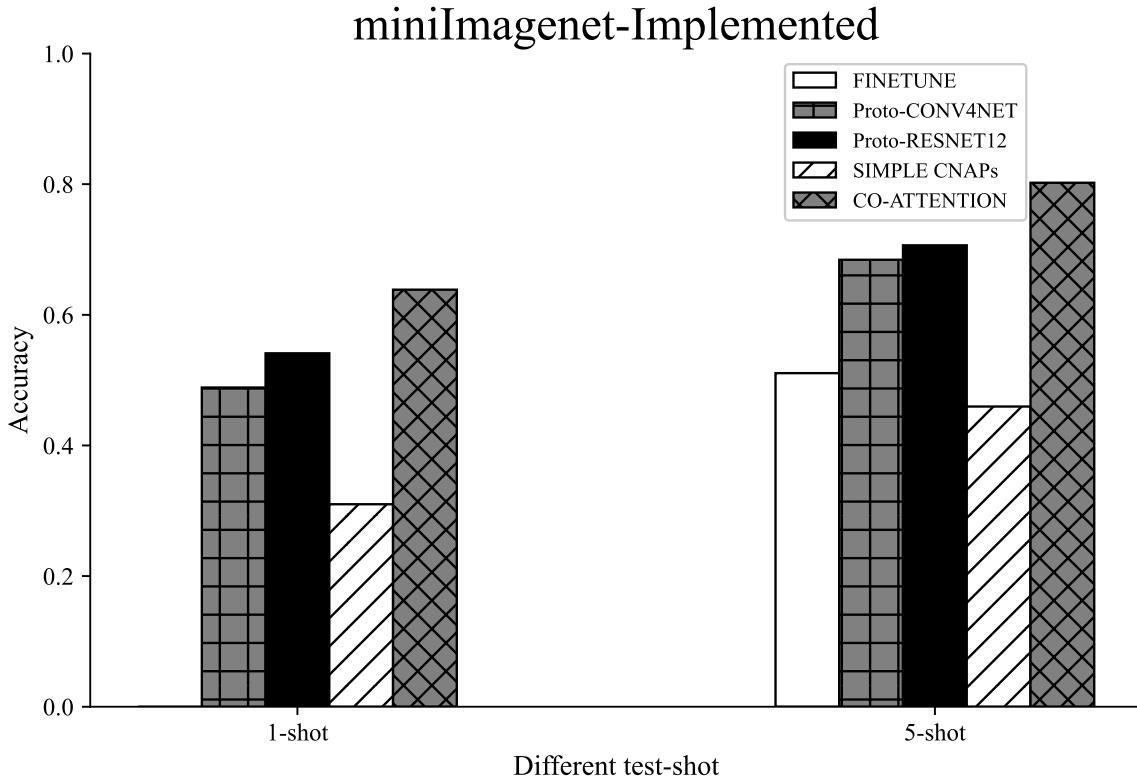


图 18: miniImagenet 实现的准确率

改进的方法中，将骨架替换为 ResNet12 可以获得一定的提升；SimpleCNAPs 方法准确率较低，可能因为 miniImagenet 图像数据集蕴含的信息很多，训练的数据量不足以完成“预训练 + 元训练”过程。在这篇论文的报告中使用 ImageNet 的其他数据全部做预训练，miniImagenet 训练集只做元训练，准确

率可以高达 90%；Attention 方法的准确率有明显提升，说明了复杂的现实生活图像中，注意力机制以及去除“不相关的相关”是能够起到很大作用的，对于这类蕴含信息多、信息比较隐含的目标，可以多尝试应用注意力机制。

5.3 gesture 上的应用创新

gesture 上实现的各方法准确率如表 4 和图 19 所示，可以看到这个数据集和任务下 SimpleCNAPs 准确率达到了最高水平。

表 4: gesture 实现的准确率

Result	cross-envir 10-way Acc.		cross-label 10-way Acc.	
	2-shot	-	2-shot	-
RESNET18 BASE	41.38%	-	-	-
PROTOTYPICAL NETWORKS (CONV4NET)*	50.24%	-	61.50%	-
PROTOTYPICAL NETWORKS (RESNET12)	57.33%	-	69.77%	-
SIMPLE CNAPs	74.06%	-	91.25%	-
CO-ATTENTION PROTOTYPE	56.27%	-	58.57%	-

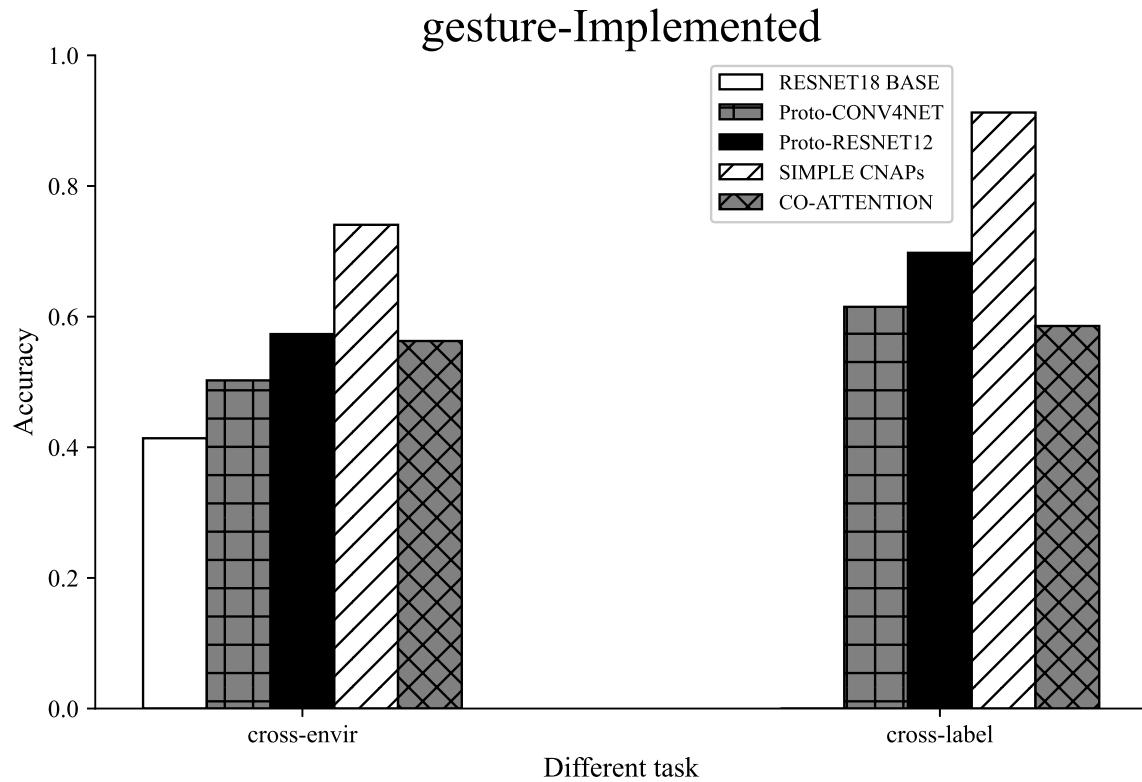


图 19: gesture 实现的准确率

可以看到，在手势频谱图这一简单图像下，图像蕴含的信息直观并且内容量较少，背景单一确定，内容集中在中心区域下，注意力机制的表现不佳甚至不如直接改进 backbone 而不做其他改进的准确率。SimpleCNAPs 方法由于所需的数据量已经足够，这种“预训练 + 元训练”的方式就可以表现出优异的性能。这也说明了不同的问题、不同的数据集下，不同方法会有不同的效果，不一定一种方法始终是性能优秀的，需要实际尝试才可以确定。

6 总结与展望

6.1 总结

全部工作的总结如图 20 所示：

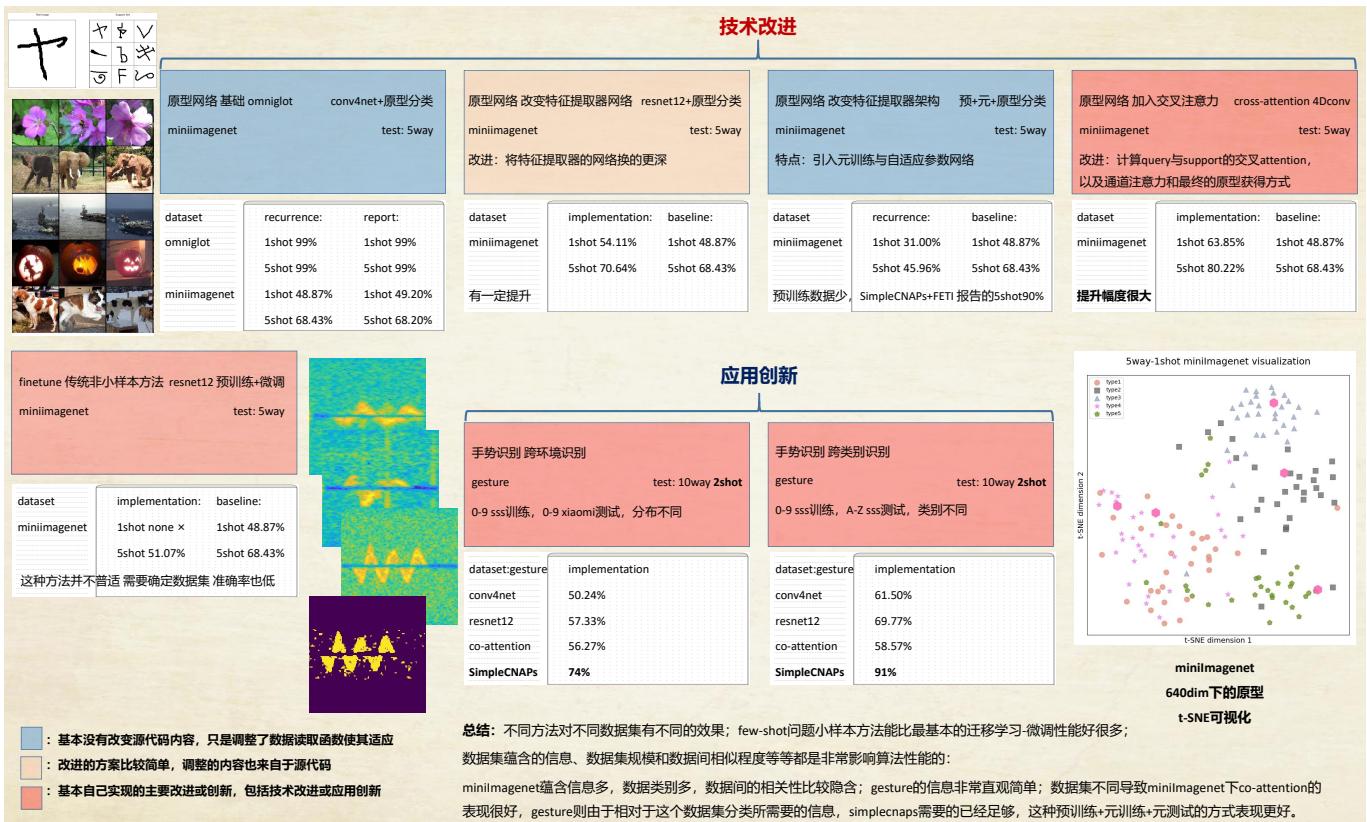


图 20: 复现工作的完整总结 (矢量形式, 可以放大查看该图)

关于复现工作以及工作量的总结:

- 复现部分的工作量:** 理解原型网络思想; 调研小样本学习的各种方法; 完成代码的重构与实现, 并展示执行结果;
- 技术改进部分的工作量:** 实现了微调方法 (Finetune) 并对比; 改变了特征提取器骨架并获得了一定效果; 尝试了 SimpleCNAPs 的开源代码; 综合了各种注意力机制的内容组合出一个插入原型提取器后的 4D 卷积-交叉注意力模块, 明显提升了 miniImagenet 的识别准确率;
- 应用创新部分的工作量:** 应用各种小样本方法到一个现实应用: 基于声波的手势频谱图识别, 对跨实验环境的识别和跨书写内容的识别 (跨分布与跨标签) 都有明显效果。

关于复现工作得到的结果和启示: 不同方法对不同数据集有不同的效果; few-shot 问题小样本方法能比最基本的迁移学习-微调性能好很多; 此外, 数据集蕴含的信息、数据集规模和数据间相似程度等等都是非常影响算法性能的: miniImagenet 蕴含信息多, 数据类别多, 数据间的相关性比较隐含, 而 gesture 的信息就非常直观简单; 数据集不同导致 miniImagenet 下 co-attention 的表现很好, gesture 则由于相对于这个数据集分类所需要的信息, simplecnaps 需要的已经足够, 这种预训练 + 元训练 + 元测试的方式表现更好。

除此之外, 跨域, 即跨分布问题有时比类别的改变更难解决; 不同分布域是物联网里很需要解决的一个问题, 研究意义很重大, 之后也会继续研究下去。

6.2 展望

未来的工作主要可能有两方面, 一方面引入注意力机制后的手势频谱图效果不甚理想, 可能可以通过 Class Activation Mapping (与全局平均池化方法有关) 来可视化注意力机制关注的内容, 帮助看一下为什么手势频谱图下 attention 的表现不理想, 注意力机制实际上注意到哪些位置了; 另一个方

面就是原型网络的半监督改进，也是目前我主要研究的方向。本次工作介绍的原型网络是一个监督学习工作，实际上如果支撑集中包括了没有标签的数据，问题就将转化为半监督小样本问题。事实上半监督原型网络方法在原型网络工作的次年就由这一工作的作者所做出^[33]，他们的核心思想如图 21 所示，利用无标签的真实数据，调整原型在空间的分布位置，更大程度的利用数据。图 21 中右侧展示的就是半监督原型网络的思路，图中原来的 2-shot 相当于可以提供一个比较好的初始随机聚类中心，unlabeled set 的数据对原型位置有一个调整作用，聚类中心即原型的中心可以直接求平均。这种方法在 miniImagenet 数据集上也有明显的效果提升，这也将是我之后研究的主要内容。

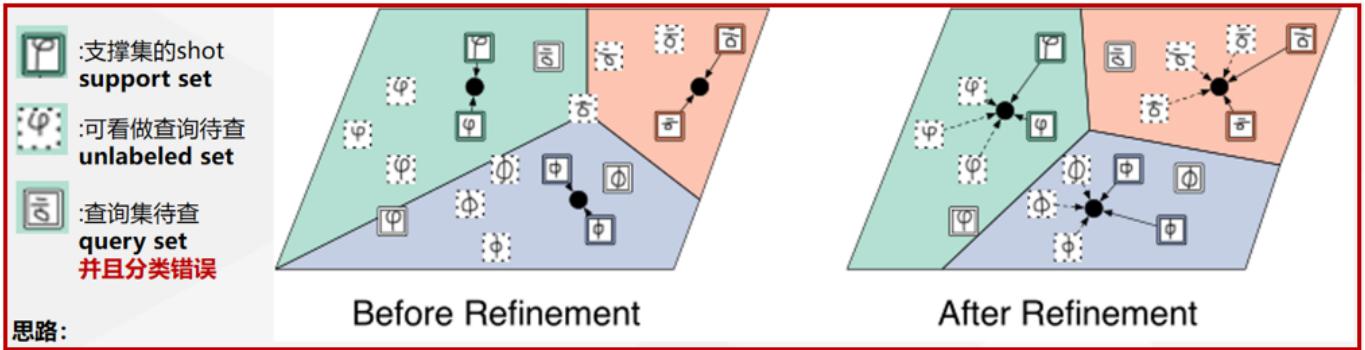


图 21: 半监督原型网络的主要思路^[33]

本次工作其实也没有特别重点或者大的关注点，算一个综合性质的复现，也希望能和大家多多交流；自己的创新其实也是组合其他已有想法的“搭积木”，但是能在现实应用中起到好的效果，应该就可以实现物联网的价值。

最后，非常感谢罗胜老师在前沿技术课过程中的付出，老师在选题阶段就认真负责的帮助我确定选题，讨论选题内容；PPT 演讲时也对我的工作进行了细致的指导和鼓励。再次感谢罗胜老师和本次课程！

参考文献

- [1] WANG Y, YAO Q, KWOK J T, et al. Generalizing from a few examples: A survey on few-shot learning [J]. ACM computing surveys (csur), 2020, 53(3): 1-34.
- [2] RAVI S, LAROCHELLE H. Optimization as a model for few-shot learning[J]., 2016.
- [3] OUYANG W, WANG X, ZHANG C, et al. Factors in finetuning deep model for object detection with long-tail distribution[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 864-873.
- [4] WANG Y, YAO Q. Few-shot learning: A survey[J]., 2019.
- [5] BATENI P, GOYAL R, MASRANI V, et al. Improved few-shot visual classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14493-14502.
- [6] RUSU A A, RAO D, SYGNOWSKI J, et al. Meta-learning with latent embedding optimization[J]. arXiv preprint arXiv:1807.05960, 2018.
- [7] TRIANTAFILLOU E, ZHU T, DUMOULIN V, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples[J]. arXiv preprint arXiv:1903.03096, 2019.
- [8] REQUEIMA J, GORDON J, BRONSKILL J, et al. Fast and flexible multi-task classification using conditional neural adaptive processes[J]. Advances in Neural Information Processing Systems, 2019, 32.
- [9] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks [C]//International conference on machine learning. 2017: 1126-1135.
- [10] BANERJEE A, MERUGU S, DHILLON I S, et al. Clustering with Bregman divergences.[J]. Journal of machine learning research, 2005, 6(10).
- [11] SUNG F, YANG Y, ZHANG L, et al. Learning to compare: Relation network for few-shot learning[C] //Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1199-1208.
- [12] KOCH G, ZEMEL R, SALAKHUTDINOV R, et al. Siamese neural networks for one-shot image recognition[C]//ICML deep learning workshop: vol. 2. 2015: 0.
- [13] VINYALS O, BLUNDELL C, LILLICRAP T, et al. Matching networks for one shot learning[J]. Advances in neural information processing systems, 2016, 29.
- [14] MISHRA N, ROHANINEJAD M, CHEN X, et al. Meta-learning with temporal convolutions[J]. arXiv preprint arXiv:1707.03141, 2017, 2(7): 23.
- [15] GARNELO M, ROSENBAUM D, MADDISON C, et al. Conditional neural processes[C]// International Conference on Machine Learning. 2018: 1704-1713.
- [16] PEREZ E, STRUB F, DE VRIES H, et al. Film: Visual reasoning with a general conditioning layer[C] //Proceedings of the AAAI Conference on Artificial Intelligence: vol. 32: 1. 2018.

- [17] VASWANI A, SHAZER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [18] KALGAONKAR K, RAJ B. One-handed gesture recognition using ultrasonic Doppler sonar[C]//2009 IEEE International Conference on Acoustics, Speech and Signal Processing. 2009: 1889-1892.
- [19] GUPTA S, MORRIS D, PATEL S, et al. Soundwave: using the doppler effect to sense gestures[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2012: 1911-1914.
- [20] QIFAN Y, HAO T, XUEBING Z, et al. Dolphin: Ultrasonic-based gesture recognition on smartphone platform[C]//2014 IEEE 17th International Conference on Computational Science and Engineering. 2014: 1461-1468.
- [21] WANG W, LIU A X, SUN K. Device-free gesture tracking using acoustic signals[C]//Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking. 2016: 82-94.
- [22] WANG X, SUN K, ZHAO T, et al. Dynamic speed warping: Similarity-based one-shot learning for device-free gesture signals[C]//IEEE INFOCOM 2020-IEEE Conference on Computer Communications. 2020: 556-565.
- [23] WEISS K, KHOSHGOFTAAR T M, WANG D. A survey of transfer learning[J]. Journal of Big data, 2016, 3(1): 1-40.
- [24] OZCAN T, BASTURK A. Transfer learning-based convolutional neural networks with heuristic optimization for hand gesture recognition[J]. Neural Computing and Applications, 2019, 31(12): 8955-8970.
- [25] RAHIMIAN E, ZABIHI S, ASIF A, et al. Fs-hgr: Few-shot learning for hand gesture recognition via electromyography[J]. IEEE transactions on neural systems and rehabilitation engineering, 2021, 29: 1004-1015.
- [26] WANG J, ZHANG L, WANG C, et al. Device-free human gesture recognition with generative adversarial networks[J]. IEEE Internet of Things Journal, 2020, 7(8): 7678-7688.
- [27] LING K, DAI H, LIU Y, et al. Ultragesture: Fine-grained gesture sensing and recognition[J]. IEEE Transactions on Mobile Computing, 2020.
- [28] WANG Y, SHEN J, ZHENG Y. Push the limit of acoustic gesture recognition[J]. IEEE Transactions on Mobile Computing, 2020.
- [29] LIU C, WANG P, JIANG R, et al. Amt: Acoustic multi-target tracking with smartphone mimo system [C]//IEEE INFOCOM 2021-IEEE Conference on Computer Communications. 2021: 1-10.
- [30] HOU R, CHANG H, MA B, et al. Cross Attention Network for Few-shot Classification[C]//NeurIPS. 2019.
- [31] ROCCO I, CIMPOI M, ARANDJELOVIĆ R, et al. Neighbourhood consensus networks[J]. Advances in neural information processing systems, 2018, 31.

- [32] MIN J, CHO M. Convolutional Hough Matching Networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 2940-2950.
- [33] REN M, TRIANTAFILLOU E, RAVI S, et al. Meta-learning for semi-supervised few-shot classification [J]. arXiv preprint arXiv:1803.00676, 2018.