一种针对 One-Shot Learning 的神经网络模型

2017110220　曹亚帅
yashcao@bupt.edu.cn

17th December 2017

1 Introduction

参考原文：《Matching Networks for One Shot Learning》[*].
论文性质：NIPS(Neural Information Processing Systems) 国际会议，2016.

主要问题：

1. 机器学习领域虽然近年来在 CV 和 NLP 等方面取得了很多进展，但特点依旧是：大量的数据 + 简单的算法 > 少量的数据 + 复杂的算法。尤其是大数据驱动的深度学习，相比之下人类可以通过几个例子学会一个概念，但是深度学习却需要大量的数据才能达到准确地识别。论文解决的就是**小样本学习问题**。

2. 论文的工作是：借鉴了深度神经特征的度量学习 (metric learning) 的思想，以及外部记忆增强神经网络和 attention 机制的思想。设计的模型可将一个小的标注集以及一个未标注的测试样例映射到它对应的标签，在这个过程中避免了对于新的标签类别进行调整的需求。

本文的整体结构如下: 第1节说明了本文要解决的问题。在第2节中简述本文涉及到的机器学习相关的基础理论，联系了本学期课程。第3节分析了提出方法的创新点。第4节给出训练的方法。

2 Theorem

2.1 参数模型 和 非参数模型

参数模型即给定选择的目标函数形式，并学习一系列固定个数的参数尽可能表征这些数据的模式。参数个数不会随样本量的增大而增加。参数模型往往对数据分布有较强的假设。

对于目标函数形式不作过多假设的算法称为非参数的机器学习算法。通过不做假设，算法可以自由的从训练数据中学习任意形式的函数。非参的方法寻找最合适的训练数据，同时保留一些对没有见过的数据的泛化能力。

[*]论文地址：http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning

2.2 Deep Neural Networks

深度神经网络是由许多相连的神经元组成的，一般包含输入层，隐层，输出层。假设输入样本为 $\boldsymbol{x}$，中间经过权重 $\boldsymbol{w}$ 和偏置 b 的计算得到加权输出 $\boldsymbol{w}^T\boldsymbol{x}+b=z$，每一层对 z 经过激活函数 $f(\cdot)$ 的处理得到预测 $y=f(z)$。

数据集样本 $\boldsymbol{x}$ 对应的标签为 $\boldsymbol{t}$，通过 t 和 y 的误差得到 loss 函数，loss 函数经过梯度下降等方法优化 $\boldsymbol{w}$、b 最终使深度神经网络可以更加接近准确的预测。

深度学习虽然效果很好，但需要庞大的数据集且学习速度很慢。在数据集上使用随机梯度下降进行权重更新可以理解为通过模型将训练样本逐步学习到其大量参数中。

2.3 The Fully Conditional Embedding

嵌入函数是用来将图像或文本表示成某种向量的形式。从输入空间映射到将要进行距离比较的特征空间。论文中的嵌入函数 $f(\hat{x})$，$g(S)$ 分别是测试样本和训练样本的特征提取函数 (embedding)，使用深度网络实现。

2.4 The Attention Kernel

度量（或距离函数）是表示特征之间相似度距离的函数。度量学习 (Metric Learning) 的目的是为了衡量样本之间的相近程度，这是模式识别的核心问题之一。常见的度量距离有欧氏距离，余弦距离，甚至是 SVM 的核矩阵等。

attention 机制的方法是参考了人的注意力，比如我们看一幅性感图片，我们会很自然地把注意力集中在关键位置。因此利用以往的任务来训练一个 attention 模型, 从而面对新的任务能够直接关注最重要的部分。借鉴了类似 KNN 的的思路。

充当权重的 attention 函数 a 是测试样本和每个训练样本之间余弦距离 C 的 softmax（归一化）：

$$a(\hat{x},x_i)=\frac{\exp[C(f(\hat{x}),g(x_i))]}{\sum_{j=1}^{k}\exp[C(f(\hat{x}),g(x_j))]}$$

3 Model Analysis

3.1 解决方案

鉴于非参数模型的特点，论文提出了一种新的方法，旨在结合参数模型和非参数模型的最佳特性。论文提出方法特点如下。

创新点：

1. 建模：为了能够从单个标记样本中学习一个类，需要依赖非参数方法（使用记忆组件而非仅仅依靠训练的权重）。论文使用 Matching Net，这种神经网络借鉴了 attention 的思想实现快速学习。

2. 训练：基于简单的机器学习原则，即测试和训练条件必须匹配。为了训练达到快速学习，通过每一类仅提供几个样本来训练，将任务从一小批切换到一小批。训练过程中，尽量模拟测试流程，使用小样本构造 minibatch。

3.2 Matching Net 模型

S 是一个小的支撑集，包含 k 个样本：$\{(x_i, y_i)\}_{i=1}^{k}$，模型 Matching Net(简称 MN) 通过训练集得到的分类器是 $S \rightarrow c_S(\cdot)$。利用分类器可以对测试样本 $\hat{x}$ 预测 $c_S(\hat{x})$，预测的概率分布为 $P(\hat{y} \mid \hat{x}, S)$。概率函数 P 由神经网络参数化，因此直接得到的预测输出是 $arg\ \underset{y}{max} P(y \mid \hat{x}, S)$，关键是描述 P，因此与之匹配的网络模型 Matching Net 计算估计的预测表示如下：

$$\hat{y} = \sum_{i=1}^{k} a(\hat{x}, x_i) y_i$$

采用线性标签的组合可以举例理解：$0.9 \times cat + 0.1 \times dog$，输入 y_i 是一个独热向量。

根据上面公式的思想，MN 的结构如下图所示。

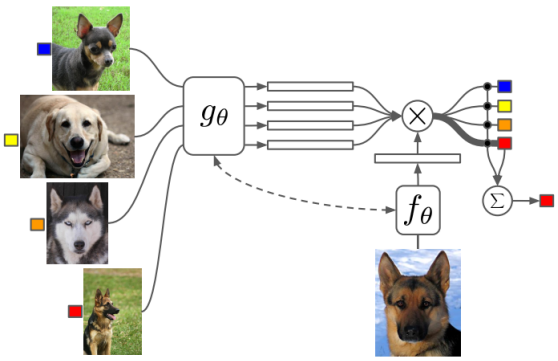

Figure 1: Matching Net 结构

4 Algorithm Implementation

4.1 训练策略

MN 模型预测的结果形式为 $P_\theta(\cdot \mid \hat{x}, S)$，MN 的参数集合 θ 包括嵌入函数 f 和 g 中的参数。首先定义任务 T，根据 T 抽样出标签集 L，记为 $L \sim T$。例如一个任务是图片“动物分类”，候选动物有 10 类 ($|T| = 10$)。首先选择 2 个类（如 L 为“cat”，“dog”），每个类选 5 个样本图片产生支持集 S，再抽样产生 batch B（标签也必须是已标注的 cat 和 dog，样本数量可以多一些），Matching Net 的目标是最大化根据支持集 S 预测 B 中标签的概率，如下所示：

$$\theta = arg\ \underset{\theta}{max} E_{L \sim T}[E_{S \sim L, B \sim L}[\sum_{(x,y) \in B} \log P_\theta(y, \mid x, S)]]$$

训练过程中，迭代一次的流程：

1. 选择少数几个类（如 2 类），在每个类中抽取少量样本；
2. 将选出的集合划分：支撑集 S，测试集 B；

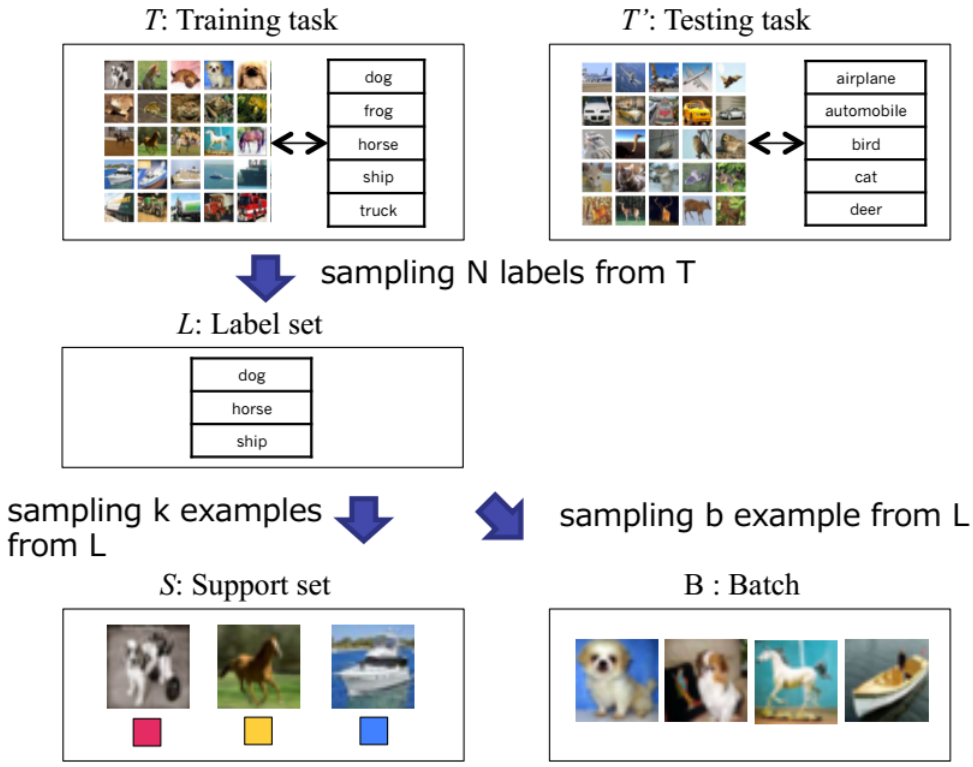

Figure 2: 训练流程

3. 利用本次迭代的支撑集，计算测试集的误差；

4. 计算梯度，更新参数。

特别注意的是，在完成训练之后，所有训练中用过的类别，都不能出现在后续**真正测试**中，即训练集和测试集的类别互不包含。

在测试过程中，同样遵守此流程：

1. 选择少数几个类别，在每个类别中选择少量样本；

2. 将选出的集合划分：支撑集，测试集；

3. 利用本次迭代的支撑集，计算测试集的误差。

4.2 小结

参考的这篇论文被审稿人评价论文叙述啰嗦，细节不清楚。但是论文提出的一个创新点甚至让 Feifei-Li 的高徒认为很有意思：训练一个端到端的最近邻分类器。由于近 2 年来 one-shot learning 被看好，所以很有借鉴意义。但同样这篇论文的写作硬伤让人很容易看的思路乱，所以我搜集了很多网络资料整合理解，也大概理解了基本思路。