

# Headline

大家好：

这是2018年度第7篇Arxiv Weekly。

本文是 人脸数据 方向的文章。

## Highlight

利用半监督学习和主动学习的思想，依托已有的标注数据，将海量无标注数据自动标注后引入训练，提升训练效果

## Information

### Title

*Consensus-Driven Propagation in Massive Unlabeled Data for Face Recognition*

### Link

<https://arxiv.org/pdf/1809.01407.pdf>

### Codes

<https://github.com/XiaohangZhan/cdp/>

### Source

- 香港中文大学-商汤联合实验室 (CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong)
- 商汤科技 (SenseTime Group Limited)
- 南洋理工大学 (Nanyang Technological University )

## Introduction

近年来，随着模型性能的飞跃和大量标注数据的采集，人脸识别的performance有了很大的提高。然而，进一步倍增人脸的数据集变得非常困难[基数太大，增加得少没用，增加得多成本高昂]

本文发现未标注的人脸数据也能像标注的数据一样发挥价值。为了通过实验说明这个问题，本文模拟了真实世界中简易图像采集的场景，也即 在无限制的场景下采集大量未标注的数据，且和我们已经获取的标注数据没有直接的联系。**显然这个场景的图像获取比较而言几乎没有成本的，同样在传统意义上也是没有用的。**本文的思路就是如何“废物利用”，“变废为宝”。

本文主要采用的方式是自底向上建立一个relational graph，来衡量自由图像和标注图像代表之间的语义相关度，从而将已有的label最可靠地赋予自由图像。文章提出了一致性驱动传播 (**Consensus-Driven Propagation, CDP**) 概念，利用committee (委员会) 和mediator (调停者) 两个模块来鲁邦德选择正样本对。

扩展性实验证了两个模块的有效性，尤其在无视边缘地匹配hard positive pair上。利用CDP，我们使用9%的labeled数据在MegaFace上训练得到了78.18%的精度。（无label下精度61.78%；全label下精度78.52%）。

## Keys

### 1. 问题论述

真正在工程当中，人脸的数据是要采集的，要标注的，要好多钱的……这就是本文试图缓解的问题。

## 2. 理论背景：SSL和AL

本文要采用的方法类似于半监督学习(semi-supervised learning, SSL)。显然无条件限制地大量采集人脸数据是非常便宜的。这样获得的数据无法直接用于训练，基于它们训练甚至比传统的SSL更加tough。

- 数据是从无限制环境采集的，意味着姿态角度、照明条件、遮挡条件都有极大的可变化范围，给判断两张脸之间的similarity带来了挑战。
- 搜集到的数据和已有的labeled数据之间没有identity的相似度保证，也即完全未必是同一群人。这使得一些传统的SSL算法失去用武之地。

这里插播一则主动学习 (Active Learning, AL)，因为这个问题和主动学习有一点神似，有些思路能相互借鉴。

正如之前讲到过的，**主动学习是一种动态的“待标记数据选择选择-标记后重训”的网络优化过程。旨在根据部分已知的标记，挑选未标记实例获取标记，并由此获得比静态随机采样监督学习更高的预测准确率**

实际上主动学习的具体思路有如下几个大类：

1. uncertainty sampling：选择最有信息量的数据。

这种想法是最为直接的，挑选已有网络inference时候置信度最低的数据作为最有价值进一步标记的数据。例如：

least confidence。二分类问题中越靠近0.5概率的样本越不确定；

margin sampling。选择每个实例最大和次大置信度之差，这个值越小样本越不确定；  
等等.....

2. expected error reduction：最小化期望误差。

一种更为中二的思路，遍历待标记数据集，挑选出加入标记数据集后能够使得error期望最小化的那些数据进行标记。显然这个思路非常time-consuming，除非有很特殊的条件约束否则不实用。

3. least version space：最小化解释空间。

对于一个已标记的实例集  $\Gamma$ ，与其一致的所有统计学习模型称为  $\Gamma$  的解释空间(version space)。解释空间越大，就意味着我们有越多的模型可以选择；当解释空间只有一个点时，统计学习的模型也可以唯一确定。

因此，可以总是选择那些能够最大程度缩小解释空间的实例进行标记，使得我们的模型朝着唯一化方向训练。  
这方法的合理性建立在“NN模型足够刻画要拟合的非线性关系，要解决的问题是其刻画的冗余性”的信仰上。

其中思路3引出了一种重要的算法，也即委员会质询 (query-by-committee, QBC) 算法。

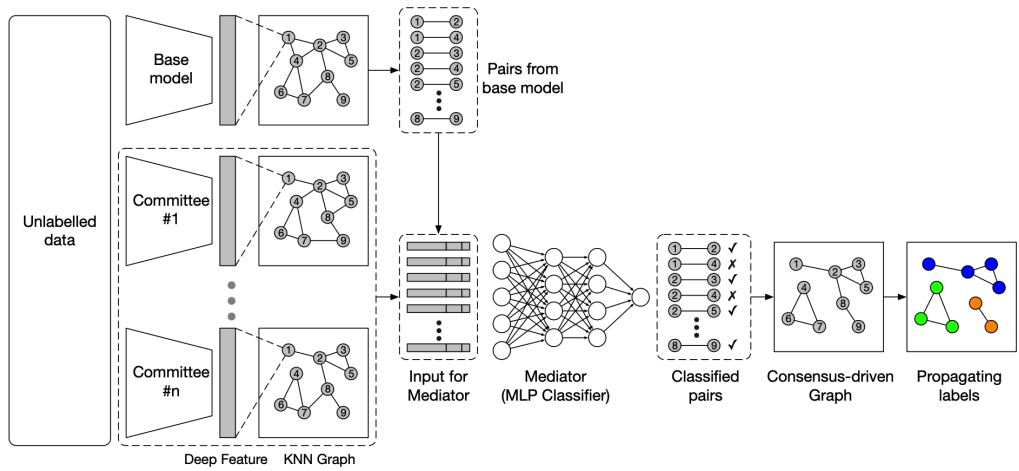
顾名思义，QBC算法实际上建立了一个“模型委员会”，里面存在多个独立的模型。在完成监督学习后，每个独立模型会对未标记数据进行一轮“质询”，也即inference。其中有些数据委员们的意见一致，证明这些数据无意义；另一些数据委员们的意见分歧很大，证明这些数据有价值。

量化模型预测结果的“分歧”可以使用信息熵算法。

$$x_{V,E}^* = \operatorname{argmax}_{x \in \mu} - \sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C}$$

其中模型数C一般取到3以上，效果就很好了。

## 3. 系统pipeline



**Fig. 1: Consensus-Driven Propagation.** We use a base model and committee models to extract features from unlabeled data and create k-NN graphs. The input to the mediator is constructed by various local statistics of the k-NN graphs of the base model and committee. Pairs that are selected by the mediator compose the “consensus-driven graph”. Finally, we propagate labels in the graph, and the propagation for each category ends by recursively eliminating low-confidence edges.

系统的pipeline如上图所示。总结来说分为如下的几个步骤。

- 首先利用supervised learning进行base model和committee的训练作为初始化。
- 将一系列unlabeled data输入base model和所有committee，得到一系列的KNN Graph。
- 将KNN Graph输入后续的mediator，调停者将调停委员会的意见，将至融合为一个统一的输出，也即consensus-driven graph。
- 基于consensus-driven graph给每个unlabeled data以伪标注，并且输入下阶段的multi-task网络进行训练，提升网络性能。

## 4. Committee及其IO

所谓的Committee其实就是一系列不同的backbone。对于每个输入的unlabeled data，每个committee都会输出对应的feature map。

最后，将每个数据的feature map展开为向量，用cosine similarity来度量数据间两两的相似性。以数据本身为结点，以余弦相似性为边的权。

对每个结点只保留相似性最高的k条边，就构成了committee的基本输出：KNN Graph。

## 5. Mediator及其IO

为了方便表述，我们引用原文当中的符号如下：

- $B$  for base model and  $C_i, i \in \{1, 2, \dots, N\}$  for all  $N$  committee members.
- $D_u$  for unlabeled dataset.  $Z$  for feature domain.
- $\mathcal{F}_{B/C_i}(D_l)$  is a function for  $D_l \mapsto Z$  where  $D_l \in D_u$ .

显然，committee直接输出的KNN Graph没办法漂亮地作为一个“委员会意见”，最多是一个“委员会研讨会议记录”。

在操作中，我们采用如下的方式，将KNN Graph中关键信息抽离作为“意见”输入到调停者中，以便形成统一意见。

1. Relationship，也即两个数据是否是k近邻。

$$R_{C_i}^{(n_0, n_1)} = \begin{cases} 1 & \text{if } (n_0, n_1) \in \xi(\mathcal{G}_{C_i}) \\ 0 & \text{otherwise} \end{cases} \quad i = 1, 2, 3, \dots \quad (1)$$

where  $\mathcal{G}_{C_i}$  is the k-NN graph of  $i^{th}$  committee model and  $\xi$  denotes all edges of a graph.

2. Affinity，也即两个数据具体的“密切程度”。

$$A_{C_i}^{(n_0, n_1)} = \cos(\langle \mathcal{F}_{C_i}(n_0), \mathcal{F}_{C_i}(n_1) \rangle), \quad i = 1, 2, \dots, N \quad (2)$$

其中， $\cos(a, b)$ 代表余弦相似度。其定义方式是两个向量高维空间夹角的余弦值。

两个向量有相同的指向时，余弦相似度的值为1；两个向量夹角为90°时，余弦相似度的值为0。

余弦相似性最常用于高维正空间。例如在信息检索中，每个词项被赋予不同的维度，而一个文档由一个向量表示，其各个维度上的值对应于该词项在文档中出现的频率。余弦相似度因此可以给出两篇文档在其主题方面的相似度。

3. local structure，也即每个数据的k近邻的统计特征，刻画这个数据的局部统计特性。

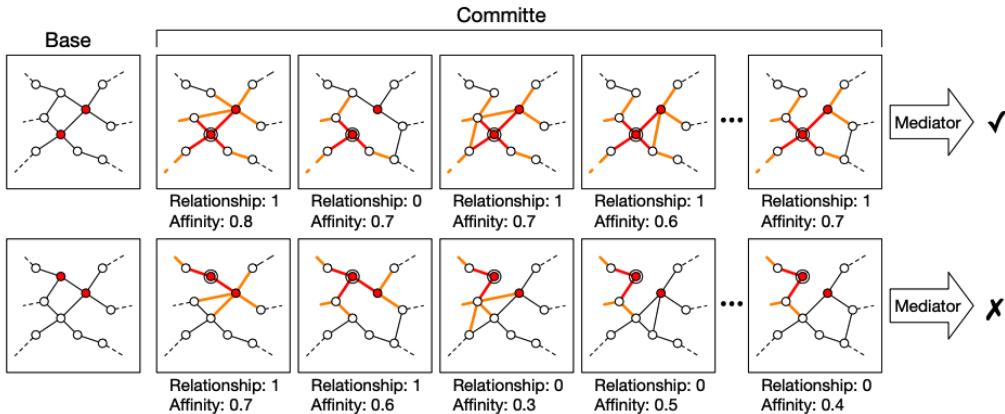
定义式子的K邻域空间相似度集合如式(3)。

$$D_{C_i}^x = \{\cos(\langle \mathcal{F}_{C_i}(n_0), \mathcal{F}_{C_i}(n_1) \rangle), k = 1, 2, \dots, K\}, i = 1, 2, \dots, N \quad (3)$$

定义了K邻域后，取每个committee针对每对结点K邻域的均值和方差，作为mediator的local structure输入。

$$I_{D_{mean}} = (\dots E(D_{C_i}^{n_0}), \dots, E(D_{C_i}^{n_1}) \dots), i = 0, 1, \dots, N \quad (4)$$

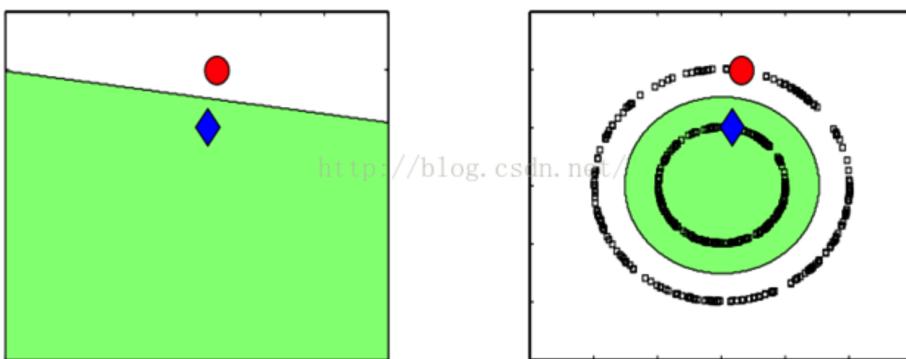
$$I_{D_{var}} = (\dots \sigma(D_{C_i}^{n_0}), \dots, \sigma(D_{C_i}^{n_1}) \dots), i = 0, 1, \dots, N \quad (5)$$



**Fig. 2: Committee and Mediator.** This figure illustrates the mechanisms of committee and mediator. The figure shows some sampled nodes in different versions of graphs brought by the base model and the committee. In each row, the two red nodes are candidate pairs. The pair in the first row is classified as positive by the mediator, while the pair in the second row is considered as negative. The committee provides diverse opinions on “relationship”, “affinity”, and “local structure”. The “local structure” is represented as the distribution of first-order (red edges) and second-order (orange edges) neighbors. Note that the figure only shows the “local structure” centered on one of the two nodes (the node with double circles).

## 6. 伪标注和联合训练

这里插播一则传统Label Propagation算法。



可以看到，如果只有两个labeled数据，那么进行最优分类的时候结果可能非常差；但如果还有一系列的无标签数据作为辅助，则能够给出漂亮的分类结果。

经典的标签传播（Label Propagation, LP）算法十分简单，首先可以得到一个结点（数据）的邻接矩阵，代表不同结点之间的相关性。然后想办法把其中labeled结点的label“传播”到全图，也即给unlabeled结点打上标签。

我们设 $w_{ij}$ 为邻接矩阵的元素。其值越大，代表两个结点越相似。

由此我们可以定义转移概率矩阵如下：

$$P_{ij} = P(i \rightarrow j) = \frac{w_{ij}}{\sum_{k=1}^n w_{ik}}$$

并且我们定义label矩阵 $Y_L \in \mathcal{M}_{L \times C}$ 和 $Y_U \in \mathcal{M}_{U \times C}$ , 并合并记作 $F = [Y_L; Y_U]$ 。其中C为类数量; L为labeled数据数量, U为unlabeled数据数量。

实际上,  $F_{ij}$ 代表第i个数据gt标签为j类的概率。显然对于 $Y_L$ 而言, 每行应该是 $(0, \dots, 1, \dots, 0)$ 的形式, 且是固定不变的。而 $Y_U$ 则是我们LP算法需要求解的对象。

利用如下的迭代, 能够合理地给出 $Y_U$

- 执行传播  $F = PF$
- 重置样本标签为真值  $F_L = Y_L$
- 重复上两步直到F收敛

实际上为了更好地理解传播过程, 我们拆解之, 写出下式

$$F_{ij} = \sum_{k=1}^N P_{ik} F_{kj}$$

从含义上, 就是第i个结点标签为j的概率, 应该是结点k标签为j的概率乘以结点k转移到结点i的概率, 这样的k应该遍历所有可行结点, 概率自然应该求和。

然而这样的求解过程存在计算的冗余, 考虑到 $Y_L$ 的不变性, 利用分块矩阵将之剔除, 可以得到下式:

$$f_U \leftarrow P_{UU} f_U + P_{UL} Y_L$$

其中算法收敛性是可以证明的, 而且会收敛到一个显式解:

$$f_U = (I - P_{UU})^{-1} P_{UL} Y_L$$

迭代法高效, 显式解需要求逆但是准确, 各有千秋。

本文采取的方式和传统的LP不同。其原因是在在海量人脸数据中, 不能指望新数据和老数据的标签相同。因此本文的方法需要突破这个障碍。

文章其实采用了对生成的consensus-driven graph进行简单聚类的方式进行伪标注。也即给定一个数据量阈值。在consensus-driven graph中, 凡是连通子图结点数小于该阈值的, 一律打上相同的伪标签。如果连通子图结点数大于该阈值, 则去掉其中的弱连接关系, 直到新的连通子图结点数小于该阈值为止。

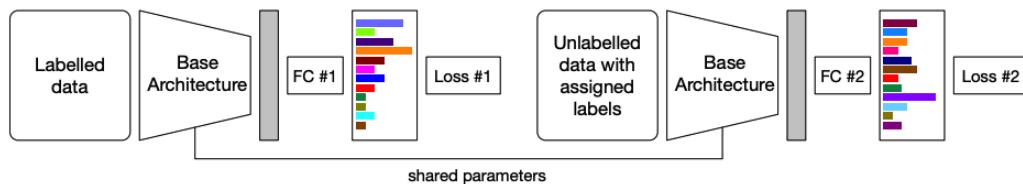


Fig. 3: Model updating in multi-task fashion. The weights of two CNNs are shared. “FC” denotes fully-connected classifier. In our experiments we use weighted Cross-Entropy loss as the objective.

整体的标签-伪标签联合训练pipeline如上图。两个base architecture结构和参数一模一样, 而两个loss用一个超参数(weight balance) 加权后求和。

# Results

## 1. 实验条件

- training set: MS-Celeb-1M数据集
- testing set: MegaFace数据集+InsightFace数据清洗
- Committee setup:

Table 1: Performance and the number of parameters of the base model and the committee members.

	Architecture	MegaFace	IJB-A	Parameters
Base	Tiny NASNet-A	<b>61.78</b>	<b>75.87</b>	20.1M
Committee	VGG16	50.22	70.75	75.6M
	ResNet18	51.48	69.23	23.5M
	ResNet34	52.44	72.52	33.6M
	Inception V3	52.82	75.53	33.0M
	ResNet50	56.16	73.21	36.3M
	ResNet101	57.87	74.52	55.3M
	Inception-ResNet V2	58.68	75.13	66.1M
	DesNet121	60.77	69.78	28.9M
Ensemble	(multiple)	69.86	76.97	-

## 2. 实验结果

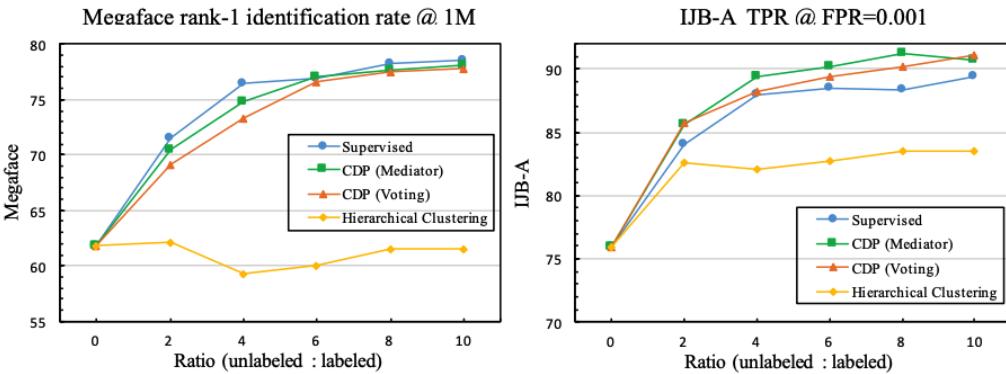


Fig. 4: Performance comparison on MegaFace identification task and IJB-A verification task with different ratios of unlabeled data added to one portion of labeled data. CDP is proven to 1) obtain large improvements over the lower bound (ratio of unlabeled:labeled is 0:1); 2) surpass the clustering method by a large margin; 3) obtain competitive or even higher results over the fully-supervised counterpart.

从结果来看，CDP的结果比Fully Supervised的结果不差太多。如introduction中所说，MegaFace上CDP的训练得到了78.18%的精度。而无label精度61.78%；全label精度78.52%

CDP相比于无label训练提升了16.4%，相比于全label只下降了0.34%。



Fig. 6: This figure shows two groups of faces in the unlabeled data. All faces in a group has the same identity according to the original annotations. The number on the top-left corner of each face is the label assigned by our proposed method, and the faces in red boxes are discarded by our method. The results suggest the high precision of our method in identifying persons of the same identity. Interestingly, our method is robust in pinpointing wrongly annotated faces (group 1), extremely low-quality faces (e.g., heavily blurred face, cartoon in group 2), which do not help training. See supplementary materials for more visual results.

另一件有意思的事情是，本文的方法相比于直接用标注的数据进行训练，能够剔除一些极端无意义的样本，例如错误的标注、夸张的卡通、极度模糊或者黑暗的数据。这些数据对网络学习提取信息没有太大的帮助，反而会使得网络

confused, 性能下降。

## Insights

总体而言，本文提出了一整套针对人脸的半监督学习策略，能够利用少量标注数据和海量无标注数据联合训练，得到的效果和十倍标注数据集的训练效果相当。

另外，文章的方法其实对于人脸数据清洗、人脸数据主动学习等实用课题也很有帮助。