Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心


# Active Self-Paced Learning

秦枭喃　硕士

2019年02月24日

# 内容提要

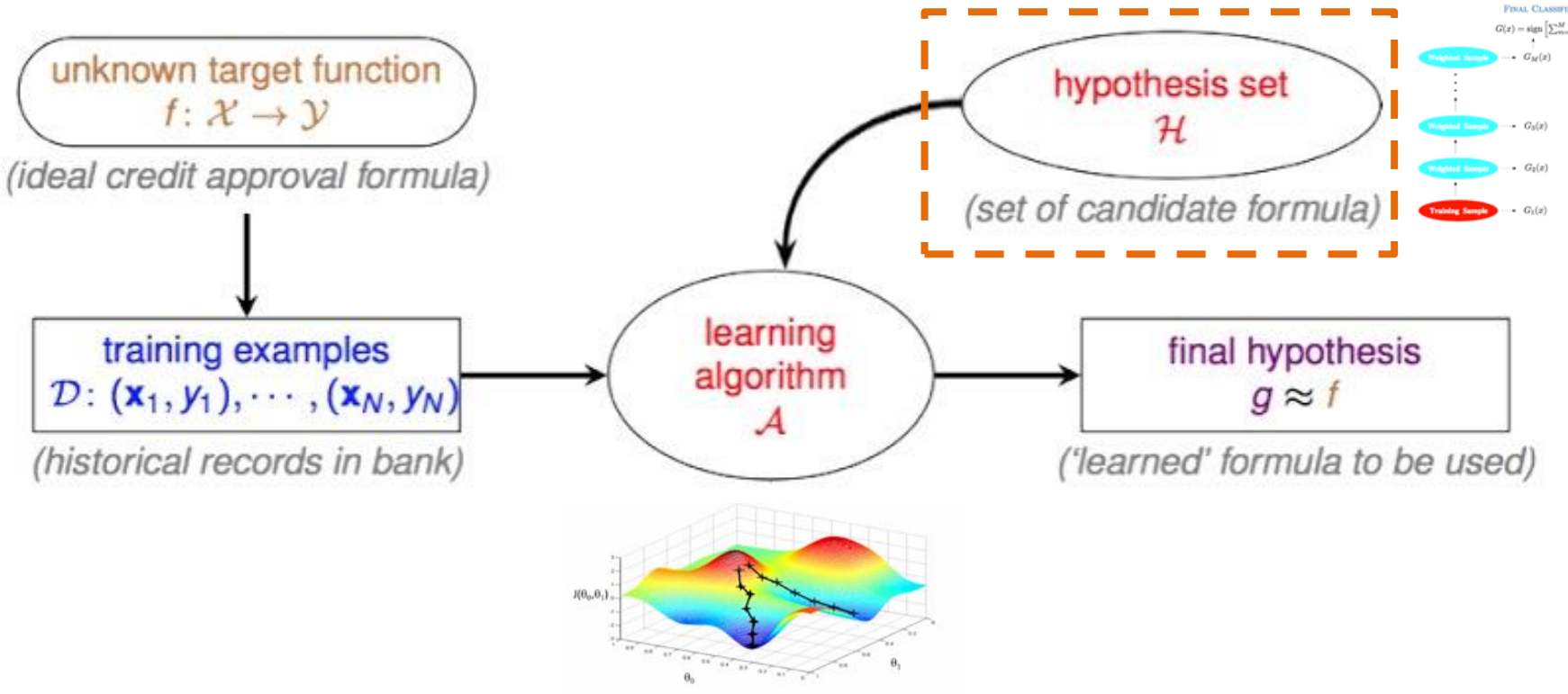
- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 模型总结
- 参考文献

# 预期收获

- 预期收获
  - 1. 回顾主动学习与自步学习的基本原理
  - 2. 理解自步学习与主动学习结合的模型框架

# 背景介绍

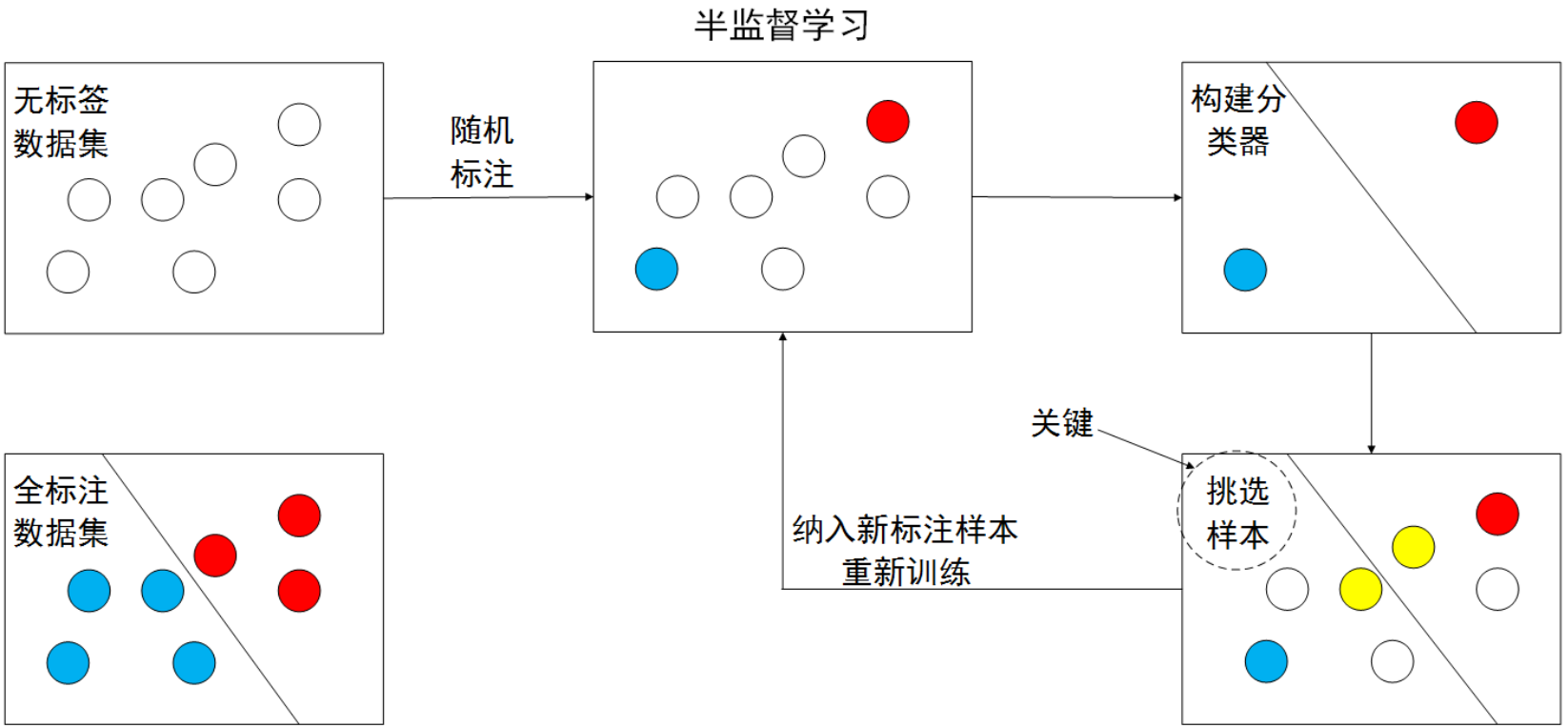
- 机器学习基础架构[1]




[1] 林轩田，机器学习基石

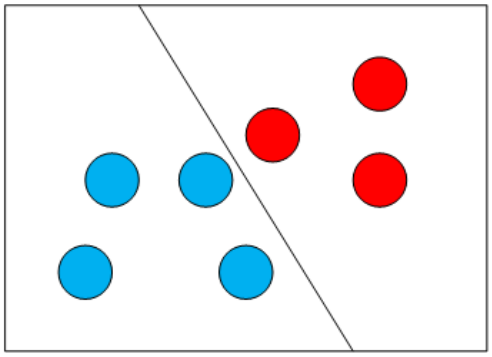
# 背景简介

- 主动学习：如何使用尽可能少的标注数据集训练一个模型，这个模型的性能可以达到一个由大量的标注数据集按照普通方法训练得到的模型的性能。

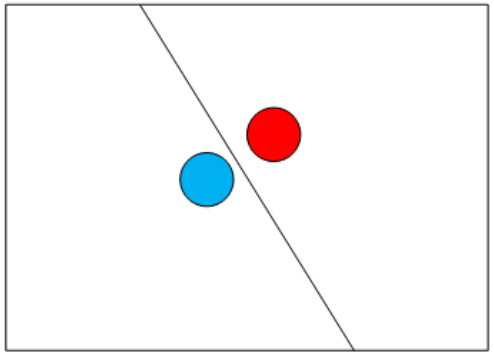


# 背景简介

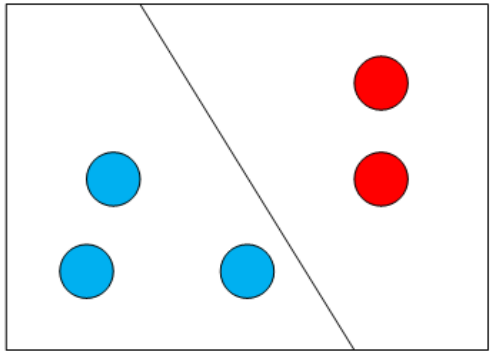
- 主动学习：如何使用尽可能少的标注数据集训练一个模型，这个模型的性能可以达到一个由大量的标注数据集按照普通方法训练得到的模型的性能。



全样本



低置信度样本



高置信度样本

# 背景简介

- 传统的主动学习方法常常强调低置信度高信息量的样本的挑选而忽略大部分高置信度样本的标注和使用，若使用模型直接对数据进行预测标注并纳入后续训练，容易导致模型陷入局部最优的状态。
  - 如何有效标注高置信度样本?
  - 如何利用高置信度样本帮助模型挑选低置信度样本?

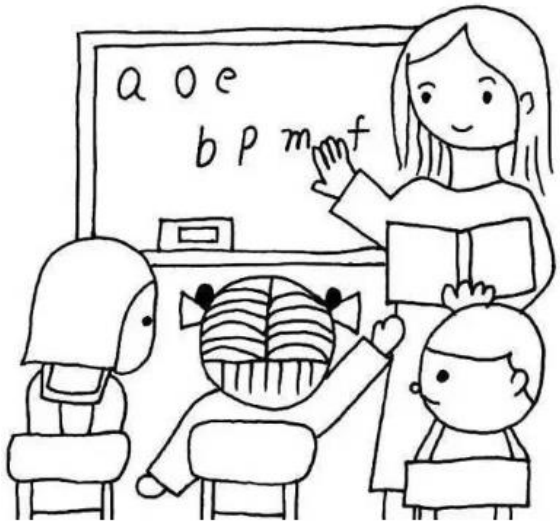
# 基本概念

- 自步学习：模拟人的认知，根据先验知识赋予样本学习先后顺序。
  - 帮助解决非凸优化问题
  - 具有较强的鲁棒性
  - 通过赋予样本权重有效区分高置信度与低置信度样本

# 算法原理

# 算法原理

- Active Self-Paced Learning（ASPL）
  - **功能：将主动学习与自步学习相结合，使模型能够自动标注新实例并将低置信度的部分纳入专家重新认证中，减少标注成本。**




老师上课教学——人工标注

学生课后自学——自动标注

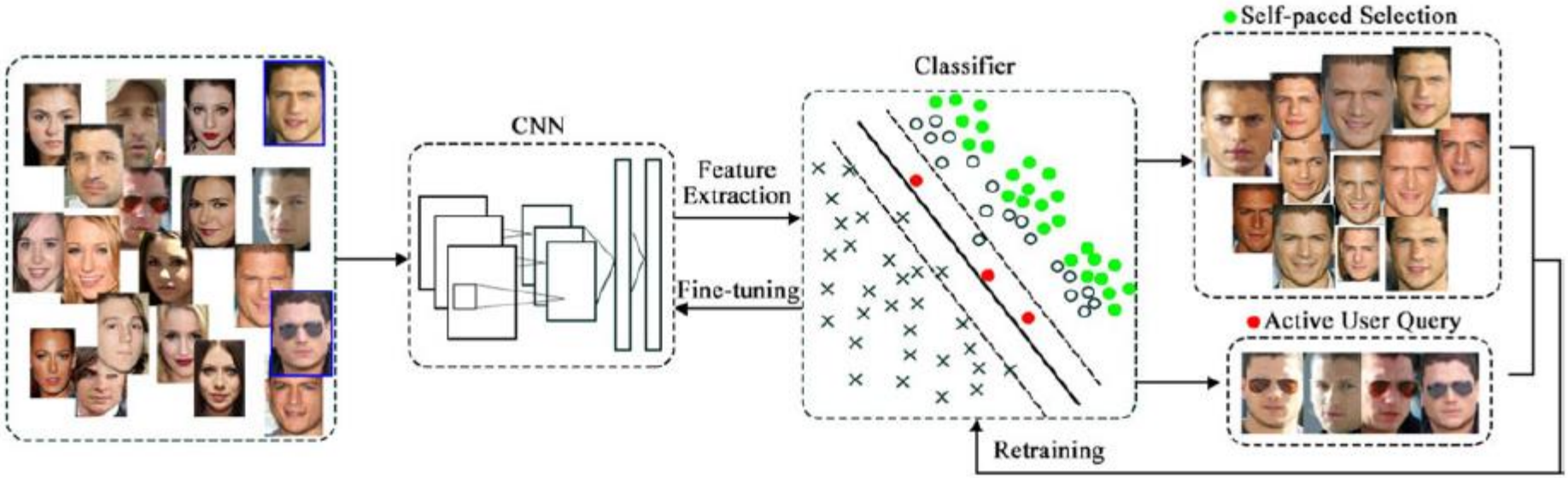
# 算法原理

| | |
|---|---|
| T | 基于无标签数据，以最小标注代价完成分类器学习 |
| I | 无标签样本，预训练模型 |
| P | For {<br>1. 自步学习进行自动标注<br>2. 主动学习进行人工标注<br>3. 更新模型<br>} |
| O | 高性能分类器 |

| | |
|---|---|
| P | 减少模型对人工标注样本需求 |
| C | 具备大量无标签样本，存在可正确标注样本的专家 |
| D | 如何实现数据的自动标注和最具信息量样本选取 |
| L | IEEE A类会议 |

# 算法原理

- ASPL的结构框图




- 模型分为四部分：CNN，Classifier，Active User Query， Self-paced Selection

# 算法原理

- ASPL算法的基本流程

**Algorithm 1.** The Sketch of ASPL Framework

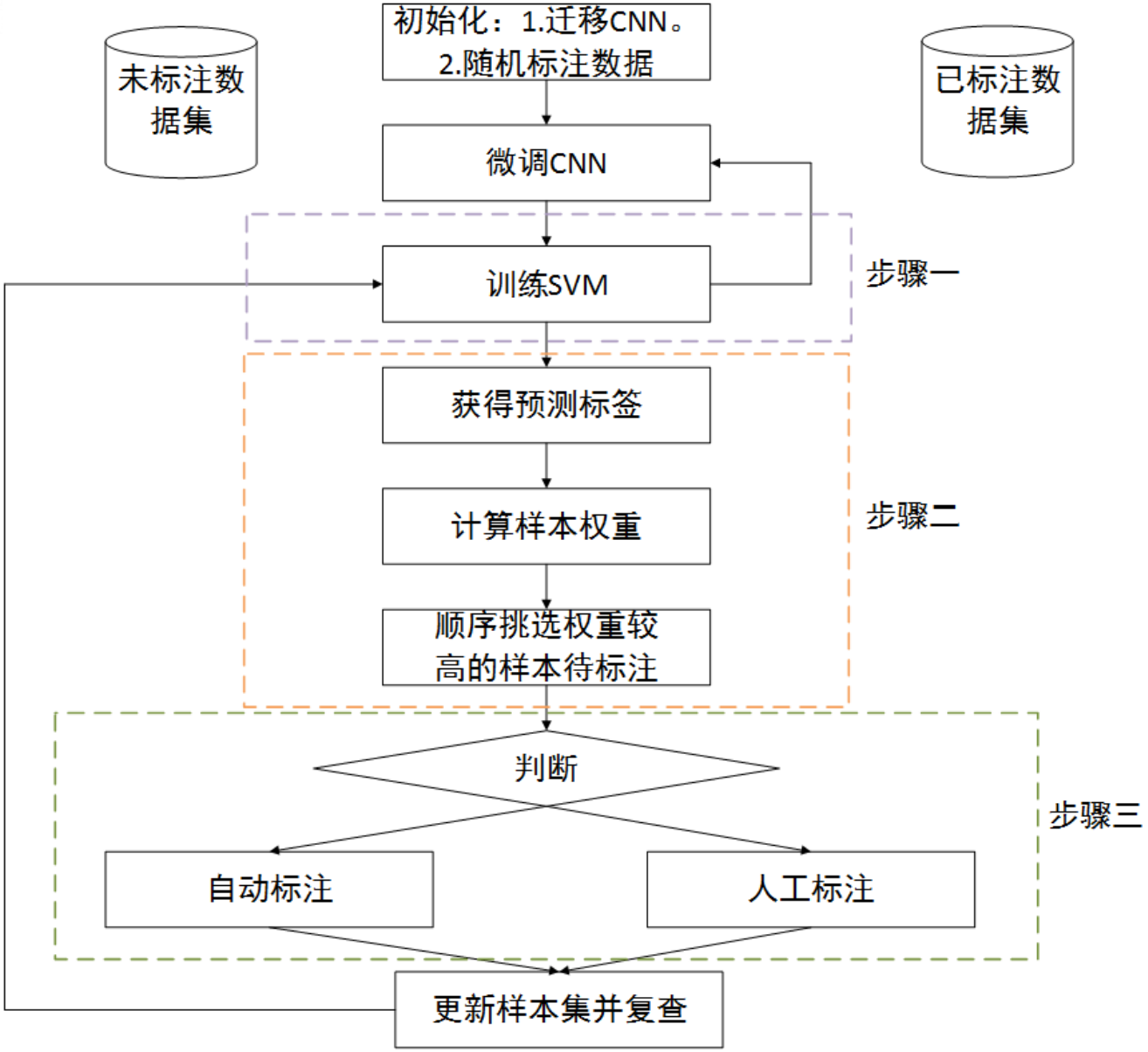
**Input:** Input dataset $\{\mathbf{x}_i\}_{i=1}^n$
**Output:** Model parameters $\mathbf{w}$, $\mathbf{b}$

Use pre-trained CNN to extract feature representations of $\{\mathbf{x}_i\}_{i=1}^n$. Initialize multiple annotated samples into the curriculum $\mathbf{\Psi}^{\boldsymbol{\lambda}}$ and corresponding $\{\mathbf{y}_i\}_{i=1}^n$ and $\mathbf{v}$. Set an initial pace parameter $\boldsymbol{\lambda} = \{\lambda^0\}^m$.
**while** not converged **do**
  Update $\mathbf{w}, \mathbf{b}$ by one-vs-all SVM
  Update $\mathbf{v}$ by the SPL via Eqn. (7)
  Pseudo-label high-confidence samples $\{\mathbf{y}_i\}_{i\in\mathcal{S}}$ by the reranking via Eqn. (8)
  Update the unclear class set $U$
  Verify the annotated samples by AL.
  Update low-confidence samples $\{\mathbf{y}_i, \mathbf{\Psi}_i^{\lambda}\}_{i\in\phi}$ by the AL
  **if** $u$ unseen classes have labeled,
    Handle $u$ new classes via the steps in Section 4.1
    Go to the step 2
  **end if**
  In every $T$ iterations:

- Update $\{\mathbf{x}_i\}_{i=1}^n$ through fine-tuning CNN
- Update $\boldsymbol{\lambda}$ according to Eqn. (10)

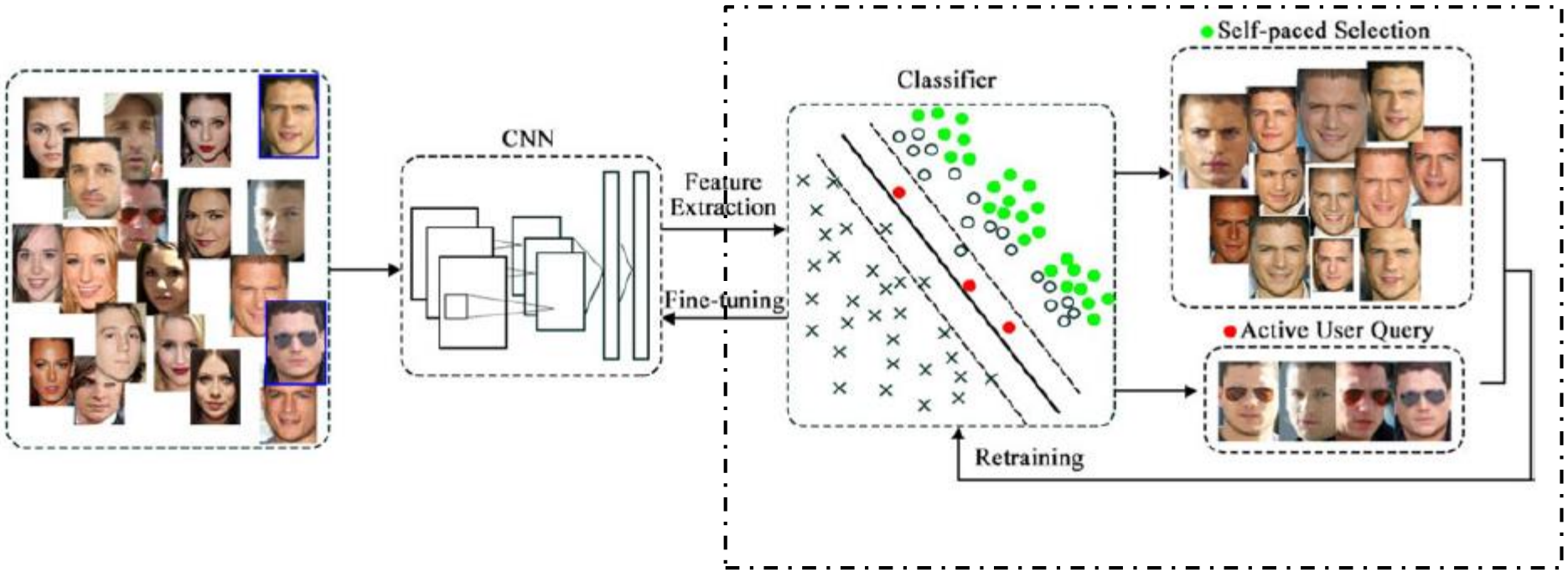
**end while**
**return** $\mathbf{w}, \mathbf{b}$;

# 算法原理


初始化：1.迁移CNN。
2.随机标注数据
未标注数据集
已标注数据集
微调CNN
训练SVM
步骤一
获得预测标签
计算样本权重
顺序挑选权重较高的样本待标注
步骤二
判断
步骤三
自动标注
人工标注
更新样本集并复查

# 算法原理

- ASPL的结构框图和公式



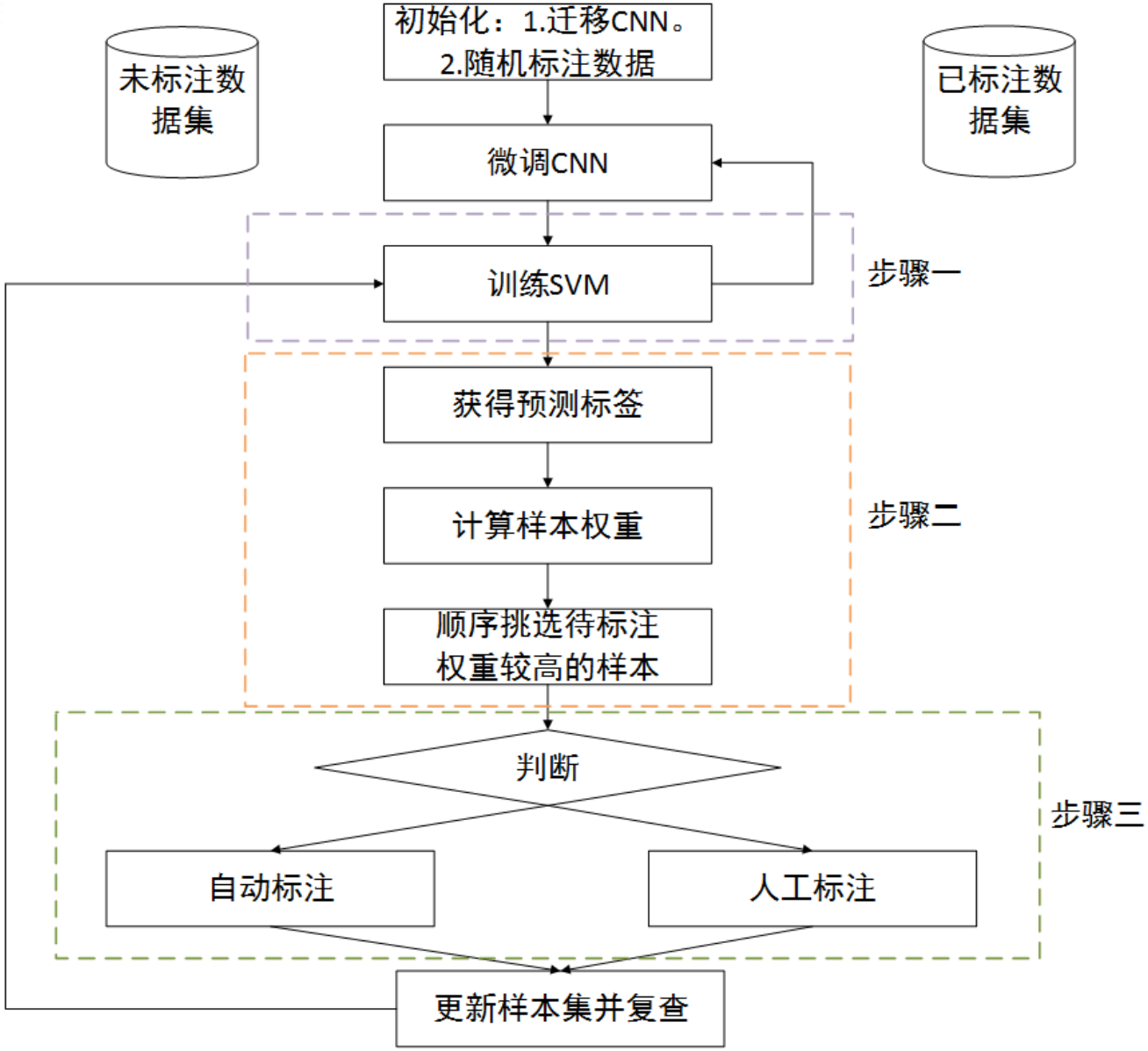

$$\min_{\{\mathbf{w},\mathbf{b},\mathbf{v},\mathbf{y}_i\in\{-1,1\}^m, i\notin\Omega^{\lambda}\}} \sum_{j=1}^{m} \frac{1}{2}\|\mathbf{w}^{(j)}\|_2^2 + C\cdot L\left(\mathbf{w}^{(j)}, b^{(j)}, \mathcal{D}, \mathbf{y}^{(j)}, \mathbf{v}^{(j)}\right) + f\left(\mathbf{v}^{(j)}; \lambda_j\right)$$
$$s.t. \quad \mathbf{v}\in\Psi^{\lambda},$$

# 算法原理

$$\min_{\{\mathbf{w},\mathbf{b},\mathbf{v},\mathbf{y}_i\in\{-1,1\}^m, i\notin\Omega^{\lambda}\}} \sum_{j=1}^{m} \frac{1}{2}\|\mathbf{w}^{(j)}\|_2^2 + C\cdot L\left(\mathbf{w}^{(j)}, b^{(j)}, \mathcal{D}, \mathbf{y}^{(j)}, \mathbf{v}^{(j)}\right) + f\left(\mathbf{v}^{(j)}; \lambda_j\right)$$
$$s.t. \quad \mathbf{v}\in\Psi^{\lambda},$$

- $\frac{1}{2}\|\mathbf{w}^{(j)}\|_2^2$：SVM里的正则项
- $C\cdot L\left(\mathbf{w}^{(j)}, b^{(j)}, \mathcal{D}, \mathbf{y}^{(j)}, \mathbf{v}^{(j)}\right)$：SVM的损失函数，hinge loss
- $f\left(\mathbf{v}^{(j)}; \lambda_j\right)$：自步学习的正则项
- 每次循环中将顺序进行三次交替优化，直到获得最终分类器的参数。

# 算法原理


初始化：1.迁移CNN。
2.随机标注数据
未标注数据集
已标注数据集
微调CNN
训练SVM
步骤一
获得预测标签
计算样本权重
顺序挑选待标注权重较高的样本
步骤二
判断
自动标注
人工标注
步骤三
更新样本集并复查

# 算法原理

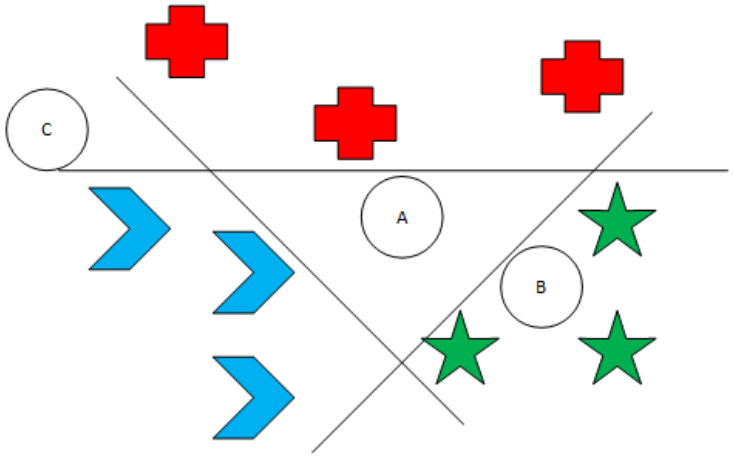
- 步骤1：构建分类器
  - 输入：已标注的样本数据，固定v，C，x，y
  - 输出：优化更新分类器的w，b
  - 利用已标注的样本构建分类模型SVM

$$\min_{\mathbf{w},\mathbf{b}} \sum_{j=1}^{m} \frac{1}{2} \|\mathbf{w}^{(j)}\|_2^2 + C \sum_{i=1}^{n} v_i^{(j)} l\left(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\right)$$

# 算法原理

- 步骤2：挑选待标注样本集
  - 输入：未标注样本数据，固定w，b，x，y（预测标签）
  - 输出：未标注数据的样本权重v和待标注样本集S
  - 使用SVM预测未标注数据，计算未标注数据的损失值，通过自步学习的方法赋予样本权重，并挑出待标注样本

$$\min_{\mathbf{v}\in[0,1]} \sum_{j=1}^{m} C \sum_{i=1}^{n} v_i^{(j)} l\left(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\right) + f\left(\mathbf{v}^{(j)}; \lambda_j\right)$$



# 算法原理

- 步骤2：挑选待标注样本
  - 输入：未标注样本数据，固定w，b，x，y（预测标签）
  - 输出：未标注数据的样本权重v和待标注样本集S
  - 问题：由于SVM预测一些标签会使模型性能变差，需要判断修正。

$$\min_{\mathbf{v}\in[0,1]} \sum_{j=1}^{m} C \sum_{i=1}^{n} v_i^{(j)} l\left(\mathbf{w}^{(j)}, b^{(j)}; \mathbf{x}_i, y_i^{(j)}\right) + f\left(\mathbf{v}^{(j)}; \lambda_j\right)$$

$$f(\mathbf{v}^{(j)}, \lambda_j) = \lambda_j \left( \frac{1}{2} \|\mathbf{v}^{(j)}\|_2^2 - \sum_{i=1}^{n} v_i^{(j)} \right) \longrightarrow$$

线性软加权正则项

$$v_i^{(j)} = \begin{cases} -\frac{C\ell_{ij}}{\lambda_j} + 1, & C\ell_{ij} < \lambda_j \\ 0, & \text{otherwise}, \end{cases}$$

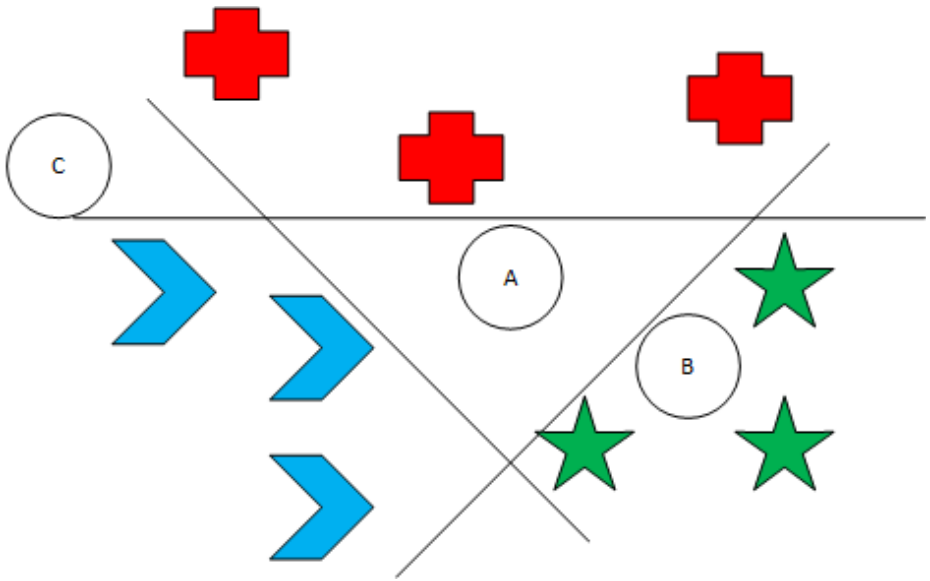
# 算法原理

- 步骤3：标注样本数据
  - 输入：待进行标注的样本集S
  - 输出：此样本集的所有样本标签y
- 通过使用SVM分析每个样本属于每一类样本的损失值情况，来优化目标函数。

$$\min_{\mathbf{y}_i \in \{-1,1\}^m, i \in \mathcal{S}} \sum_{i=1}^{n} \sum_{j=1}^{m} v_i^{(j)} \ell_{ij}$$

$$\text{s.t.}, \sum_{j=1}^{m} |y_i^{(j)} + 1| \leq 2.$$

# 算法原理




$$\min_{\mathbf{y}_i \in \{-1,1\}^m, i \in \mathcal{S}} \sum_{i=1}^{n} \sum_{j=1}^{m} v_i^{(j)} \ell_{ij}$$

$$\text{s.t.}, \sum_{j=1}^{m} |y_i^{(j)} + 1| \leq 2.$$

**Theorem 1.**

*a)* *If* $\forall j \in M$, $\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)} < 0$, *Eqn. (8) has a solution:*

$$y_i^{(j)} = -1, \quad j = 1, \ldots, m;$$

*b)* *When* $\forall j \in M$ *except* $j = j^*$, $\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)} < 0$, *i.e.,* $v_i^{(j^*)} \ell_{ij^*} > 0$, *then Eqn. (8) has a solution:*

$$y_i^{(j)} = \begin{cases} -1, & j \neq j^* \\ 1, & j = j^* \end{cases};$$

*c)* *Otherwise, Eqn. (8) has a solution:*

$$y_i^{(j)} = \begin{cases} -1, & j \neq j^* \\ 1, & j = j^* \end{cases},$$

*where*

$$j^* = \arg \min_{1 \leq j \leq m} v_i^{(j)} \left( \ell_{ij} - \left( 1 + (\mathbf{w}^{(j)T}\mathbf{x}_i + b^{(j)}) \right)_+ \right). \qquad (9)$$

# 模型中的小细节

- 样本权重v的设置：
  - 人工标注下的样本权重v的取值为[1]。
  - 自动标注下的样本权重v的取值为[0,1]。

- 复查标记样本：
  - 在实际生活中由于存在人工标记错误或是系统自动标注错误的情况，每次循环的最后需要在标注数据集中挑出预测分数最低的n个样本进行人工复查。

- 新类别的出现：
  - 若在循环中出现新类别的样本，将采用KNN算法，找到相似数据进行标注，以提高新类别标注样本的数量。

# 优劣分析

# 优劣分析

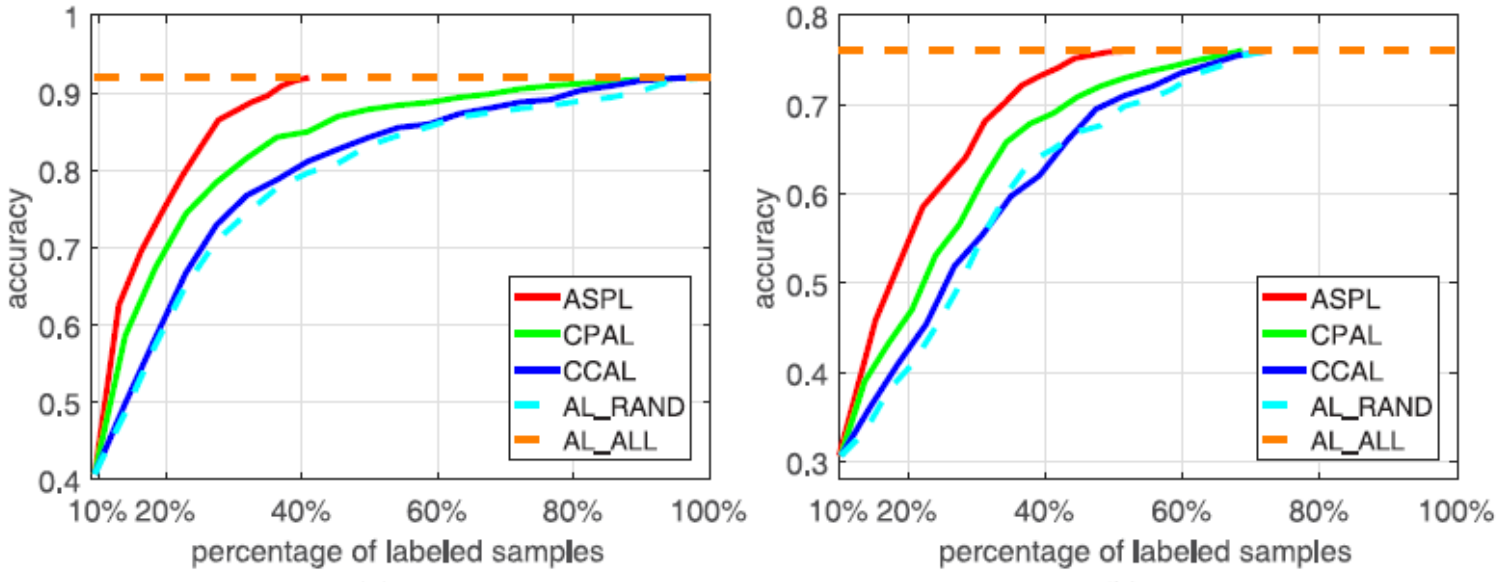
- 数据集

| 数据集 | 图片数量 | 人物类别 | 图像/类别 |
| --- | --- | --- | --- |
| CACD | 56138 | 500 | 79~306 |
| CASIA-WebFace-Sub | 181901 | 925 | 100~804 |

- 对比方法
  - CPAL：主动学习中一种方法（2012）。
  - CCAL：主动学习中一种方法（2002）。
  - AL_RAND：随机挑选样本进行人工标注。
  - AL_ALL：全样本标注下的模型训练。

# 优劣分析

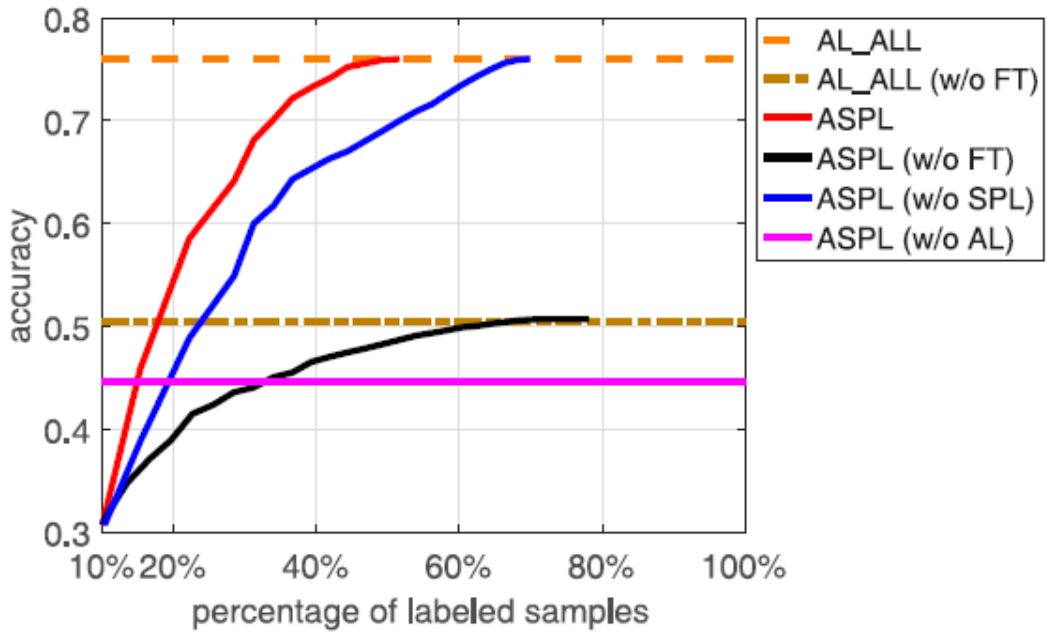
- 实验结果：




– 在两个数据集中，ASPL模型到达性能上限所需标注样本最少，优于CPAL和CCAL算法。

# 优劣分析

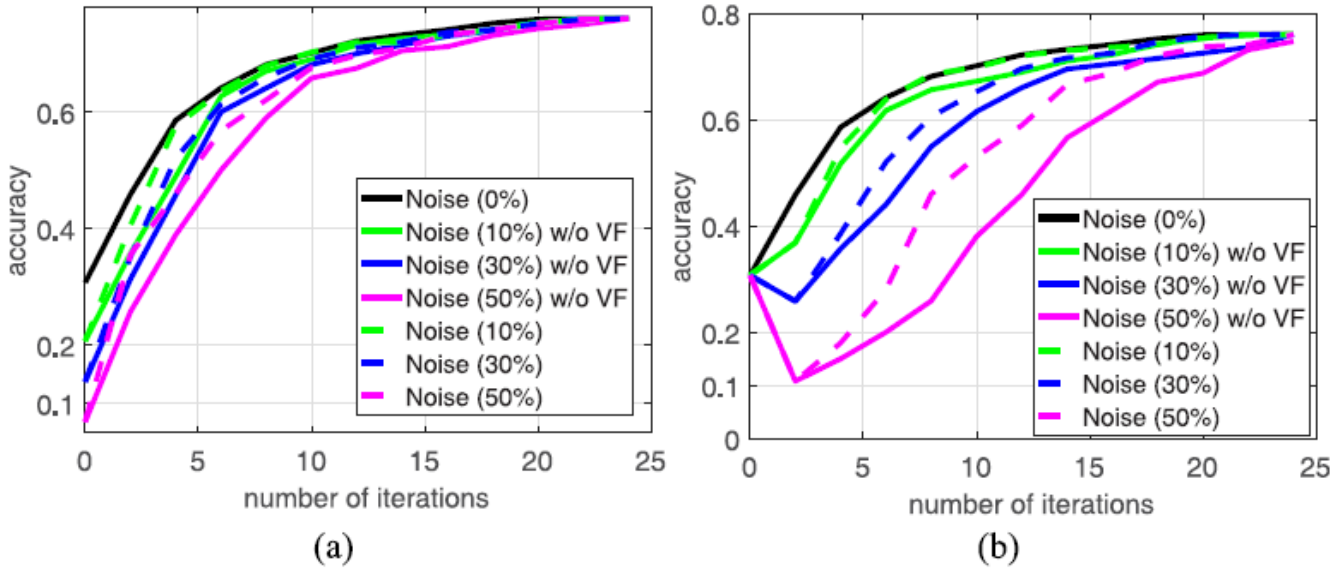
- 成分分析：分析不同模块对性能的贡献




  - 微调CNN，自动标注样本，都有助于减少标注数据，提高模型准确率。
  - 自步学习通过自动利用大多数高置信度样本进行特征学习，显著提高识别准确度并减少注释样本数量。

# 优劣分析

- 鲁棒性分析：测量模型抗噪声的能力




- 左边实验是在初始已标注数据中加入不同数量的噪声，右边实验是在第二次循环中加入不同数量噪声。
- 随着ASPL训练的进行，当迭代次数增加时，可获得与原始清洁数据相当的精确度。且人工复查已标记数据能够有效恢复噪声数据，证明此过程的有效性。

# ASPL总结

- 优点：
  - 将自步学习、主动学习、CNN相结合，有效的减少标注成本。
  - 具有较强的鲁棒性。
  - 对大/小数据集都适用，具有较强的实用性。

- 缺点：
  - 不适用于存在数据不平衡的数据集。
  - 模型固化，不易于泛化到大多数领域。

# 参考文献


- Active Self-Paced Learning for Cost-Effective and Progressive Face Identification
- 干货 | Active Learning: 一个降低深度学习时间，空间，经济成本的解决方案
  https://cloud.tencent.com/developer/article/1135260
- 主动学习年度进展|VALSE2018
  https://www.jiqizhixin.com/articles/2018-06-20-14

知人者智，自知者明。
胜人者有力，自胜者
强。知足者富。强行
者有志。不失其所者
久。死而不亡者，寿。

# 谢谢！