Sampling matters

1. Introduction

1.1 Motivation

在神经网络的应用中, 我们平时更多的是关注网络结构. 而在这里, 作者选择从分析另外的两个很重要的地方, loss函数以及sampling方案.

这个论文, 首先通过实验分析了sampling和loss的重要性比较, 通过的是控制变量法. 其次从几何角度等等对sampling的影响过程进行了分析.

并且分析显示, sampling方案的选择要比loss重要.

1.2 Task

这里主要将任务集中于图像处理上, 即, 通过深度学习将图像转换为有丰富语义的表征. 有以下的表征 .

- zero-shot learning
- visual search
- face recognition
- fine-grained retrieval

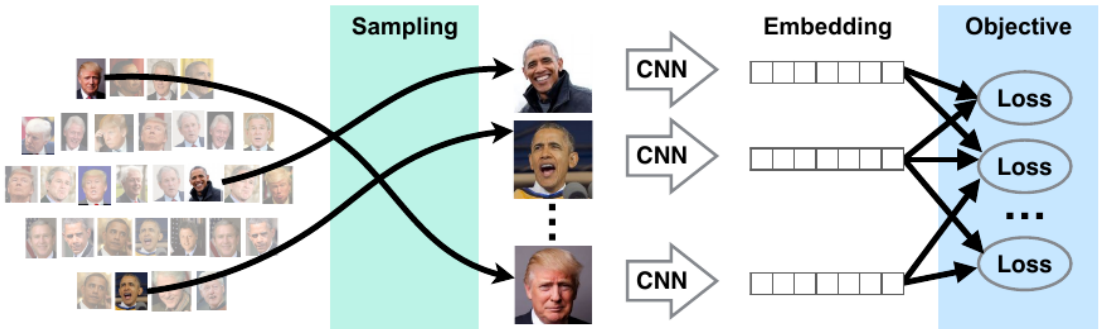


Figure 1: An overview of deep embedding learning: The first stage samples images and forms a batch. A deep network then transforms the images into embeddings. Finally, a loss function measures the quality of our embedding. Note that both the sampling and the loss function influence the overall training objective.

1.3 About Method - learn embedding

1) Insight

insight 还是相当简单的:

使得含有同一目标, 即标注为相近的目标图片的距离尽可能相近. 反之, 则使其远.

2) Insight in methods - loss function

- contrastive loss : positive->1, negative->0
- pairwise losses : 在之后介绍
- triplet loss : 不仅仅loss function相比于第一个有所改变, 还改变了positive和negative例子采样的方法.
- 等等

3) From here

我们知道影响loss function的有两个点:

- loss function
- sampling strategy

1.4 Conclusion

本文的总结预览:

- sample selection 的重要性大于 loss.
- 关于**采样策略**方面:
  - 分析了现有的采样策略的优劣
  - 提出一个新的采样策略:
    - 提出
    - 分析为什么好 : corrects the bias induced by the geometry of embedding space
    - 效果: a lower variance of gradients $\rightarrow$ stabilizes training
- 关于**loss function**方面:
  - 指出loss function也是有影响的
  - 分析了现在主流loss function的优劣
  - 提出了一个新的loss function:
    - It relaxes the loss, making it more robust
    - isotonic regression

2. Related Work

2.1 About loss function

1) 介绍了一些loss function

- triplet losses
  - introduction
  - 加上一些限制的版本 : PDDM , Histogram Loss
  - 使用更多example的版本 : n-pair loss , Lifted Structure

> defines constraints on all images in a batch

- other loss func:
  - Structural Clustering： optimizes for **clustering quality**
  - PDDM : proposes a new module to model **local feature structure**.
  - HDC： trains an ensemble to model examples of **different "hard levels"**

2) 作者观点

一个简单的 pairwise loss 其实就足够了, 关键的是sampling strategy.

2.2 About sampling strategy

- 一般的方法(随机采样)
- hard negative mining:

  introduction, 基本步骤如下:
  - 通过一些方法收集或者制作一批负样本 : Creat a batch of negative samples
  - 与positive examples一起训练Model :Train model on it
  - 利用这次训练中的fasle positive 去作为下一次的negative samples : Use fasle positive as negative sample

    > fasle positive : 应该是负样本但被检测为正样本的样本

  - 收集和制作新的负样本 : Creat new negative samples

  可以看到, 这个方法的特点是, negative examples经过模型预测的结果永远是偏向正的. 好处就是提高效率, 减少因为无效负样本花费的时间.
- **semi-hard negative mining:**

  described in chapter 3

3. Preliminaries

3.1 Notations

- Data point : $x_i \in \mathbb{R}^N$
- Deep network : $f : \mathbb{R}^N \to \mathbb{R}^D$
- the distance between two datapoints : $D_{ij} := ||f(x_i) - f(x_j)||$
- Euclidean norm : $||\ ||$
- Positive/negative value : $y_{ij} = 1/0$

3.2 loss

1) contrastive loss

$$\ell^{\text{contrast}}(i,j) := y_{ij} D_{ij}^2 + (1 - y_{ij}) \left[\alpha - D_{ij}\right]_+^2$$

2) triplet loss

$$\ell^{\text{triplet}}(a,p,n) := \left[D_{ap}^2 - D_{an}^2 + \alpha\right]_+ .$$

3.3 Compution effienicy

1) Risk minimization

风险最小化, 现在考虑两个loss function的计算量.假设dataset中有n个数据:

- For constractive loss : $O(n^2)$ pairs
- For triplet loss : $O(n^3)$ triplets

因此, 想要进行全部的学习是不现实的.

3.4 Convergences

1)问题所在

一旦模型开始收敛之后:

- most samples contribute in a minor way : 大部分样本都贡献很小
- very few of the negative margins are violated :大部分负样本失效

2) 可以利用sampling strategy加快收敛

- For contrastive loss : hard negative mining
- For triplet loss : semi-hard negative mining

> - hard negative mining in triplet loss also lead to a collaspe model:**all images have the same embedding.**
> - About **semi-hard negative mining**:
>
> $$n_{ap}^{\star} := \underset{n:D(a,n)>D(a,p)}{\operatorname{argmin}} D_{an},$$
>
> 其中, a,n,p分别是 anchor, negative, positive, anchor就是作为基底进行比较的那个像素框样本, negative是与anchor不一样类别的像素框. positive反之.
>
> n_{ap}^{*} : obtain a negative instance n within a batch
>
> 可以看到, 这个方法下, 负样本都是一批中最小的(但是要大于positive example)

4. Distance Weighted Margin-Based Loss

这一节分为两部分. 4.1, 4.2对以往的sampling strategy 和loss function进行了分析.

4.3,4.4提出了自己的新方案, 以及对其的分析.

4.1 About common sampling

1) Sampling strategy

sampling negative uniformly

2) Distance distribution

这里以constrastive loss视角进行分析, 将一个example pair(anchor和positive/negative)之间的距离作为变量, 下面是距离的分布:

$$q\left(d\right) \propto d^{n-2}\left[1-\frac{1}{4}d^{2}\right]^{\frac{n-3}{2}}.$$

> n : dimensions

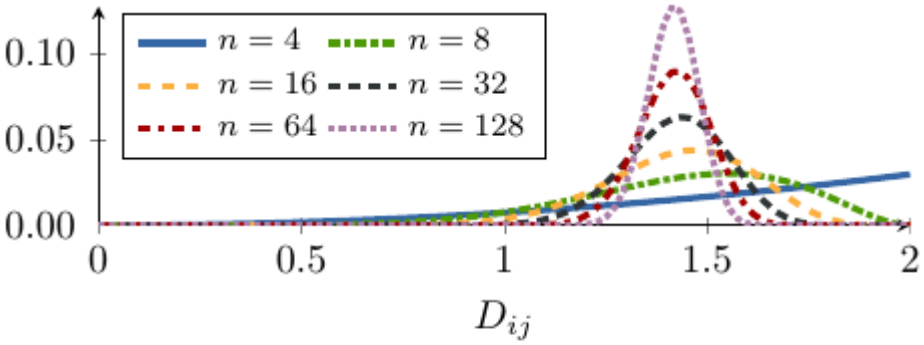


Figure 2: Density of datapoints on the D-dimensional unit sphere. Note the concentration of measure as the dimensionality increases — most points are almost equidistant.

4.2 About hard negative mining

1) loss 的微分

这里考虑triplet loss函数:

$$\ell^{\text{triplet}}(a,p,n) := \left[D_{ap}^{2} - D_{an}^{2} + \alpha\right]_{+}.$$

这个loss function关于 $f(x)$ 的微分是:

$$\partial_{f(x_n)}\ell^{(\cdot)} = \frac{h_{an}}{\|h_{an}\|}w(t)$$

> 对于绝对值的微分视为对函数 $f := \sqrt{x^2}$ 的微分即可. 利用链式法则即可解得.

2) nosie

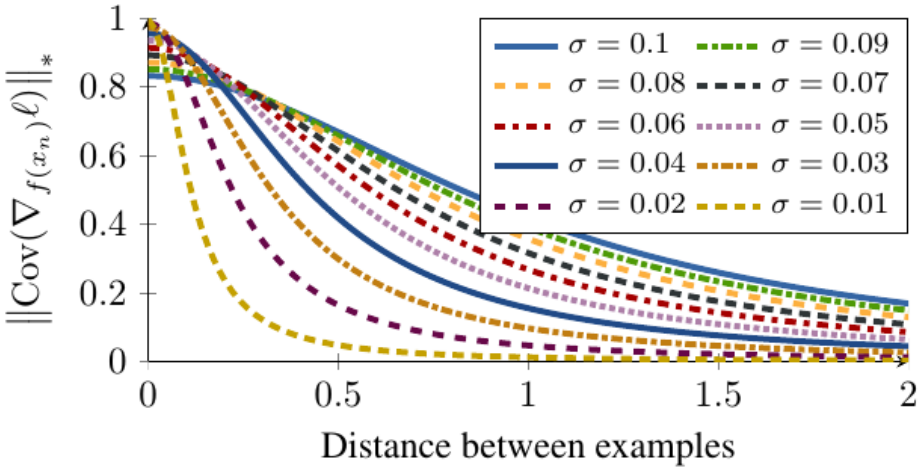
模型中多是存在噪音的, 比如 L2 正则项或是 dropout, 数据增强等等. 那么微分就会变成下面的式子:

$$\frac{h_{an}+z}{\|h_{an}+z\|}$$

如果h_{an}很小的话, 下降方向将会被噪音主导. 这就是这个loss不好的地方.

3) 定量分析影响

这里假设噪音服从一个均值为0, 方差为 σ 的分布, 下面是实验结果, 其中y轴是梯度的协方差矩阵的绝对值, 也就是梯度的方差:



(a) Variance of gradient at different noise levels.

这里y轴的值越大, 说明方差越大, 这样分布也就越平缓, 说明模型对自己下一步向那个方向下降非常没有自信. 但是并不是说, 越小越好, 因为越小学习越慢.

所以,

- y轴的值越小, 说明模型下一步下降方向越确定, 越趋近一个决定性算法
- y轴的值越大, 说明模型下一步下降方向越模糊, 越趋近一个随机算法.

4.3 Distance weighted sampling

1) New sampling methods

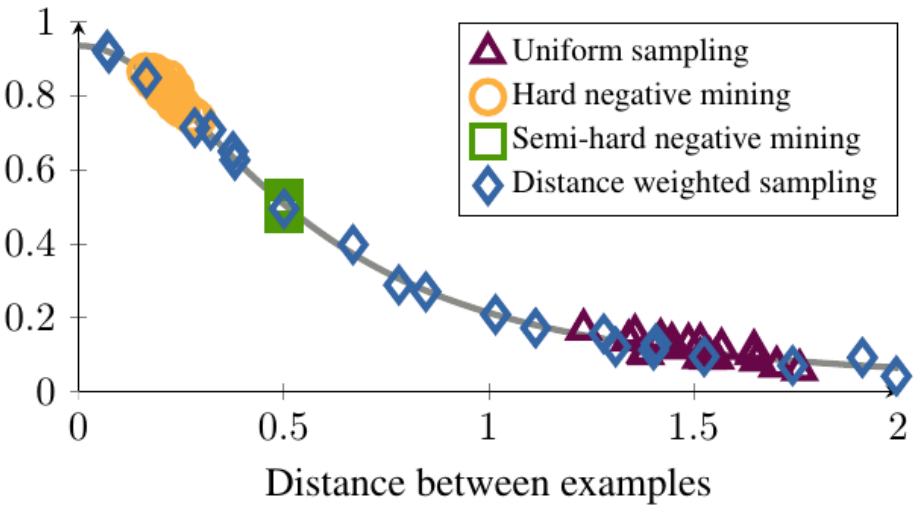
按照距离作为参数进行权重计算 $q(d)^{-1}$, 进行加权:

$$\Pr\left(n^{\star}=n|a\right) \propto \min\left(\lambda, q^{-1}\left(D_{an}\right)\right).$$

2) Comparision to other sampling methods

2.1) 比较的标准

下面是按照不同的采样策略采样出来的examples. 以及在此之上的梯度variance计算.



(b) Sample distribution for different strategies.

2.2) 对这个结果的分析

- Uniform sampling(随机采样) :
  - Property:

    由于上面分析过得, 随机采样的距离的正态分布性(如下图), 大部分距离都集中在1.2-1.7之间

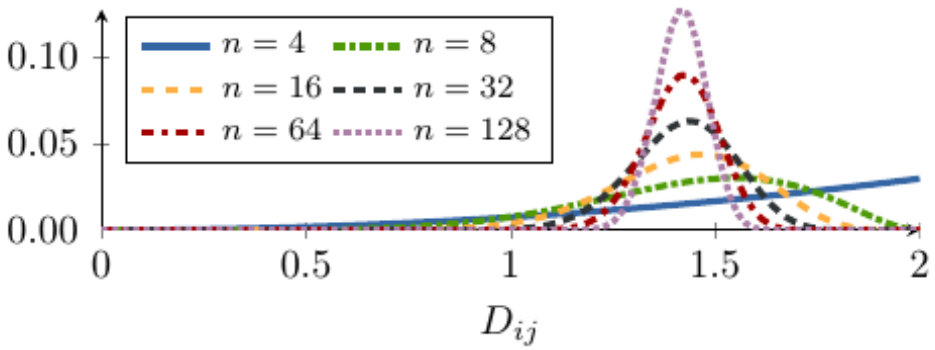


Figure 2: Density of datapoints on the D-dimensional unit sphere. Note the concentration of measure as the dimensionality increases — most points are almost equidistant.

  - Result:

    随机采样的结果会产生较低的梯度, 使得下降过小, 学习速率慢.

- Hard negative mining:
  - Property

    一直使用的是false positive.

  - Result:
    - This leads to **noisy gradients** that **cannot effectively push two examples** apart. 因此使得采取的样本距离集中在方差较大的区域. 使得学习不准确.
    - Lead to **a collapsed model**

- Semi-hard negative mining :
  - Property

    会选择一个mini-batch中最小的大于positive example距离的例子作为negative example : 因为是在一个拥有一般性的batch中选最值, 使得anchor和negative example之间的距离总是非常相近的.
  - Result:

    也许会在一开始收敛速度比较快, 但是在达到某个点后便会停止收敛.
- Distance weigthed sampling(提出的新策略)
  - Property:

    既不会有太大的方差也不会有太小的方差, 所以这个平衡的方案是最棒的.
  - Result:

    可以通过控制方差, 来稳定的提供有信息量的example

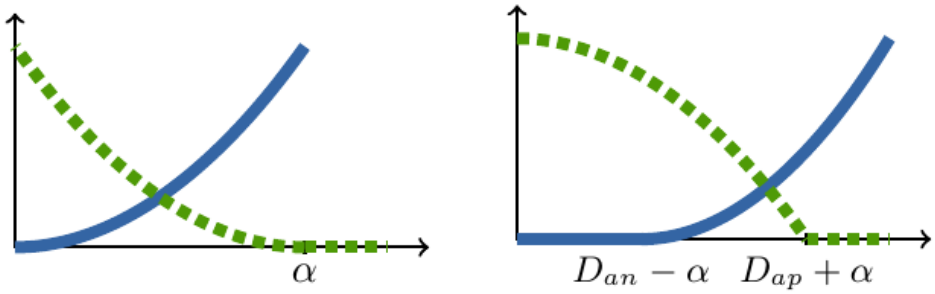
4.4 Margin based loss

1) New loss function

$$\ell^{\text{margin}}(i,j) := (\alpha + y_{ij}(D_{ij} - \beta))_+ .$$

2) Comparision to other loss functions

2.1) 为什么 triplet loss 要比 constrastive loss 好?



(a) Contrastive loss [11] (b) Triplet loss ℓ_2^2 [25]

> The solid blue : loss value for positive pairs
>
> the dotted green : loss value for negative pairs.

- **contrastive loss** 没有假定一个阈值, 也就是这个图像中的与x轴重合的部分. 这样的话, 无法区分similar和dissimilar的图像.

2.2) 为什么 hard negative mining 不适合 triplet loss?

- negative loss 中的凸形曲线

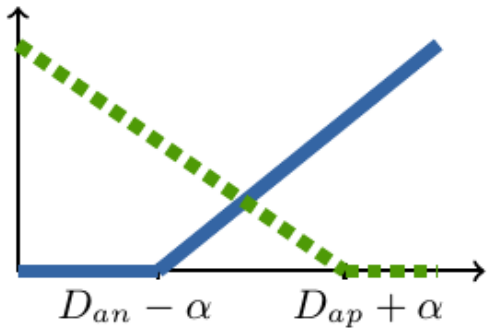
由于 **hard negative mining** 总是对于 negative samples 有较低的loss(这个在上面解释了). 那么, 依照这个曲线来看, 梯度总是很小的, 因此, 无法有效的学习.

2.3) triplet loss 的另外形式

A improvement of triplet loss:

$$\ell^{\text{triplet},\ell_2} := (D_{ap} - D_{an} + \alpha)_+ .$$

由于这里把凸形曲线换成了直线, 因此可以取得更好的效果. 其实就是将二次曲线换成了直线.

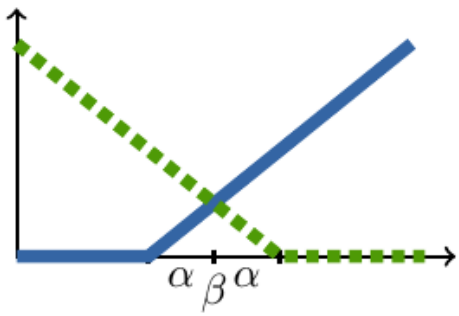


(c) Triplet loss ℓ_2

2.4) Margin based loss

- **Advantage**
  - 相比于 **contrastive loss** 而言 : 有 triplet loss 的灵活性

    > 这里的灵活性是指阈值 α 的设定, **contrastive loss** 中的 α 是一个绝对的值, 不是一个相对的值, 因此影响很大.



(d) Margin based loss

  - 相比于 **triplet loss** : 有计算上的简便性, 因为将 $O(n^3)$ 变成了 $O(n^2)$
- 如何决定 β 的值?

  这里的参数的设定虽然是两个也是相对的, 但是 β 并不像 **triplet loss** 中的一样, 有自然的绝对参照物 D_{ap}, D_{an}, 这里的 β 是需要设定的, 因此为了享有那种灵活性, 必须要将 β 自动确定. 具体如下:

$$\beta(i) := \beta^{(0)} + \beta^{(\text{class})}_{c(i)} + \beta^{(\text{img})}_{i}$$

具体的步骤看论文好了, 不想看这一段.