

极客大学机器学习训练营 神经网络训练方法

王然

众微科技 AI Lab 负责人

二〇二一年三月三十一日

1 基本训练方法

2 预训练方法

3 半监督学习

4 参考文献

1 基本训练方法

2 预训练方法

3 半监督学习

4 参考文献

- ▶ 在本讲中我们只关心有监督学习和相关内容；我们不讨论生成模型和强化学习模型的训练；
- ▶ 即使如此，深度学习的训练也比传统模型的训练要困难；
- ▶ 原因：
 - ▶ 训练时间长；
 - ▶ 训练稳定性差；
 - ▶ 训练方法繁多；
 - ▶ 训练透明性差；
 - ▶ ...

- ▶ 基本训练 trick;
- ▶ 预训练相关主题;
- ▶ 半监督学习相关主题。

- ▶ 传统优化器：SGD (+Momentum) 或 Adam；
- ▶ 其他优化器：AdamW/RAdam/Lookahead/Lars/Lamb；

- ▶ 两种情况：有权重、无权重；
- ▶ 有权重：学习率多半需要很低；增加复杂上层网络时候要注意对于已经训练好的权重和随机初始化的权重采用不同的学习率；后者实现注意 train 和 eval 的正确选择；
- ▶ 无权重：Warmup；Reduce on Plateau；Cosine Annealing 和 Snapshot Ensemble；

- ▶ 一般来说，当模型快进入到收敛的阶段时候，需要减少模型的变化程度（进行微调）；其他目的，减少算力消耗；
- ▶ 加大 Batch Size；
- ▶ 选用更慢的优化器（SGD+Momentum/Lookahead+SGD+Momentum）；

- ▶ 在一些情况下，自然的损失函数未必是最有效的；
- ▶ 其他损失函数可以在不同阶段进行引入，在一些时候，过早的引入损失函数只会使得训练更加困难；
- ▶ 一些有用的损失函数：Focal Loss；Label Smoothing；Temperature in Softmax
- ▶ 其他的函数我们会在后面提到。

- ▶ 不同的初始化 Ensemble 效果更好；
- ▶ 一些文章表明增加惩罚使得不同网络的权重不同更好。

- ▶ 一般应用于 CV 和 NLP 数据当中；
- ▶ 是比赛中最重要的手段；
- ▶ 两种数据扩充方式：增加合理的样本；增加噪声；
- ▶ 不宜过多；过多会实际导致模型记住观测；
- ▶ Prediction Time Augmentation；
- ▶ 不同数据扩充方法可以进行 ensemble。

- ▶ 一般应用与 CV 和 NLP 数据当中；
- ▶ 将不同 encoder 进行（两两）拼接并训练；
- ▶ 增加 ensemble 多样性；
- ▶ 一般来说 encoder 架构/训练方式越不同，带来的增益越大。

1 基本训练方法

2 预训练方法

3 半监督学习

4 参考文献

- ▶ 一些任务具备基本的通用性质；例如图像识别可以认为是（其他）CV 任务的基础，而语言模型的训练可以认为是 NLP 各项任务的基础；
- ▶ 我们关注的任务可能数据集并不够大；
- ▶ 使用在其他任务上训练好的模型，针对当前模型进行 finetune。

- ▶ 预训练；
- ▶ Encoding Refinement。

- ▶ 选择一个非常典型且观测量巨大的数据集进行训练；
- ▶ 典型例子：ImageNet；
- ▶ 对于 CV 来说，不同的网络架构对于不同程度的细节捕捉不一样，所以需要尝试不同的组合；
- ▶ 问题：难以找到合适的数据集；任务不具有代表性；

预训练方式 2：多任务学习（思考题）

如果单任务学习的任务不够代表性，是否可以采取多个任务呢？

- ▶ 多任务学习经常导致模型难以应付不同任务的需求 (Negative Transfer);
- ▶ 解决 Negative Transfer 的方法：
 - ▶ Soft Sharing;
 - ▶ Mixture Model (Fedus, Zoph, and Shazeer 2021)

- ▶ 采用无监督的方式进行训练;
- ▶ 思想: 使用网络将输入进行“还原”;

- ▶ 沙漏型的网络；
- ▶ 对输入增加随机噪声，要求网络恢复原始的输入；
- ▶ 应用时注意将中间的多层提出。

见Doersch (2016)。

- ▶ 主要应用于 CV 的生成任务；
- ▶ 虽然数学形式很优美，但是由于其比较复杂，实际效果不一定好；
- ▶ 重要思路极其衍生：
 - ▶ ELBO，例如 (Hafner et al. 2019)；
 - ▶ 随机输入的处理，例如 Jaegle et al. (2021)；

- ▶ 目前最火的预训练框架之一；
- ▶ 主要来源于预训练语言模型；
- ▶ 在多模态和 CV 当中也开始有应用。

- ▶ 核心思想：如果两个字/词周围的语境类似，则该字/词表示语义也接近；
- ▶ 数学上来讲 $P(o|c) = \frac{\exp(v_o^t v_c)}{\sum_{w \in V} \exp(v_w^t v_c)}$ ；
- ▶ Negative Sampling：由于可选文字很多，所以只抽取部分不属于上下文的字/词；数学形式 $\log \sigma(v_o^t v_c) - \sum_i \log \sigma(-v_i^t v_c)$ ，其中求和针对的是 Negative Sample；
- ▶ 对于结构化数据来说，在有时序的情况下可以考虑使用这种方法；

- ▶ 使用 Masked Language Model 作为目标，使用 Transformer 作为网络结构；
- ▶ 加入 [CLS] token；
- ▶ 很多其他的提升，但是核心思路类似。

多任务学习的引入

1 基本训练方法

2 预训练方法

3 半监督学习





4 参考文献

1 基本训练方法

2 预训练方法

3 半监督学习

4 参考文献

-  Doersch, Carl (2016). "Tutorial on variational autoencoders". In: *arXiv preprint arXiv:1606.05908*.
-  Fedus, William, Barret Zoph, and Noam Shazeer (2021). "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity". In: *arXiv preprint arXiv:2101.03961*.
-  Hafner, Danijar et al. (2019). "Dream to control: Learning behaviors by latent imagination". In: *arXiv preprint arXiv:1912.01603*.
-  Jaegle, Andrew et al. (2021). "Perceiver: General Perception with Iterative Attention". In: *arXiv preprint arXiv:2103.03206*.