THUHCSI
清华大学人机语音交互实验室

NIPS2020
Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect

报告人：张阳

时间：2021年6月17日

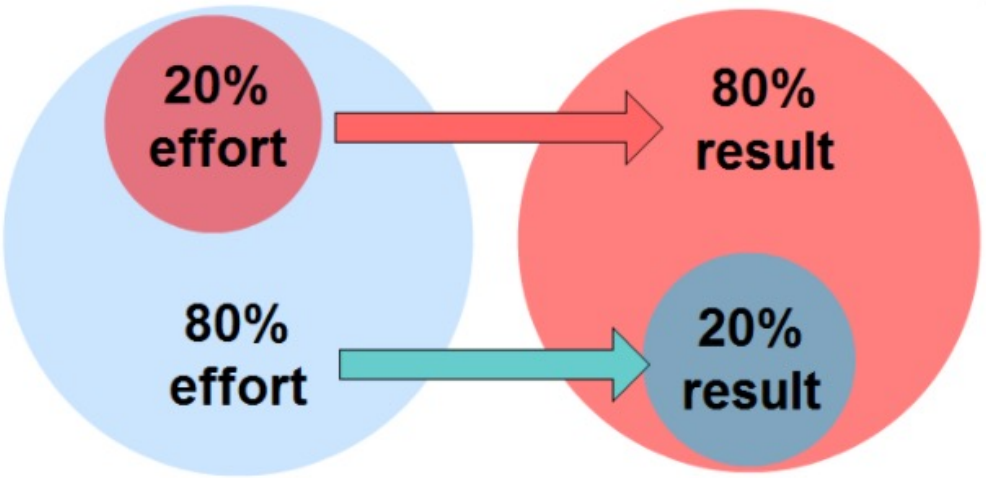
目 录

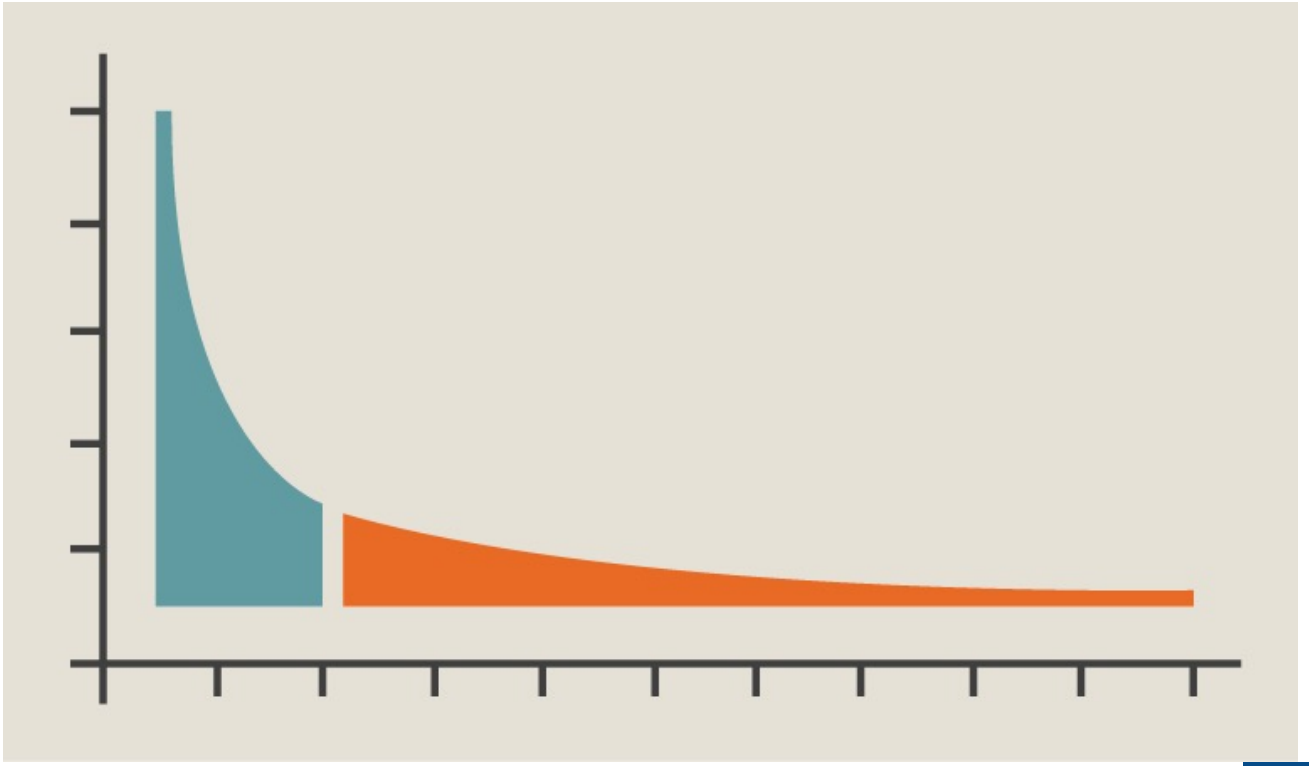
Content

THUHCSI
清华大学人机语音交互实验室

二八定律

- 经济学上，有一个著名的**帕累托法则（Pareto principle）**，又称为二八定律。
- 比如80%的财富集中在20%的人手里，图书馆里20%的书可以满足80%的顾客。于是大家往往只关注在PDF图中最左面的20%的顾客，以期满足80%





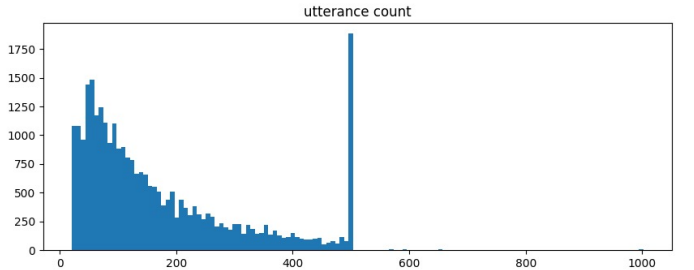
Heavy-tailed distribution

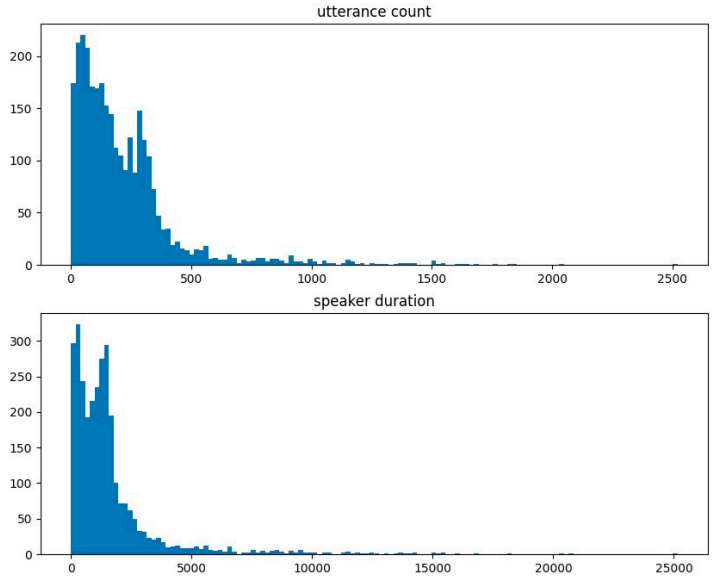
THUHCSI
清华大学人机语音交互实验室

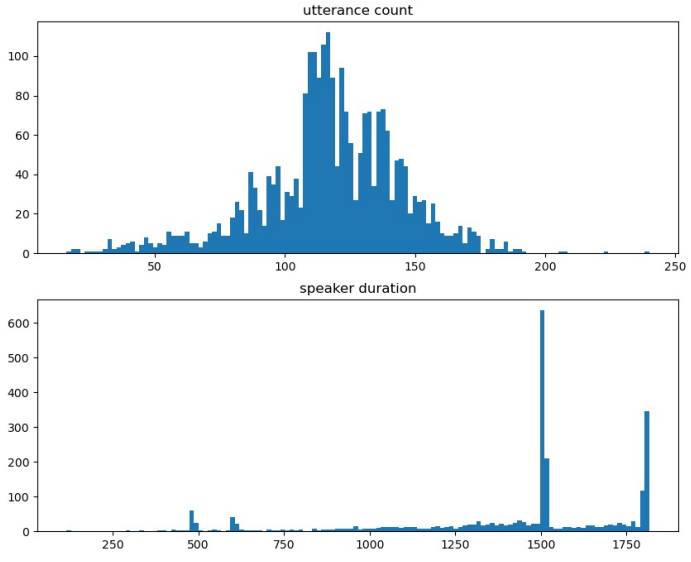
- **重尾分布（Heavy-tailed）**可以分为三个子类型【参考维基百科】：
  - ✓ 分别是长尾分布（long-tailed distributions）
  - ✓ 次指数分布（subexponential distributions）
  - ✓ 肥尾分布 (Fat-tailed distribution)

- **目前我翻到的paper都是研究long-tailed distributions下的**

Voxceleb & CN-Celeb & LibriSpeech data distribution





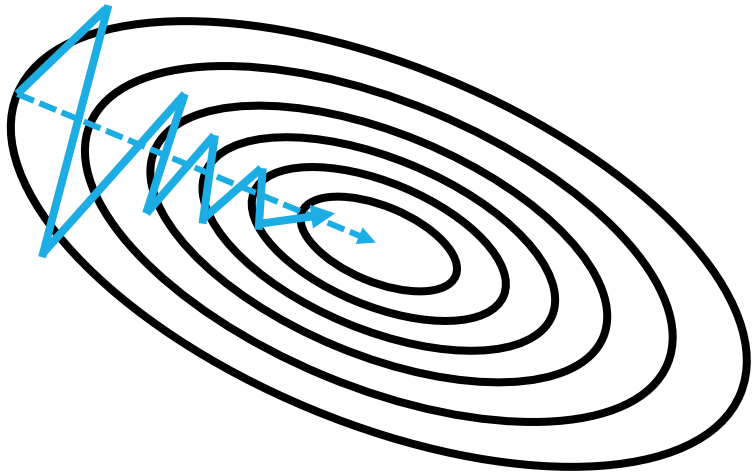


- **voxceleb1&2 dataset**
- number of speaker: 7245
- number of utterance: 1245525
- source: youtube

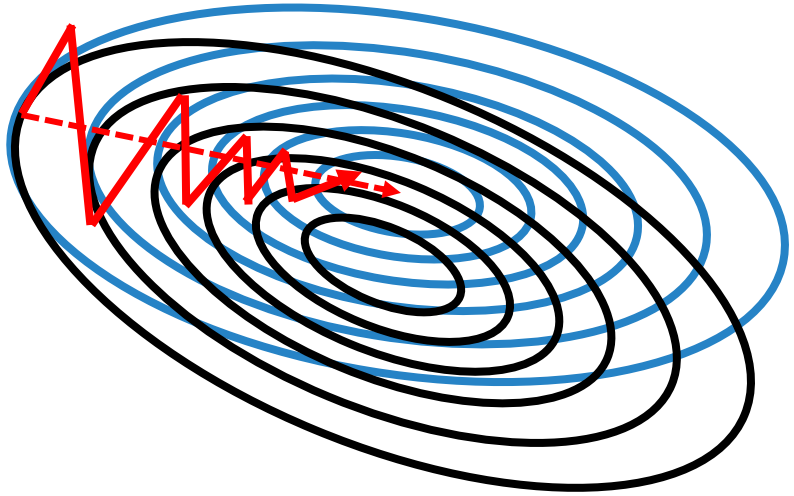
- **cnceleb1&2 dataset**
- number of speaker: 3000
- number of utterance: 659593
- source: bilibili

- **librispeech**
- number of speaker: 2484
- number of utterance: 292367
- source: recording

Accumulative Momentum Effect



SGD Momentum in
Balanced Dataset

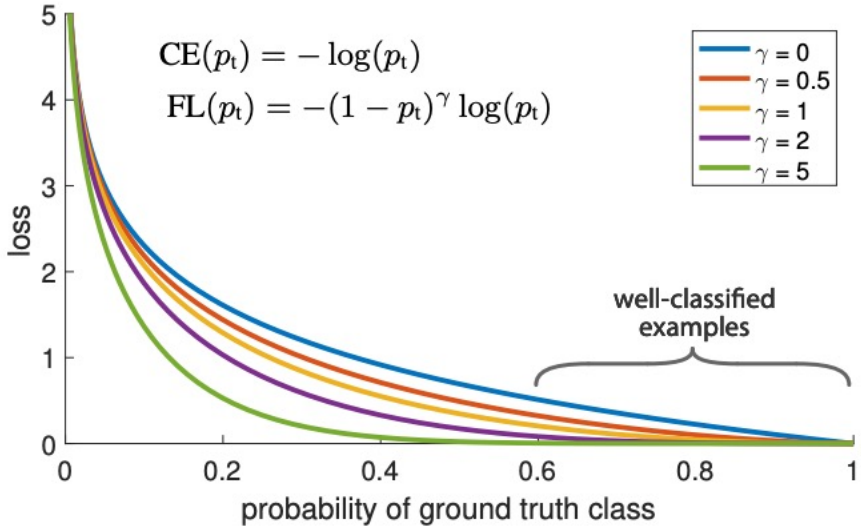


SGD Momentum in
Long-Tailed Dataset

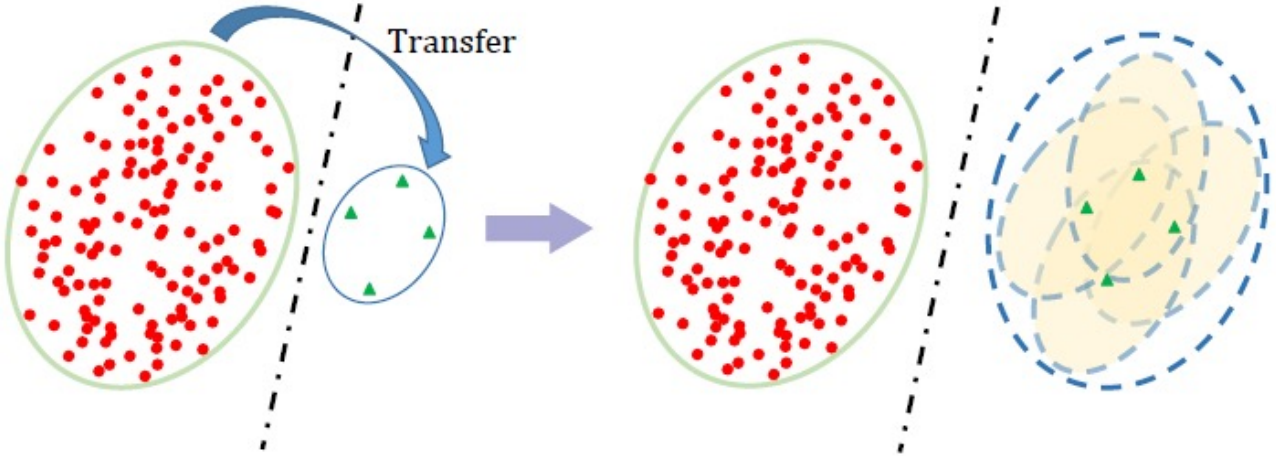
- Global Optima for All Categories
- Local Optima for Head Categories
- Momentum Direction in Balanced Data
- Momentum Direction in Long-Tailed Data

related work

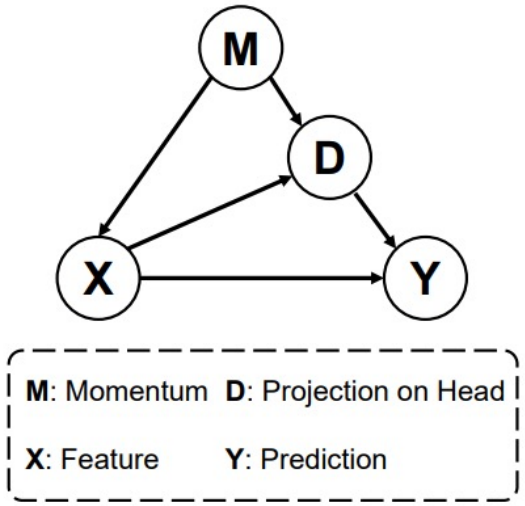
- **the most common solutions:**
  - ✓re-sampling
  - ✓re-weighting



- **other solutions:**
  - ✓focal loss[2]
  - ✓decoupling[3]
  - ✓transfer learning[4]



Causal Graph



(a) The Proposed Causal Graph

- ✓ M 是优化器的动量
- ✓ X 是backbone提取的特征
- ✓ Y 是预测
- ✓ D 是特征对头部大类的偏移量

☐ decoupling 是 two-stage的方法

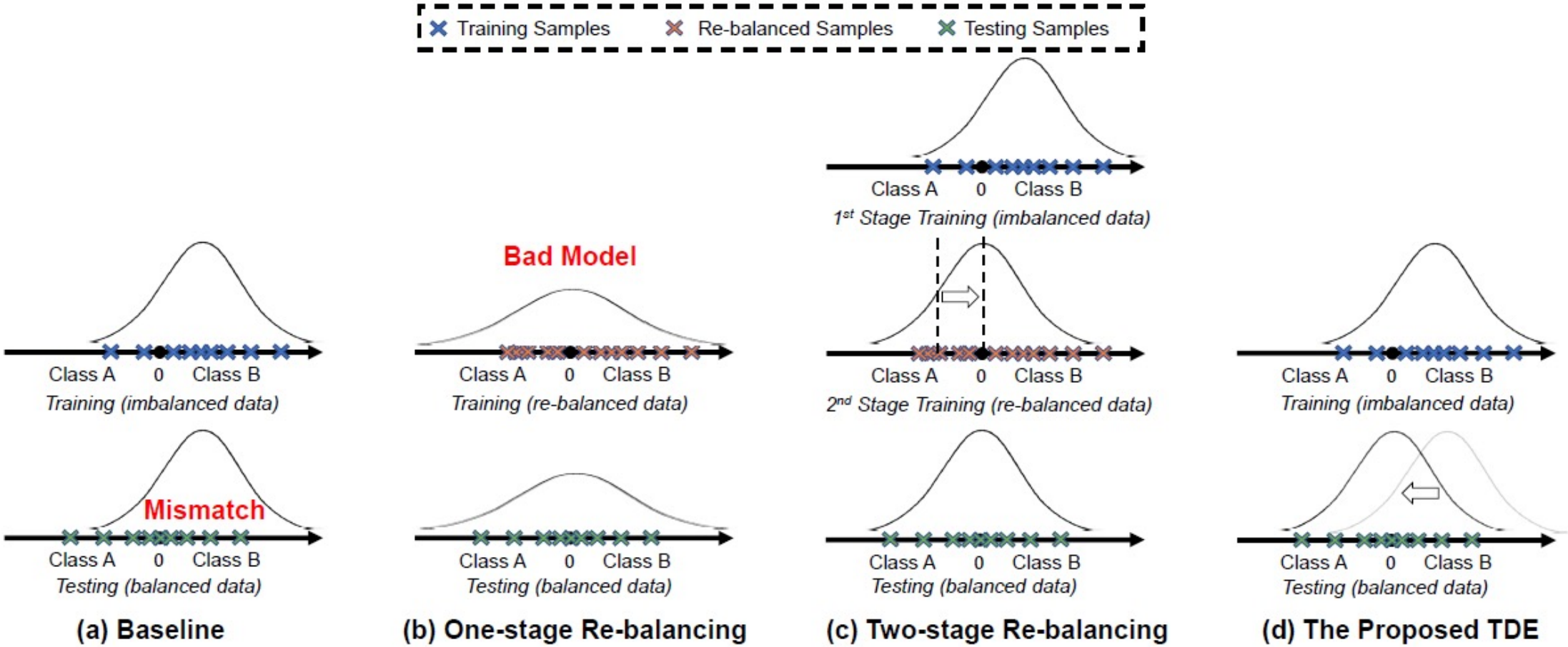
- ✓ **第一步**：不作任何再均衡直接学习
- ✓ **第二步**：将第一步学习的模型中的特征提取backbone的参数固定（不再学习），然后单独接上一个分类器（可以是不同于第一步的分类器），对分类器进行class-balanced sampling学习

☐ 作者实现one-stage的方法

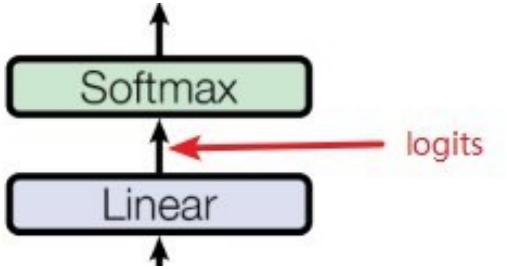
- ✓ De-confound Training
- ✓ counterfactual TDE inference

a simple example

- 0 point is the classifier's decision boundary



De-confound Training



- **linear** classifier logits

$$Y_i = (w_i)^T x + b_i$$

- **cosine** classifier logits

$$Y_i = \frac{(w_i)^T x}{||w_i||||x_i||}$$

- **Causal Norm** classifier logits

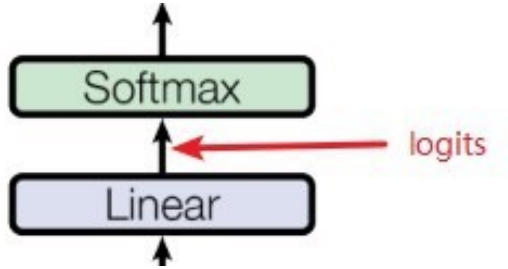
$$Y_i = \frac{(w_i)^T x}{(||w_i|| + \gamma)||x_i||}$$

- **Multi-head CausalNorm Classifier**

$$Y_i = \frac{\tau}{K} \sum_{k=1}^{K} \frac{(w_i^k)^T x^k}{(||w_i^k|| + \gamma)||x^k||}$$

- ✓ K: number of multi-head
- ✓ 用因果干预进行de-confound的训练
- ✓ 通过multi-head多重采样来近似
- ✓ 训练时统计一个移动平均特征 $\tilde{x}$

counterfactual TDE inference



- **linear** classifier logits

$$Y_i = (w_i)^T x + b_i$$

- **cosine** classifier logits

$$\mathrm{Y}_i = \frac{(w_i)^T x}{||w_i||||x_i||}$$

- **Causal Norm** classifier logits

$$\mathrm{Y}_i = \frac{(w_i)^T x}{(||w_i|| + \gamma)||x_i||}$$

- **Counterfactual TDE inference**
  - ✓ TDE方法则直接 矫正特征本身的分布
  - ✓ $\tilde{\mathrm{x}}$ 平均特征， 他的单位方向看作是特征对头部类的倾向方向

$$\hat{d} = \bar{x}/||\bar{x}||$$

  - ✓ 从logits中剔除对头部类过度倾向的部分

$$TDE(Y_i) = \frac{\tau}{K}\sum_{k=1}^{K}\left(\frac{(w_i^k)^T x^k}{(||w_i^k|| + \gamma)||x^k||} - \alpha \cdot \frac{cos(x^k, \hat{d}^k) \cdot (w_i^k)^T \hat{d}^k}{||w_i^k|| + \gamma}\right)$$

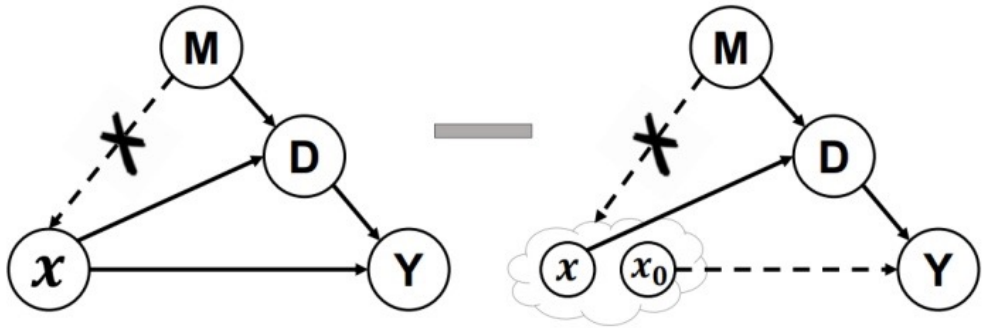


Image Classification: ImageNet-LT

Methods	Many-shot	Medium-shot	Few-shot	Overall
Focal Loss† [24]	64.3	37.1	8.2	43.7
OLTR† [8]	51.0	40.8	20.8	41.9
Decouple-OLTR† [8, 10]	59.9	45.8	27.6	48.7
Decouple-Joint [10]	65.9	37.5	7.7	44.4
Decouple-NCM [10]	56.6	45.3	28.1	47.3
Decouple-cRT [10]	61.8	46.2	27.4	49.6
Decouple-τ-norm [10]	59.1	46.9	30.7	49.4
Decouple-LWS [10]	60.2	47.2	30.3	49.9
Baseline	66.1	38.4	8.9	45.0
Cosine† [38, 39]	67.3	41.3	14.0	47.6
Capsule† [8, 42]	67.1	40.0	11.2	46.5
(Ours) De-confound	**67.9**	42.7	14.7	48.6
(Ours) Cosine-TDE	61.8	47.1	30.4	50.5
(Ours) Capsule-TDE	62.3	46.9	30.6	50.6
(Ours) De-confound-TDE	62.7	**48.8**	**31.6**	**51.8**

- Many-shot
  - ✓ > 100
- Medium-shot
  - ✓ >= 20 & < 100
- Few-shot
  - ✓ < 20

Image Classification: ImageNet-LT

- the improvement come from **multi-head trick**

Methods	#heads K	Many-shot	Medium-shot	Few-shot	Overall
Cosine† [5, 6]	1	67.3	41.3	14.0	47.6
Cosine† [5, 6]	2	67.5	42.1	14.1	48.1
Capsule† [8, 10]	1	67.1	40.0	11.2	46.5
Capsule† [8, 10]	2	67.7	41.3	12.6	47.6
(Ours) De-confound	1	67.3	41.8	15.0	47.9
(Ours) De-confound	2	**67.9**	42.7	14.7	48.6
(Ours) Cosine-TDE	1	61.8	47.1	30.4	50.5
(Ours) Cosine-TDE	2	63.0	47.3	31.0	51.1
(Ours) Capsule-TDE	1	62.3	46.9	30.6	50.6
(Ours) Capsule-TDE	2	62.4	47.9	31.5	51.2
(Ours) De-confound-TDE	1	62.5	47.8	**32.8**	51.4
(Ours) De-confound-TDE	2	62.7	**48.8**	31.6	**51.8**

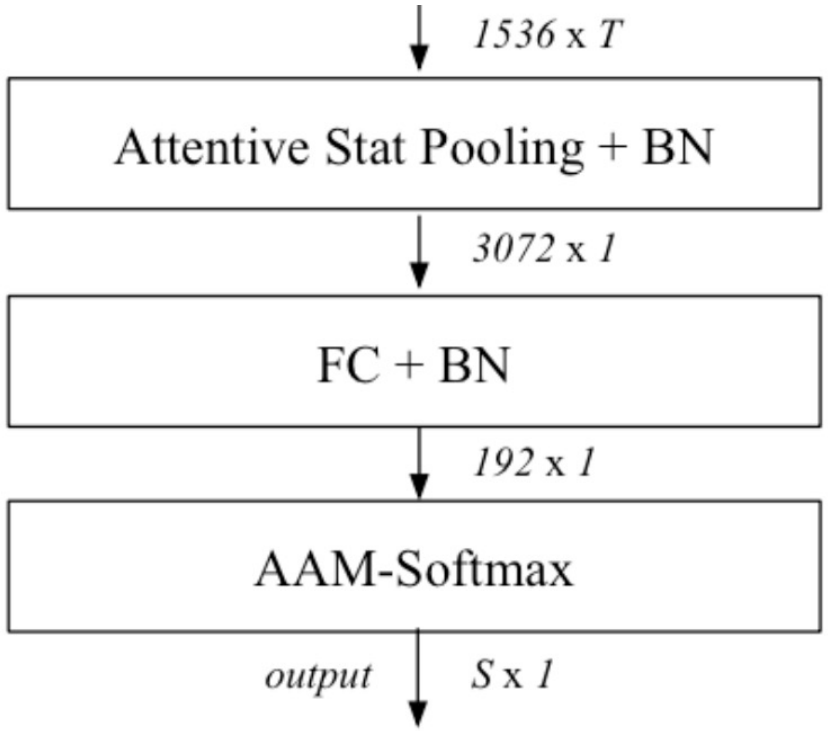
Image Classification: ImageNet-LT

- improvement can be consistent across different backbones

Methods	Backbone	Many-shot	Medium-shot	Few-shot	Overall
Baseline	ResNeXt-50	66.1	38.4	8.9	45.0
De-confound	ResNeXt-50	67.9	42.7	14.7	48.6
De-confound-TDE	ResNeXt-50	62.7	48.8	31.6	51.8
Baseline	ResNeXt-101	68.7	42.5	11.8	48.4
De-confound	ResNeXt-101	**68.9**	44.3	16.5	50.0
De-confound-TDE	ResNeXt-101	64.7	**50.0**	**33.0**	**53.3**

Advantages

- One-stage training.
- Do not need to know the data distribution during training.
- Great performance with fewer compute resource.



Reference

THUHCSI 清华大学人机语音交互实验室

相关论文

- [1] Tang K, Huang J, Zhang H. Long-tailed classification by keeping the good and removing the bad momentum causal effect[J]. arXiv preprint arXiv:2009.12991, 2020.
- [2] Lin T Y , Goyal P , Girshick R , et al. Focal Loss for Dense Object Detection[C]// IEEE Transactions on Pattern Analysis & Machine Intelligence. IEEE, 2017:2999-3007.
- [3] Kang B , Xie S , Rohrbach M , et al. Decoupling Representation and Classifier for Long-Tailed Recognition[J]. 2019.
- [4] Liu J , Sun Y , Han C , et al. Deep Representation Learning on Long-tailed Data: A Learnable Embedding Augmentation Perspective[J]. 2020.

参考资料

- https://github.com/KaihuaTang/Long-Tailed-Recognition.pytorch
- https://github.com/GZWQ/Awesome-Long-Tailed
- https://zhuanlan.zhihu.com/p/259569655

THUHCSI
清华大学人机语音交互实验室

谢谢！Q&A