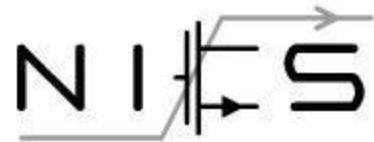




# Hardware-Friendly Optimization For Semi-Supervised Neural Network Training

Tianchen Zhao





# N Questions

- 1. What Is It? 想做个啥
- 2. Why That? 这玩意有啥用?
- 3. How To Implement? 咋个做法?
- 4. Where Do I Begin? 我准备做啥?



# Overview

## 1. What Is It? 想做个啥



# Overview

- 利用神经网络在终端的Online Training, 利用Unlabelled的数据, 对其训练, 以提高性能以及一定意义上的迁移学习
- FAIR的一篇新文章《Billion Scale Semi-Supervised Learning》提供了一种Workflow感觉对硬件终端部署有一定应用价值
  - 准备对其从硬件角度看看算法上的优化空间
  - 硬件架构实现



# Overview

## 2. Why That? 有什么用



# Background

- **Why Training**

- 接触到了实际场景的数据，目的使模型更加贴合真实应用场景
- 以提高Task的准确度 (Boost Performance)
  - Task-Specific的语义信息能够更好的协助训练
- 以更适合真实数据 (Domain Transfer)
- ~~去实际场景学习，让模型具有累积，Incremental Learning~~

- **Why Training On Device (Not On Cloud)**

- 通信受限场景
- 通信对设备功耗 (?)
- 数据隐私性，安全性



# Background

- Online Training落地的几个困难问题

- 计算量与存储其实占用很大（对比推断），与预先训练的优势？

- 对比Offline Training（提前拿GPU来训练），需要有一定的价值和意义

- 数据的Label从哪里来？

- 一个好的Example是商汤的安检闸机,以用户识别失败之后的刷卡为Label，但是这样的Label还是少
- 但是大多数时候，大多数时候没有Label

- 训练会不会把模型训练崩溃

- 训练不好控制，可能会崩溃
- 如何规划接受到的新数据
- 如果全部接收会存在，Class Imbalance的(只有几类数据经常出现)，且会有强Temporal关系(一段时间内好多个一样的类别)



# Background

- Online Training落地的几个困难问题 - Solution

- 计算量与存储其实占用很大（对比推断），与预先训练的优势？

- Exposed To 实际场景的数据，更加合理。而且Task-Specific的背景语义信息可以提点
- 证明了又一定的迁移学习能力，可以迁移到实际应用场景
- 对于本任务也可以有显著性能提升（在ImageNet上面比SOTA能提1~2个点）

- 数据的Label从哪里来？

- 通过一个Pretrain的Teacher Network为数据打标签

- 训练会不会把模型训练崩溃

- 有pretrain好的Teacher Network（参数不会被更新），有一定的保障
- 提出了一套Easy But Effective(作者原文)的训练集建立方式





# Background

- 如果实际在硬件终端上部署，该方法的**边际成本近乎为0**
  - 在现实场景中，缺的是Label而不是Data，而获取Data可以认为是几乎0成本。
  - 在线训练对于长周期的部署来说，其可在空闲时间内完成，利用了多余的计算资源。
  - 实际增多的，就是存储数据量的资源，而由于目前大部分的FPGA的片上RAM都不足以存下整个网络的参数，需要额外的存储器件，因而额外的存储也不会带来很大的开销。



### 3. How To Implement? 如何实现



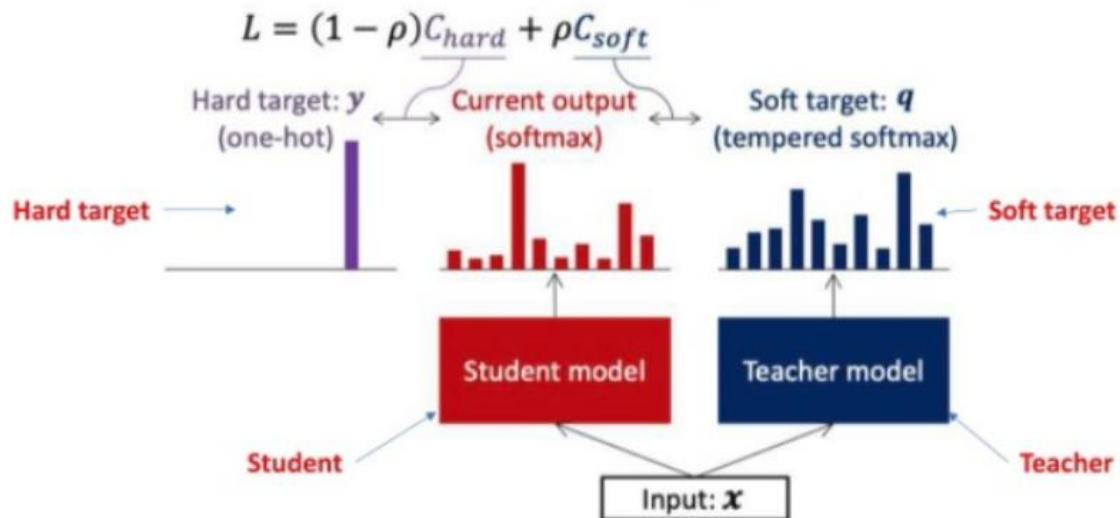
# Related Work

## • Knowledge Distillation

- Hinton在2014年提出的，实际是一个类似Teacher-Student的结构，去利用大网络对样本输出的Softmax的结果作为Soft Label(与原本One-Hot的Hard Label相对应)，去训练一个小的网络(将大网络的知识“蒸馏”提炼到网络中来进行网络)，可以认为小网络被指引去模仿大网络。
- 可以认为保留了“什么类别更容易被混淆”这一信息，因而能更好地训练

### Loss function

Transfer set = unlabeled data + original training set





# Related Work

- Knowledge Distillation — New Loss Function

- 提出了一个新的Loss Function，叫做Softmax With Temperature，本质上就是让Softmax本来差距很大的数值变得更接近
- 由于网络的输出差异过大 [0.96,0.01,0.01,0.01,0.01...]

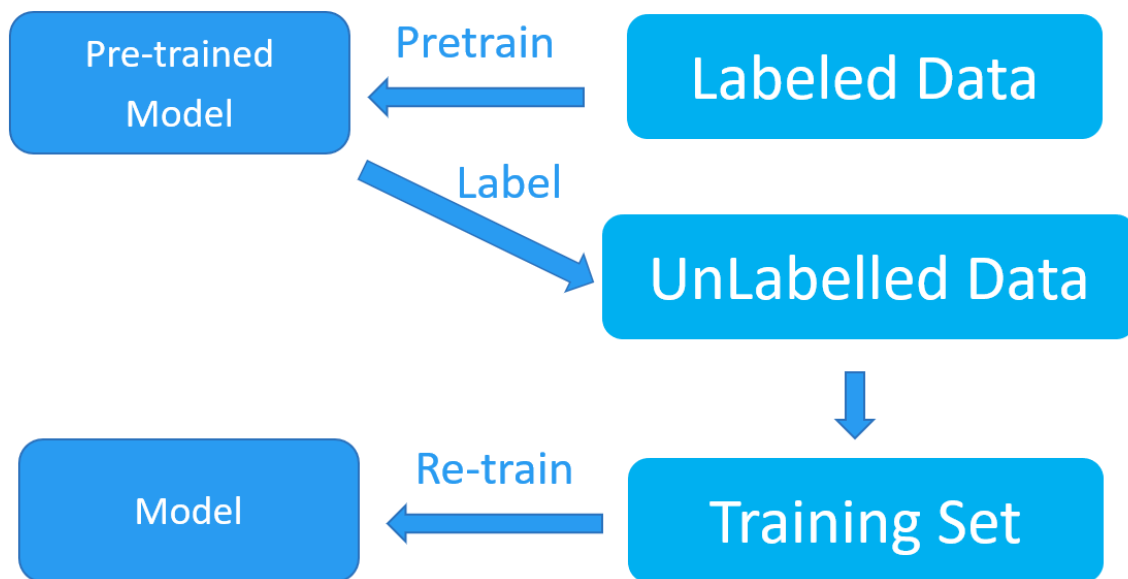
$$q_i = \frac{e^{\frac{z_i}{T}}}{\sum_j e^{\frac{z_j}{T}}}$$



# Related Work

- Semi-Supervised Learning 半监督学习

- 传统的Training Scheme属于全监督学习，而半监督学习的特征在于训练数据只有少部分具有标签，而需要从大量没有标签的数据中进行学习





# Related Work

- Semi-Supervised Learning 半监督学习

- 传统的Training Scheme属于全监督学习，而半监督学习的特征在于训练数据只有少部分具有标签，而需要从大量没有标签的数据中进行学习
- 与其他领域的联系
  - Distillation的过程可以被认为是一种半监督学习 (但是本文没有采用)
  - 半监督学习可以被认为是一种特殊的Data Augmentation、
  - 由于两者的训练数据有差异，可以被认为是弱的Transfer Learning
  - 当semi-supervised可以学到新的类别时可以认为是一种Incremental Learning
  - 由于训练数据只有少量标签，经常和Few Shot Learning相联系

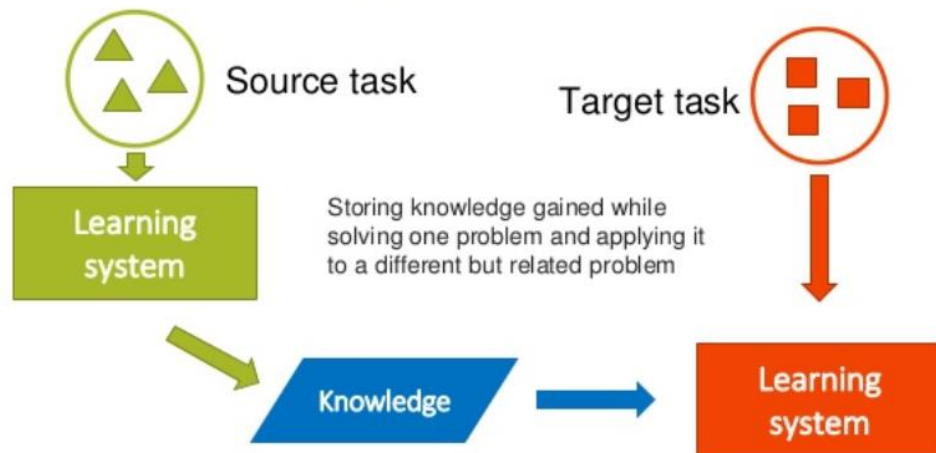


# Related Work

## • Transfer Learning 迁移学习

- 分为多个层次，本质上的要求是数据域(Domain)的转移，当实际场景的数据Unavailable时
- 我们的任务涉及的是比较弱的（这样比较稳），实际数据和训练数据有一定的差异性（但是物理上都是光学图像），存在一个数据域迁移(Domain Adaptation)的问题)

### Transfer learning





# 《Billion-scale semi-supervised learning for image classification》

- FAIR的比较新的工作
- <https://arxiv.org/abs/1905.00546> (未发表)
- 利用Teacher-Student架构为Unlabel的数据标注来训练Student Network, 以提高点数

## Billion-scale semi-supervised learning for image classification

I. Zeki Yalniz

Hervé Jégou

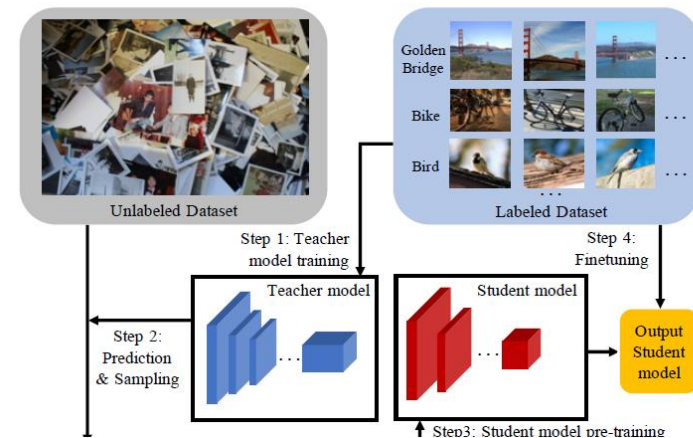
Kan Chen  
Facebook AI

Manohar Paluri

Dhruv Mahajan

### Abstract

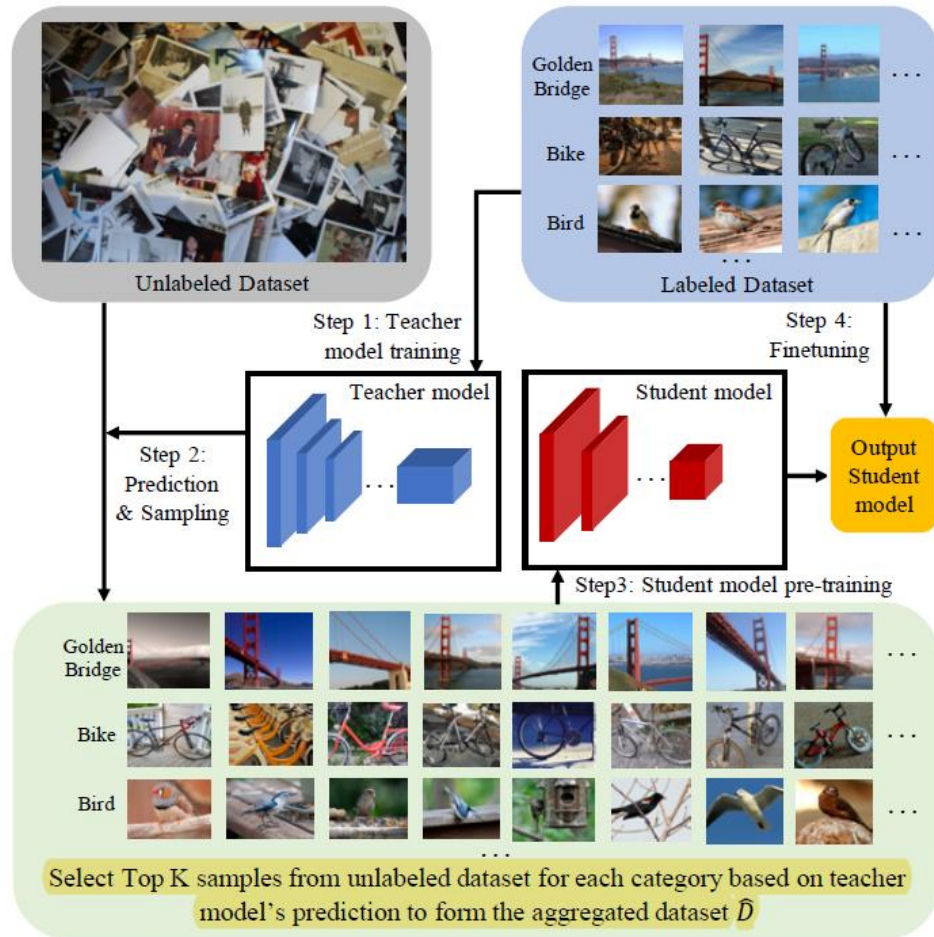
*This paper presents a study of semi-supervised learning with large convolutional networks. We propose a pipeline, based on a teacher/student paradigm, that leverages a large collection of unlabelled images (up to 1 billion). Our main goal is to improve the performance for a given target architecture, like ResNet-50 or ResNext. We provide an extensive analysis of the success factors of our approach, which leads us to formulate some recommendations to produce high-accuracy models for image classification with semi-supervised learning. As a result, our approach brings im-*







# 《Billion-scale semi-supervised learning for image classification》



1. Train Teacher On ImageNet

2. Use Pre-trained Teacher Model To Eval The YFCC(Unlabelled Dataset)

3. Using The Label Created To Sample An New Training Set

3. Using The New Training Set To Train Student From Scratch

4. Use The ImageNet(Clean Label) To Finetune The New Network



- Unlabel的数据集

- YFCC-100M
- 来自雅虎Flickr的Social Media的图片

p. 18



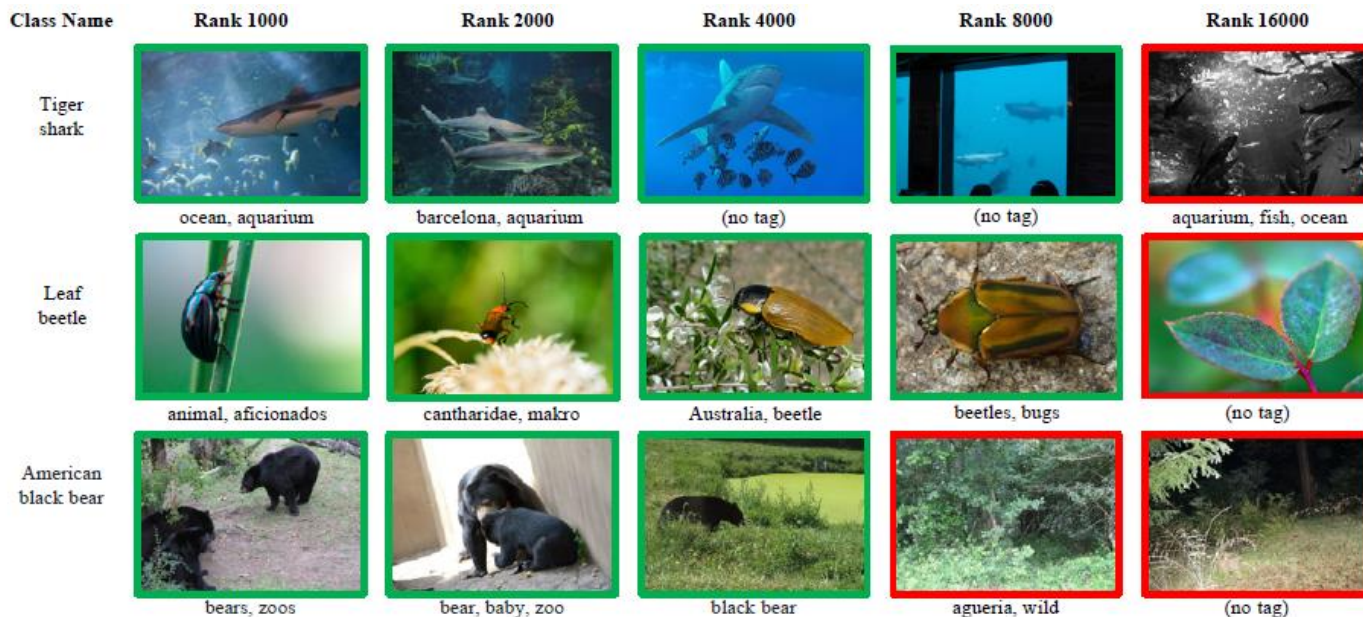
# 《Billion-scale semi-supervised learning for image classification》

- 新的数据集如何构建？

- 首先确定几个参数

- 从Unlabel数据集中采样大小M (100M/50M/25M)
- 每一类图片取前K个标注为样本 (大概在几k的量级)
- 每张图片可以最多可被识别为P类 (最终决定P取10能够获得一个相对Balanced训练集)

- 根据以上几个指标构建出一个新的DataSet





# 《Billion-scale semi-supervised learning for image classification》

- 实验结果
  - 普遍能够比SOTA高1~2个点（相当显著了）

Student Model	Ours: Semi-supervised		Fully Supervised
	Pre-training	Fine-tuned	
ResNet-18	68.7	<b>72.6</b>	70.6
ResNet-50	75.9	<b>79.1</b>	76.4
ResNext-50-32x4	76.7	<b>79.9</b>	77.6
ResNext-101-32x4	77.5	<b>80.8</b>	78.5
ResNext-101-32x8	78.1	<b>81.2</b>	79.1
ResNext-101-32x16	78.5	<b>81.2</b>	79.6

Table 2: ImageNet1k-val top-1 accuracy for students models of varying capacity before and after fine-tuning compared to corresponding fully-supervised baseline models.



# 《Billion-scale semi-supervised learning for image classification》

## • 分析结果

- 对不同的Teacher架构, student的效率有一定不同的提升(图中Student都为Res50)
- 随着Unlabel Dataset的不断增大, 准确度增长几乎线性

Model	Teacher # Params	top-1	Student top-1	Gain (%)
ResNet-18	8.6M	70.6	75.7	-0.7
ResNet-50	25M	76.4	77.6	+1.2
ResNext-50-32x4	25M	77.6	78.2	+1.8
ResNext-101-32x4	43M	78.5	78.7	+2.3
ResNext-101-32x8	88M	79.1	78.7	+2.3
ResNext-101-32x16	193M	79.6	79.1	+2.7
ResNext-101-32x48	829M	79.8	79.1	+2.7

Table 3: Varying the teacher capacity for training a ResNet-50 student model with our approach. The gain is the absolute accuracy improvement over the supervised baseline.

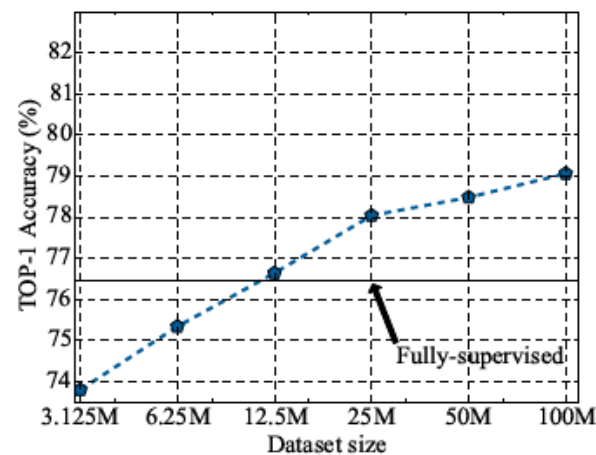


Figure 3: ResNet-50 student model accuracy as a function of the size of the unlabeled dataset  $\mathcal{U}$ .





# 《Billion-scale semi-supervised learning for image classification》

## • 实验结果

- 作者在文中还单独拎出了Self-Training（当student和teacher结构一致的时候，退化为self-training）
- 表示比较Risky(在实际应用场景内不可控制)因为误差可能会不断积累，最后的Finetune起到了很关键的效果

	ResNet-*		ResNeXt- 50-32x4	ResNeXt-101-*		
	18	50		32x4	32x8	32x16
Acc.	70.6	77.6	78.9	80.2	81.1	81.4
Gains	0.0	+1.2	+1.3	+1.7	+2.0	+1.8

Table 4: Self-training: top-1 accuracy of ResNet and ResNeXt models self-trained on the YFCC dataset. Gains refer to improvement over the fully supervised baseline.



# 《Billion-scale semi-supervised learning for image classification》

## • 实验结果

- 作者在文中还单独拎出了Self-Training（当student和teacher结构一致的时候，退化为self-training）
- 表示比较Risky(在实际应用场景内不可控制)因为误差可能会不断积累，最后的Finetune起到了很关键的效果

	ResNet-*		ResNeXt- 50-32x4	ResNeXt-101-*		
	18	50		32x4	32x8	32x16
Acc.	70.6	77.6	78.9	80.2	81.1	81.4
Gains	0.0	+1.2	+1.3	+1.7	+2.0	+1.8

Table 4: Self-training: top-1 accuracy of ResNet and ResNeXt models self-trained on the YFCC dataset. Gains refer to improvement over the fully supervised baseline.



# 《Billion-scale semi-supervised learning for image classification》

## • 分析结果

- Training的Epoch不断增加对Performance的影响
- Student的Performance随着每一类别的图片增加

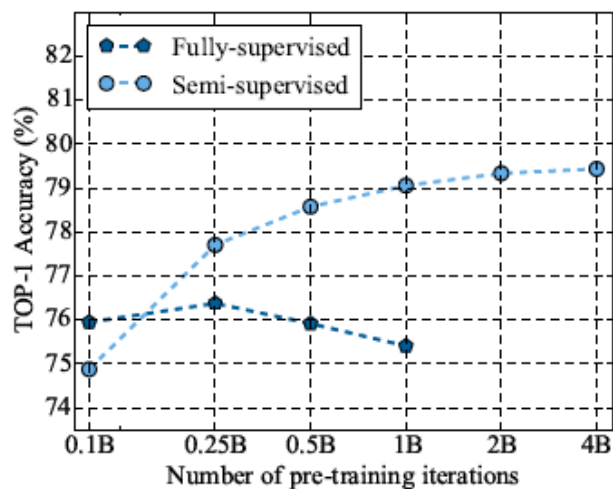


Figure 4: Effect of number of training iterations on the accuracy of fully-supervised and semi-supervised ResNet-50 student models.

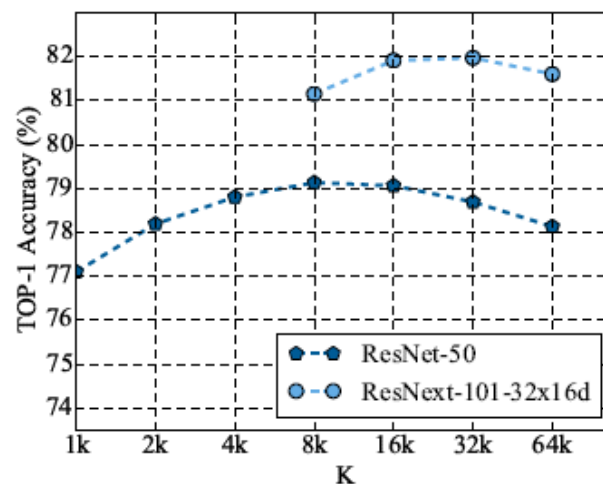


Figure 5: Student model accuracies as a function of the sampling hyper-parameter  $K$ .





## 《Billion-scale semi-supervised learning for image classification》

- 应该怎么构建数据集
  - 作者认为Rank这一步骤起到了精度提升
  - Unbalanced数据集会掉精度

	Balanced	Selection	Accuracy
balanced-ranked	Yes	Ranked	79.2
unbalanced-ranked	No	Ranked	78.4
balanced-with-tags	Yes	Random	76.8
supervised	-	-	76.4

Table 5: Analysis of the selection step on IG-1B-Targeted. All methods selects a subset of 8 million images for training, our balanced-ranked method works the best.



# 《Billion-scale semi-supervised learning for image classification》

- 除了Image Classification, 作者还尝试了多个Task尝试证明这种Workflow的准确性

- Video Classification
- Transfer Learning

Pre-trained Model	ImageNet sup.	weakly sup.	semi sup. (ours)	semi-weakly sup. (ours)
<i>full-ft</i>	82.1	83.2	83.6	<b>84.8</b>
<i>fc-only</i>	73.3	74.0	80.4	<b>80.7</b>

Table 8: CUB2011: Transfer learning accuracy (ResNet50).

Approach	Input	top-1	top-5
<i>R(2+1)D-18, clip length - 8, # params 33M, FLOPS - 21B</i>			
ours (semi-weakly sup.)	RGB	74.2	91.3
fully-supervised	RGB	64.8	84.5
weakly-supervised	RGB	71.5	89.7
<i>R(2+1)D-18, clip length - 32, # params 33M, FLOPS - 83B</i>			
ours (semi-weakly sup.)	RGB	76.7	92.3
fully-supervised	RGB	69.3	87.7
weakly-supervised	RGB	76.0	92.0
<i>R(2+1)D-34, clip length - 8, # params 64M, FLOPS - 38B</i>			
ours (semi-weakly sup.)	RGB	75.9	92.0
fully-supervised	RGB	67.0	85.8
weakly-supervised	RGB	74.8	91.2
NL I3D [42]	RGB	77.7	93.3
3 stream SATT [44]	RGB+flow+audio	77.7	93.2
I3D-Two-Stream [5]	RGB+flow	75.7	92.0



# 《Billion-scale semi-supervised learning for image classification》

- CUB2011 Dataset: 迁移学习数据集
  - Caltech-UCSD-Birds
    - 200 Class, 1,7787 Pics
  - Transfer Learning Setting
    - 把前面的卷积特征提取保留, 最后新训一个新的Softmax(也可以认为是一个Logistics Regression层)
  - 2 Method: Finetune整个网络/只finetune最后的fc


Pre-trained Model	ImageNet sup.	weakly sup.	semi sup. (ours)	semi-weakly sup. (ours)
<i>full-ft</i>	82.1	83.2	83.6	<b>84.8</b>
<i>fc-only</i>	73.3	74.0	 80.4	<b>80.7</b>

Table 8: CUB2011: Transfer learning accuracy (ResNet50).



# 《Billion-scale semi-supervised learning for image classification》

- 迁移学习

- 并没有对迁移学习本身下功夫，只是因为获得了更好的BackBone，所以最后效果舔狗了
- 如果只训练fc的时候，半监督的方法，起到了很大提升


Pre-trained Model	ImageNet sup.	weakly sup.	semi sup. (ours)	semi-weakly sup. (ours)
<i>full-ft</i>	82.1	83.2	83.6	<b>84.8</b>
<i>fc-only</i>	73.3	74.0	 80.4	<b>80.7</b>

Table 8: CUB2011: Transfer learning accuracy (ResNet50).



## 《And the Bit Goes Down: Revisiting the Quantization of Neural Networks》

- <https://arxiv.org/abs/1907.05686> (未发表)
- Code At <https://github.com/facebookresearch/kill-the-bits>
- 对NN的量化提出了一些新的思路，提出了一种针对ResNet的网络压缩方法，获得了很好的效果
- 上面的文章成果被用来服务于这篇文章

we use this particular model and refer to it as semi-supervised ResNet-50. In the low compression regime (block sizes of 4 and 9), with  $k = 256$  centroids (practical for implementation), our compressed semi-supervised ResNet-50 reaches **76.12% top-1 accuracy**. In other words, the compressed model attains the performance of a vanilla, non-compressed ResNet50 while having a size of 5 MB (vs. 97.5MB for the non-compressed ResNet-50).



## 4. Where Do I Begin? 我将做什么



# Ideas

- 2个层面

- 软件层面：针对硬件对算法进行优化

- 新的Training Set如何构建与维护(排序的方式硬件不友好)
    - 与Model Compression方法的联结

- 硬件层面：设计可行的硬件实现架构

- Training设计与外部设计配合？(待讨论...)
    - Teacher和Student可以时分复用
    - 多Procedure形式



# Ideas - Software

- New Training Set如何构建（需要复现与实验）

- 文章中的Training Set是一个很大的数据集(已有)，所以给Unlabel数据标注可以高度并行，而现实中数据需要慢慢收集
- 文章中构建Training Set是依据置信度排序的，针对硬件是否可以修改？
- 文中希望得到一个Balanced的数据集，但是针对实际应用场景是否可以只针对几类进行修正与专门训练？(需要一些实验)

- 该方法是否可以与Compression在某个层面结合

- 《And The Bits Goes Down》一文中只是先用这种方法提点，量化误差抵消掉提的这些点数，有一点“噱头”
- 可以尝试去说明经过压缩的网络也具有这样的学习能力
- 该种训练方式如何与定点与剪枝相结合？
  - How 2 定点稀疏训练的





# Ideas - Hardware

- 硬件的Training方式是否可以与Workflow相配合
  - (...)
- 具体的Training Set的存储与调用形式硬件上也可以有玩法
  - 数据和对应的Label所对应的标签的存储方式 (存储量相对还是太大了)
  - 数据的调用方式
- Teacher/Student结构两个网络在硬件上显得很冗余
  - 由于数据和Weight都是从DDR取的，并且整个semi-supervised learning的Workflow中没有T/S同时计算的，当网络结构相同的时候，可以只占用一份逻辑资源
  - 可认为是计算资源的时分复用



# Thanks for Listening!

欢迎批评指正!



# Ideas





# Reference

- [1] Three scenarios for continual learning(Not Published, Only On Arxiv)
- [2] Goodfellow, Ian J., et al. "An empirical investigation of catastrophic forgetting in gradient-based neural networks." *arXiv preprint arXiv:1312.6211* (2013).
- [3] Li, Zhizhong, and Derek Hoiem. "Learning without forgetting." *IEEE transactions on pattern analysis and machine intelligence* 40.12 (2017): 2935-2947
- [4] Kirkpatrick, James, et al. "Overcoming catastrophic forgetting in neural networks." *Proceedings of the national academy of sciences* 114.13 (2017): 3521-3526.
- [5] Rebuffi, Sylvestre-Alvise, et al. "icarl: Incremental classifier and representation learning." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.



# Reference

- [6] Zenke, Friedemann, Ben Poole, and Surya Ganguli. "Continual learning through synaptic intelligence." *Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org*, 2017.
- [7] Yoon, Jaehong, et al. "Lifelong learning with dynamically expandable networks." *arXiv preprint arXiv:1708.01547* (2017).
- [8] van de Ven, Gido M., and Andreas S. Tolias. "Generative replay with feedback connections as a general strategy for continual learning." *arXiv preprint arXiv:1809.10635* (2018).
- [9] Oswald, J.V., Henning, C., Sacramento, J., & Grewe, B.F. (2019). Continual learning with hypernetworks. ArXiv, abs/1906.00695.



# Reference

- [10] Zhang, Mengmi, et al. "Prototype Reminding for Continual Learning." *arXiv preprint arXiv:1905.09447* (2019).
- [11] Hsu, Yen-Chang et al. "Re-evaluating Continual Learning Scenarios: A Categorization and Case for Strong Baselines." *ArXiv abs/1810.12488* (2018): n. pag.



# Index

- **背景介绍**
  - 概览：大概做了什么(What)
  - 应用场景，有什么用? (Why?)
  - 需要的一些背景知识
    - Semi-Supervised Learning/Transfer Learning
    - Knowledge Distillation
- 实现方法
- 相关工作
- 工作目标





we use this particular model and refer to it as semi-supervised ResNet-50. In the low compression regime (block sizes of 4 and 9), with  $k = 256$  centroids (practical for implementation), our compressed semi-supervised ResNet-50 reaches **76.12% top-1 accuracy**. In other words, the compressed model attains the performance of a vanilla, non-compressed ResNet50 while having a size of 5 MB (vs. 97.5MB for the non-compressed ResNet-50).

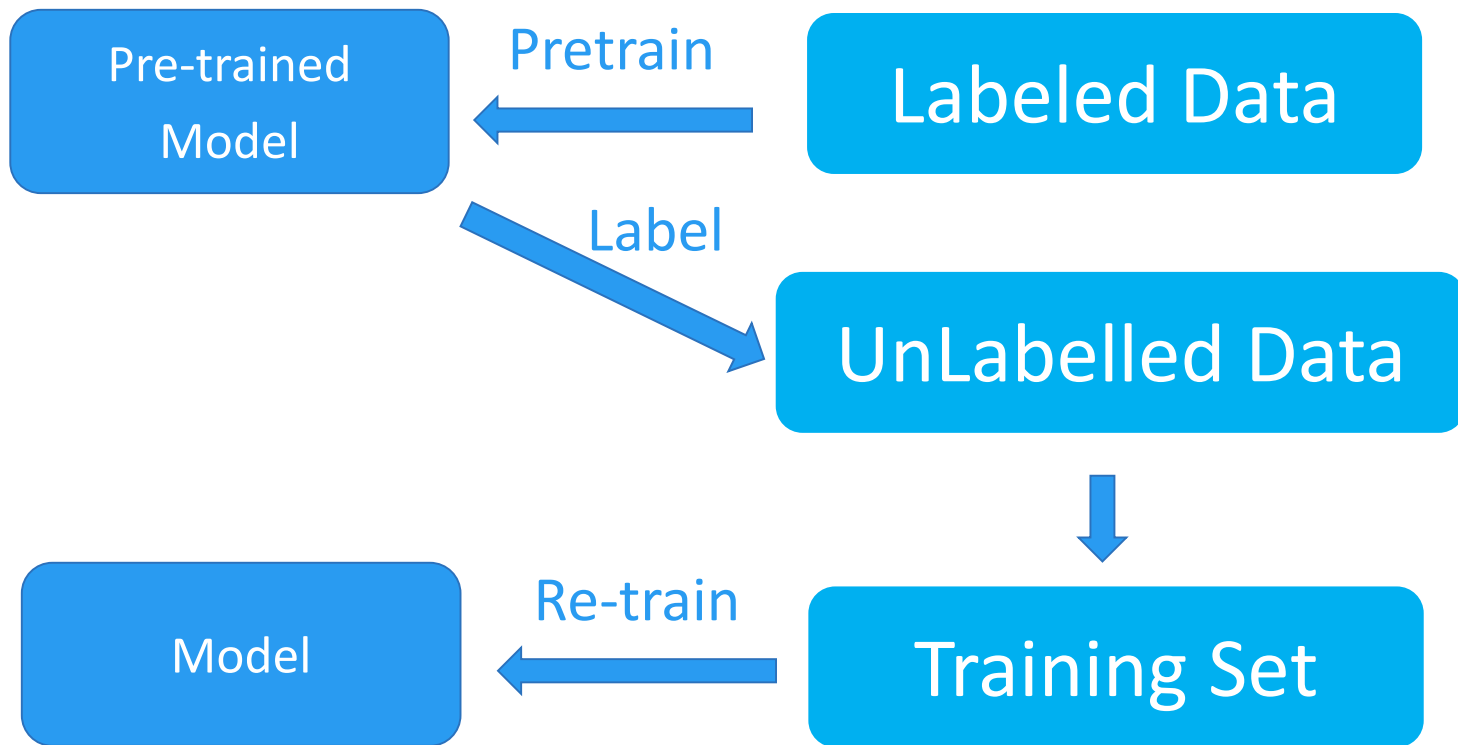
the the set of  $D_l$  of images that are henceforth considered as new positive training examples for that class. We also experiment with using raw predictions before applying the softmax for ranking, but found that they perform worse (see supplementary material). The new image training set collected for the multi-class classification problem is therefore defined as

$$\hat{\mathcal{D}} = \bigcup_{l=1}^L \hat{\mathcal{D}}_l.$$



# Index

- 背景介绍
- 实现方法
  - 对文章《Billion Scale Semi-Supervised》文章中的内容进行阐述
- 相关工作 (对比一些相关的Field以及稍微深入思考一下, 为什么要这样) – Why
  - Self-Trainig
  - Other Semi-Supervised Method
- 工作目标 (自己可能的一些下手点)
  - (目前的计划是想做硬件实现, 怕时间不够所以可能定的题目是在先算法层面做优化?)
  - 这是一种Workflow, 还未尝试过对稀疏网络或者是轻量化的网络
  - 对于数据集如何维护这一层面还未做阐释, 有实验空间
  - 如何利用Class Imbalance-本身这玩意对Training是很麻烦的一个问题





## 《Billion-scale semi-supervised learning for image classification》

- 与Weak-Supervised Warm-Up相结合

	ResNet-50	ResNeXt-101-*		
		32x4	32x8	32x16
Xie <i>et al.</i> [43]	76.1	78.8	-	-
Mixup [45]	76.7	79.9	-	-
LabelRefinery [2]	76.5	-	-	-
Autoaugment [7]	77.6	-	-	-
Weakly supervised [27]	78.2	81.2	82.7	84.2
ours (semi-supervised)	79.1	80.8	81.2	81.2
ours (semi-weakly sup.)	80.9 (81.2 <sup>†</sup> )	<b>83.4</b>	<b>84.3</b>	<b>84.8</b>