Knowledge distillation

Knowledge distillation for RNN-LM

- 《knowledge distillation for recurrent neural network language model with trust regularization》
- RNN >> n-gram
- In many applications, a large RNNLM or an ensemble of several RNNLMs is used.
- Knowledge distillation & trust regularization
- In a speech recognition N-best rescoring task, reduce the RNNLM model size to 18.5% of the baseline with no degradation in WER on WSJ

Knowledge distillation

- 《Distilling the knowledge in a neural network》

背景

ensemble model is used to improve the performance in real applications.
Cumbersome & computationally expensive in deployment

思想

Transfer learning , distillation
Compression, project

实验设计

MNIST / ASR / JTF dataset

思想&方法

思想

“蝴蝶以毛毛虫的形态吃树叶积攒能量逐渐成长，最后变换成蝴蝶这一终极形态完成繁殖。”

——《昆虫记》

蒸馏神经网络，就是从**毛毛虫到蝴蝶**

方法

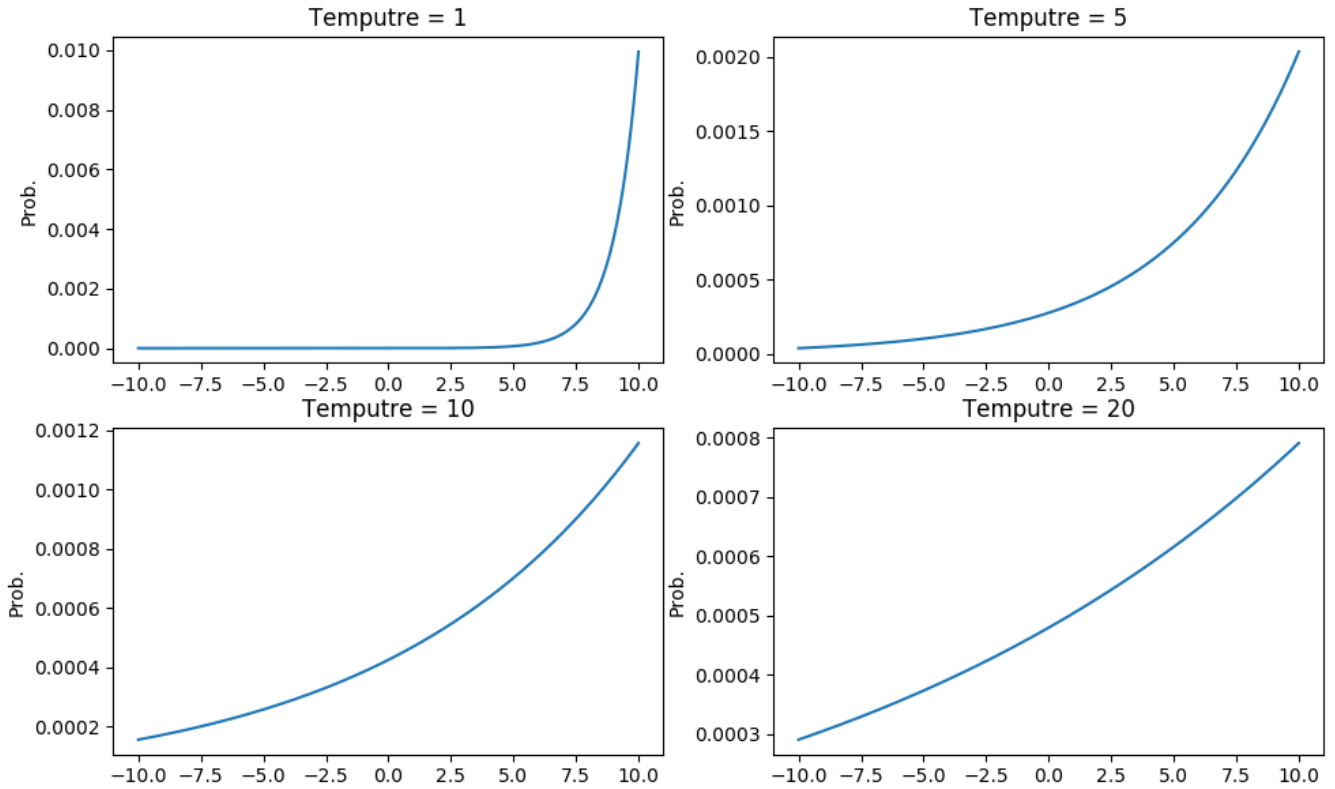
将大模型学习的知识作为先验传递给小规模的神经网络，实际应用时部署小规模神经网络。

Output of softmax → input of softmax → Knowledge distillation

核心公式

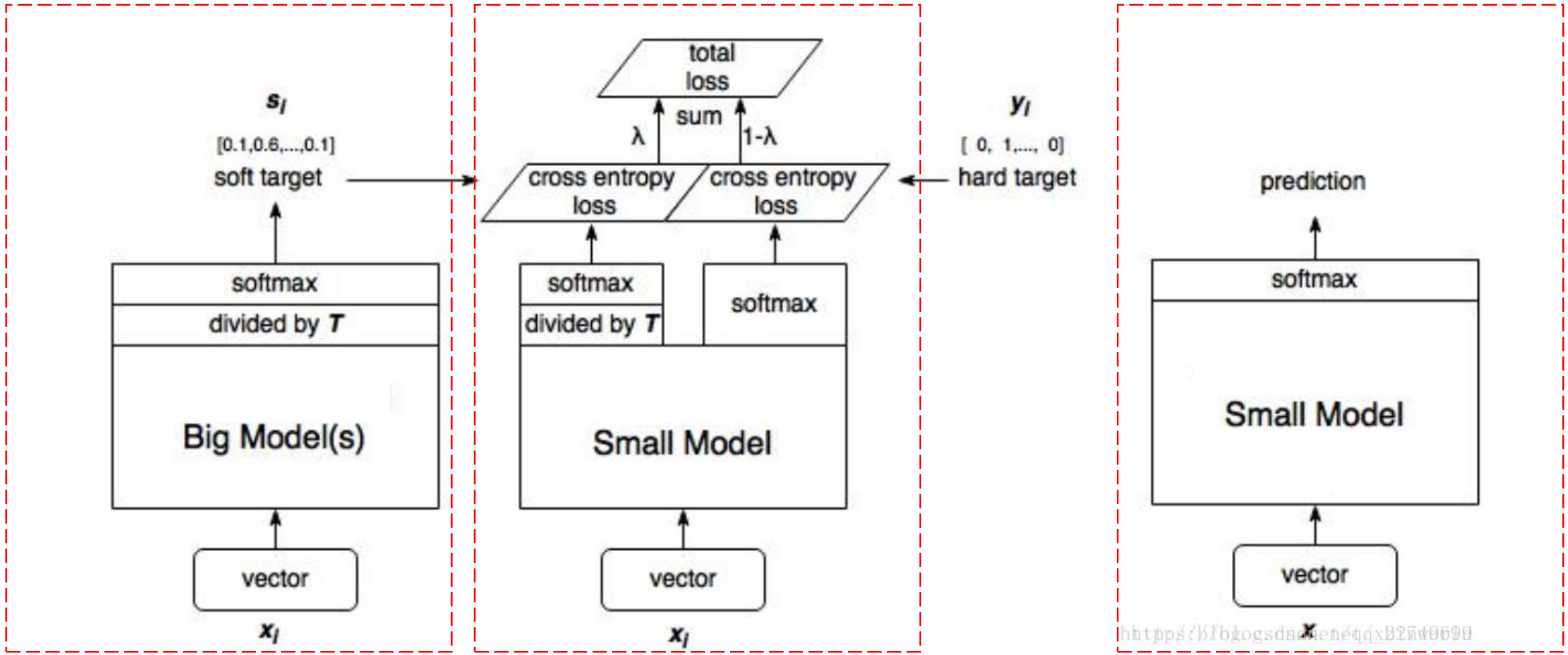
$$q_i = \frac{\exp(z_i/\boldsymbol{T})}{\sum_j \exp(z_i/\boldsymbol{T})}$$

T：温度参数



实验步骤

- 提升softmax的温度参数T，让大模型训练得到一个合适的“软目标”
- 采用同样的T训练小模型，使小模型匹配大模型的“软目标”



MNIST

	NN configs	results	conclusion
baseline	Input-1200-1200-output Dropout & weight-constrains	67	• Soft targets can transfer a great deal of knowledge to distilled model • Bigger than 300-300, T>=8 fairly similar • Input-30-30-output, T∈[2.5,4], the best
SM model	Input-800-800-output No regularization	146	
Distill(T=20)	Input-800-800-output	**74**	

	Experiment configs	results	
Distill model	Omitting all examples 3 from transfer set(**3 is never seen**)	206 (133/1010 3s)	Fine-tune bias for 3
		109(14/1010 3s)	
Distill model	Only containing 7 & 8 from training set(**only 7 & 8 are saw**)	47.3%	Fine-tune bias for 7 & 8
		13.2%	

The distilled model only makes 206 test errors of which 133 are on the 1010 threes in the test set. Most of the errors are caused by the fact that the learned bias for the 3 class is much too low. If the bias is increased by 3.5(which optimizes overall performance on the test set), the distilled model makes 109 errors of which 14 are on 3s. So with the right bias, the distilled model gets 98.6% of the test 3s correct dispite never having seen a 3 during training.

ASR

Training set: 2000h

	NN configs	Test Frame accuracy	WER
baseline	8 layers 2560 nodes 14000 labels	58.9%	10.9%
10 * ensemble		61.1%	10.7%
Distilled single	T {1,**2**,5,10} Weight 0.5 for hard	60.8%	10.7%

Vary the set of data that each model sees, useless in ensemble

More than 80% of the improvement in frame classification accuracy achieved by using an ensemble of 10 models is Transferred to the distilled model.

JFT

- 100 million labeled image with 15000 labels
- Trained for 6 months using asynchronous stochastic gradient descent on a large number of cores
- but only if a lot more cores are available

System	Conditional test accuracy	Test accuracy
Baseline	43.1%	25.0%
+61 specialist models	45.9%	26.1%

Specialist models

- Trained on examples from a very confusable subset of classes (eg: different types of mushroom)
- Class is smaller by combining all of the classes it doesn't care about into a dustbin class
- Initialized with weight of generalist model
- ½ from special subset, ½ sampled from the remainder of the training set

ASR

System & training set	Train frame accuracy	Test frame accuracy
Baseline(100% of training set)	63.4%	58.9%
Baseline(3% of training set)	67.3%	44.5%
Soft targets(3% of training set)	65.4%	57.0%

Conclusion:
soft targets allow a new model to generalize well from only 3% of the training set.

谢谢

https://blog.csdn.net/xbinworld/article/details/83063726
https://blog.csdn.net/haoji007/article/details/76777430
https://www.zhihu.com/question/50519680/answer/136406661
https://www.cnblogs.com/jins-note/p/9679450.html