Ranking Distillation: Learning Compact Ranking Models With High Performance for Recommender System

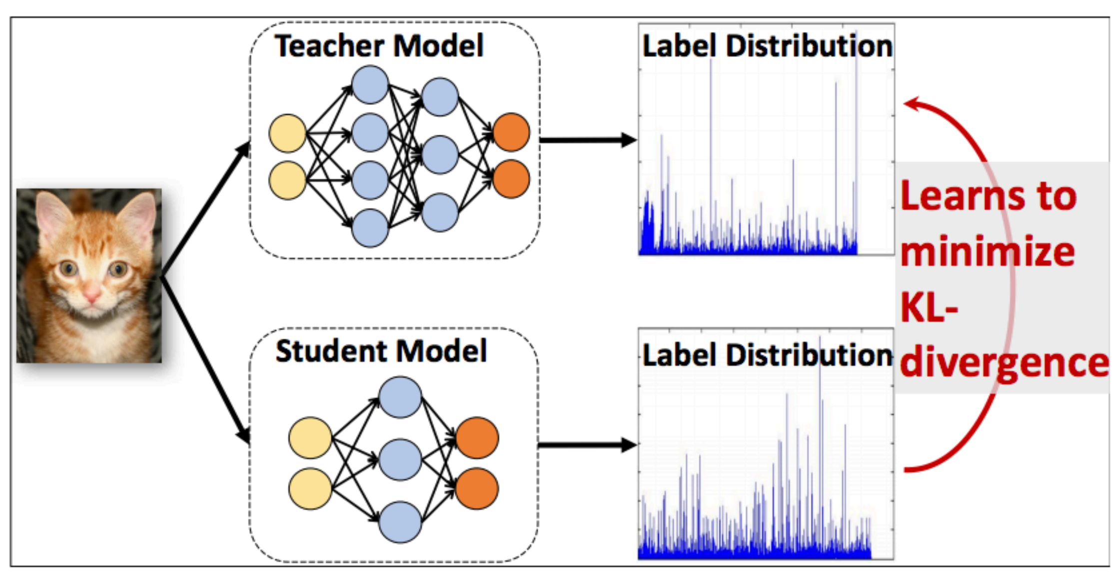
Jiaxi Tang, Ke Wang
School of Computing Science, Simon Fraser University

Abstract

- We try to make effective but expensive model to be compact while still perform well.
- We propose a training paradigm called *ranking distillation* for learning compact ranking models with high performances.
- We use our method on Recommender System, a typical ranking problem.
- Experiments on real world datasets demonstrate the effectiveness of our proposed method.

Knowledge Distillation

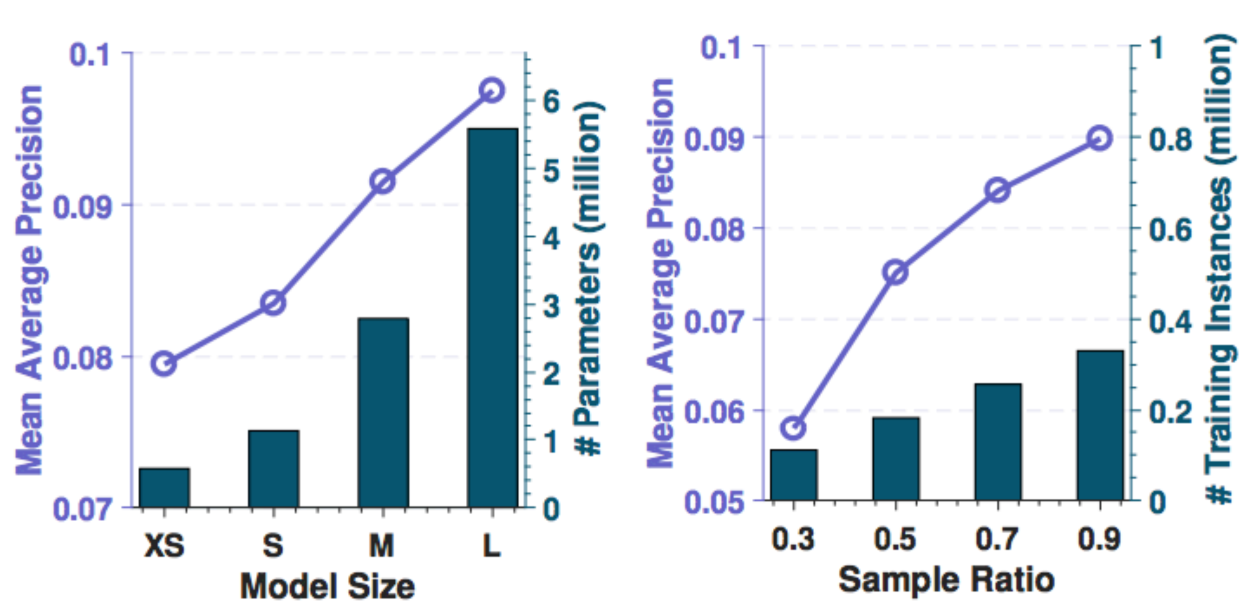
- For image classification, KD first train a teacher model from dataset with many parameters to achieve high performance.
- Then KD train a small student model from the same dataset and the teacher model.
- Eg. For a **cat** image, a well-trained teacher model also supervise the student model to predict **tiger**.



Effectiveness vs. Efficiency

- For a specific ranking model, there are typically two ways to make it perform better:

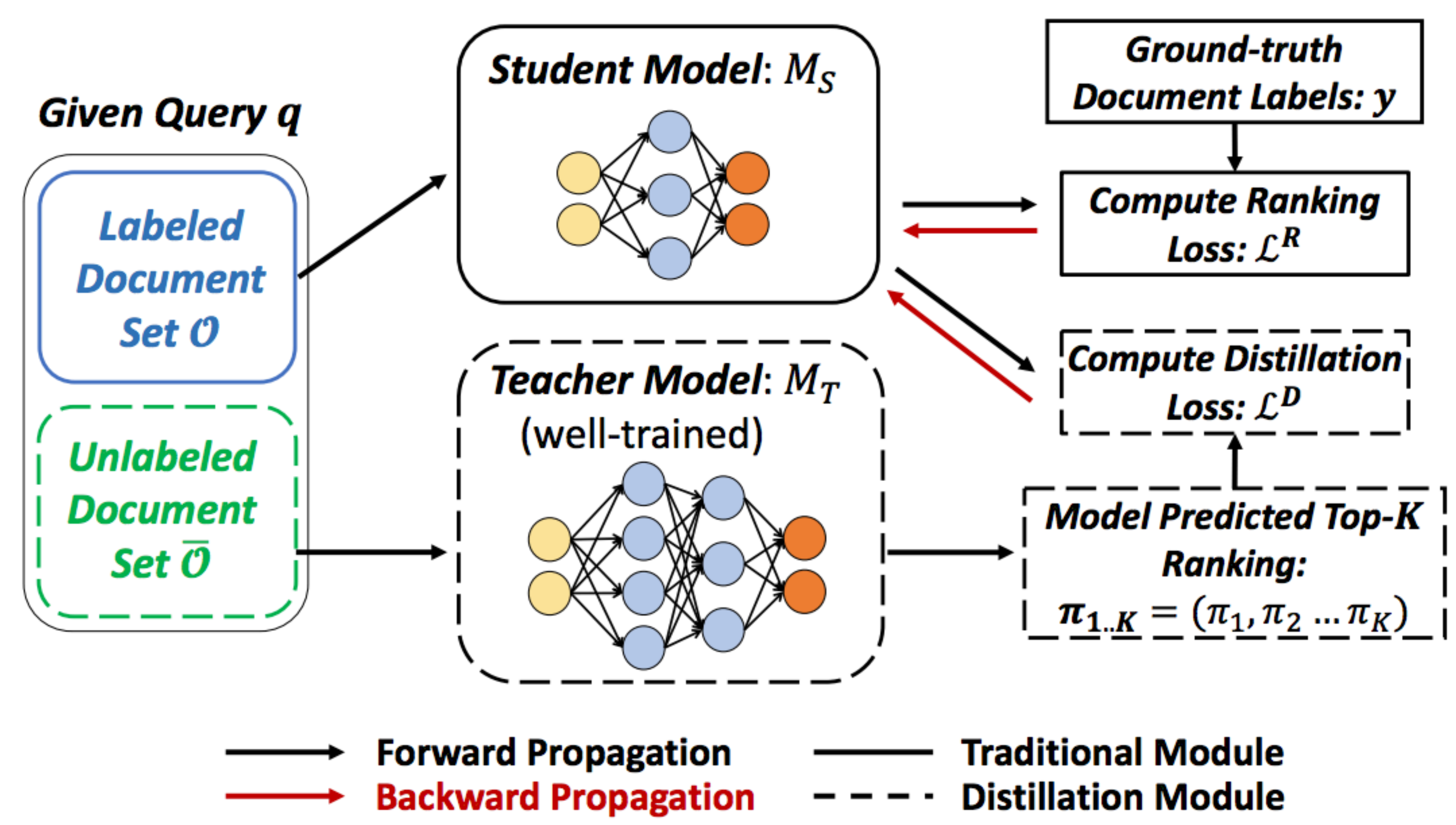
1. By having more parameters until the model get overfitted. (more flexibility and expressiveness)
2. By using more data to train the model. (more generalizable and robust for future data)



(a) MAP *vs.* model size (b) MAP *vs.* training instances

Training Paradigm of Ranking Distillation

- Inspired by KD, we use a well-trained teacher model to provide more training instances to make a student model perform better.
- For a certain query (user profile), we use the top-K ranked documents (items) as the extra positive training instances.



Weighted Point-wise Distillation Loss

- The distillation loss L^D is formulated as a weighted point-wise loss:

$$\mathcal{L}^D(\boldsymbol{\pi}_{1..K}, \hat{\boldsymbol{y}}) = -\sum_{r=1}^{K} w_r \cdot \log(P(rel = 1|\hat{y}_{\pi_r})) = -\sum_{r=1}^{K} w_r \cdot \log(\sigma(\hat{y}_{\pi_r})),$$

$\pi_{1..k}$: teacher's top-K ranked items
$\hat{y}$: student's prediction
$\sigma(\cdot)$: sigmoid function

- *Weighting by position importance w^a*
  Exponentially decayed function, with hyperparameter λ to control the decay speed.
  Assumption: Top ranked items from teacher's prediction are more correlated to the query and the ground-truth positive item

$$w_r^a \propto e^{-r/\lambda} \quad \text{and} \quad \lambda \in \mathbb{R}^+$$

- Weighting by ranking discrepancy w^b
  Non-negtive function to measure how well a student learned from its teacher, with hyperparameter μ to control the pen.
  Assumption: During the training process, we should have a dynamic weight to upweight the erroneous parts in distillation loss, and downweight the parts that already learned perfectly.

$$w^\beta = \tanh(\mu(\text{student's rank} - \text{teacher's rank}))$$

	Teacher's rank	Student's rank
π_1	1	1
π_2	2	5
π_3	3	156

$$\mathcal{L}^D = w_1^b * \log(\hat{y}_{\pi_1}) + w_2^b * \log(\hat{y}_{\pi_2}) + w_3^b * \log(\hat{y}_{\pi_3})$$

$$w_3^b >> w_2^b > w_1^b$$

Experimental Setup

- *Task*: Sequential Recommendation
- *Datasets: Gowalla & Foursquare*
- *Base Model: Fossil & Caser*
- *Baselines:*
  - *Model-T: Teacher model*
  - *Model-S: Student model*
  - *Model-RD: Student model trained with ranking distillation*
- *Evaluation Metrics:*

1) Precision@n $(n \in \{3, 5, 10\})$
2) nDCG@n $(n \in \{3\ 5, 10\})$
3) Mean Average Precision (MAP)

Experimental Results

- Evaluation on model efficiency:

Generating a recommendation list for every user.
Models with less parameters has less inference time cost.

Datasets	Model	Time (CPU)	Time (GPU)	#Params	Ratio
Gowalla	*Fossil-T*	9.32s	3.72s	1.48M	100%
	Fossil-RD	4.99s	2.11s	0.64M	43.2%
	Caser-T	38.58s	4.52s	5.58M	100%
	Caser-RD	18.63s	2.99s	2.79M	50.0%
Foursquare	*Fossil-T*	6.35s	2.47s	1.01M	100%
	Fossil-RD	3.86s	2.01s	0.54M	53.5%
	Caser-T	23.89s	2.95s	4.06M	100%
	Caser-RD	11.65s	1.96s	1.64M	40.4%

- Evaluation on model effectiveness

Models with ranking distillation, Fossil-RD and Caser-RD, always has statistically ***significant improvements*** over the student-only models, Fossil-S and Caser-S

The performance of the models with ranking distillation, Fossil-RD and Caser-RD, ***has no significant degradation*** from that of the teacher models

Gowalla

Model	Prec@3	Prec@5	Prec@10	nDCG@3	nDCG@5	nDCG@10	MAP
Fossil-T	0.1299	0.1062	0.0791	0.1429	0.1270	0.1140	0.0866
Fossil-RD	0.1355	0.1096	0.0808	0.1490	0.1314	0.1172	0.0874
Fossil-S	0.1217	0.0995	0.0739	0.1335	0.1185	0.1060	0.0792
Caser-T	0.1408	0.1149	0.0856	0.1546	0.1376	0.1251	0.0958
Caser-RD	0.1458	0.1183	0.0878	0.1603	0.1423	0.1283	0.0969
Caser-S	0.1333	0.1094	0.0818	0.1456	0.1304	0.1188	0.0919

Foursquare

Model	Prec@3	Prec@5	Prec@10	nDCG@3	nDCG@5	nDCG@10	MAP
Fossil-T	0.0859	0.0630	0.0420	0.1182	0.1085	0.1011	0.0891
Fossil-RD	0.0877	0.0648	0.0430	0.1203	0.1102	0.1023	0.0901
Fossil-S	0.0766	0.0556	0.0355	0.1079	0.0985	0.0911	0.0780
Caser-T	0.0860	0.0650	0.0438	0.1182	0.1105	0.1041	0.0941
Caser-RD	0.0923	0.0671	0.0444	0.1261	0.1155	0.1076	0.0952
Caser-S	0.0830	0.0621	0.0413	0.1134	0.1051	0.0986	0.0874

24th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 18), London, UK