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.

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:
  \[
  L^D(\pi_k; y) = \sum_{k=1}^{K} w_k \cdot \log(P(\text{rel} = 1 | \hat{y}_{\pi_k}))
  \]
- Weighting by position importance $w^p$:
  Exponentially decayed function, with hyperparameter $\lambda$ 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
  \[
  w^p_k = e^{-r/k} / \lambda \quad \text{and} \quad \lambda \in \mathbb{R}^+
  \]
- Weighting by ranking discrepancy $w^r$:
  Non-negative function to measure how well a student learned from its teacher, with hyperparameter $\mu$ to control the pen.
  Assumption: During the training process, we should have a dynamic weight to upweight the erroneous items in distillation loss, and downweight the parts that already learned perfectly.
  \[
  w^l = \tanh(\mu(\text{student’s rank} - \text{teacher’s rank}))
  \]

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

Experimental Results
- Evaluation on model efficiency: Generating a recommendation list for every user. Models with less parameters has less inference time cost.
- 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.

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)