

Introduction to Search Relevance Ranking- Session III – Knowledge Distillation

Tutorial Link: <https://dlranking.github.io/dlrr/>

Data source: <https://huggingface.co/datasets/xglue>

XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation)

Presenters: Xue Li, Keng-hao Chang @ Microsoft Ads

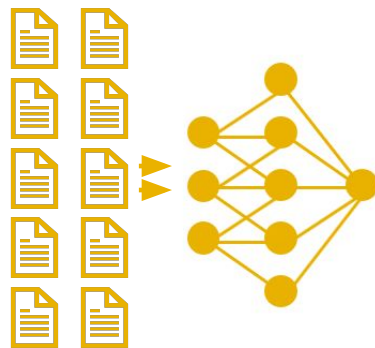
Date: August 14th, 2022

Agenda

- Knowledge distillation
- Case study in DistilBert
- Case study in Microsoft Ads for Ranking
- The colab

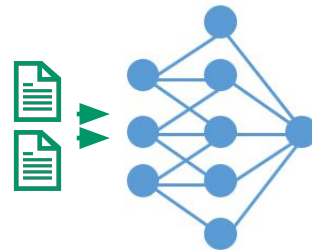
Pre-train and fine-tune

- Natural Language Processing
 - Pre-train on unsupervised tasks, e.g., Language Modeling
 - Fine-tune on downstream NLP tasks, e.g., Question Answering, search relevance
-
- Large & powerful NLP models, even beat human!



Pre-train

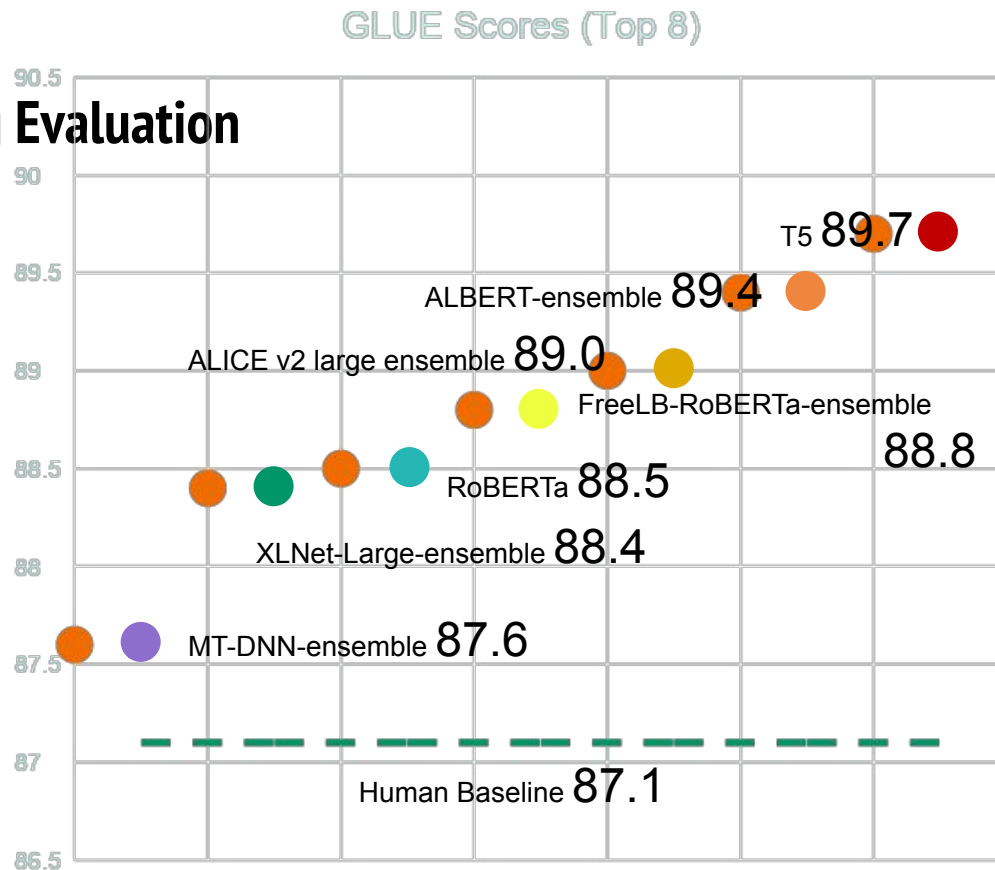
cheap large data
on related domain



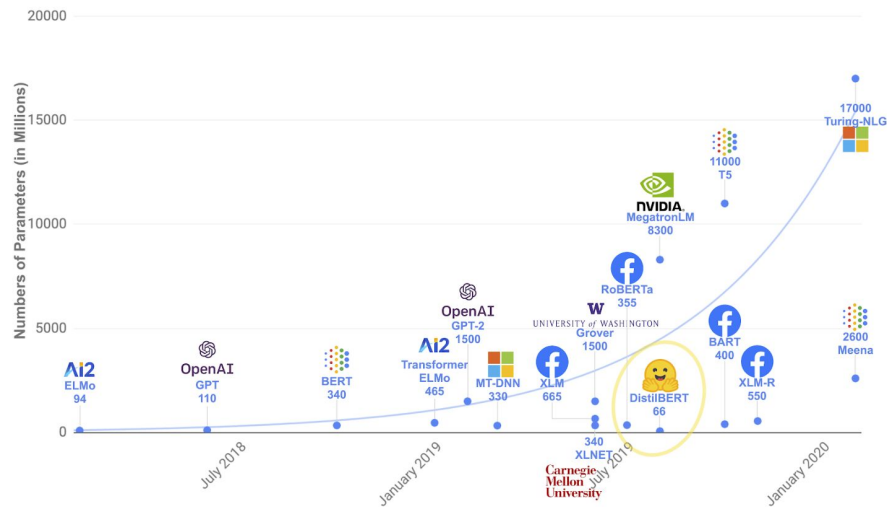
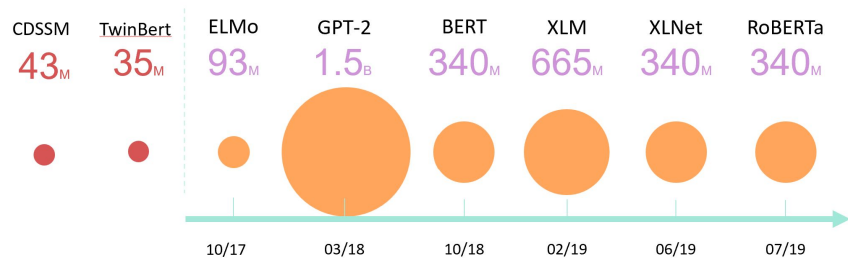
Fine-tune

Expensive well-labeled data
on downstream task

GLUE – General Language Understanding Evaluation



How large are pre-trained NLP models? (and distilled)



[Learning semantic representations using convolutional neural networks for web search | Proceedings of the 23rd International Conference on World Wide Web \(acm.org\)](#) (CDSSM)

[TwinBERT | Proceedings of the 29th ACM International Conference on Information & Knowledge Management](#)

Call outs

Will cover

- Knowledge distillation
- Practices of knowledge distillation

Will not cover

- Other lighter BERT model techniques (Albert, Electra)
- All research directions of knowledge distillation

[\[2006.05525\] Knowledge Distillation: A Survey \(arxiv.org\)](#)

Response level knowledge distillation

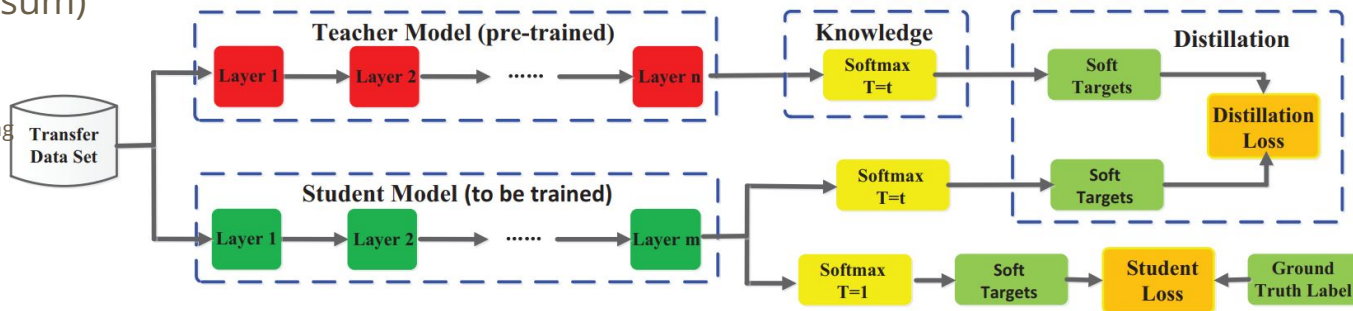
- Distil by learning softmax from teacher on a transfer set

- i.e., soft label, dark knowledge*
- KL divergence
- Vs. Logits

- Two losses (weighted sum)

- Temperature

- Analogous to label smoothing



$$L_{ResD}(z_t, z_s) = \mathcal{L}_R(z_t, z_s) , \quad (1)$$

$$p(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} , \quad (2)$$

$$L_{ResD}(p(z_t, T), p(z_s, T)) = \mathcal{L}_R(p(z_t, T), p(z_s, T)) . \quad (3)$$

[1503.02531] Distilling the Knowledge in a Neural Network (arxiv.org), Hinton

*BERT-base's predictions for a masked token in "I think this is the beginning of a beautiful [MASK]" comprise two high probability tokens (day and life) and a long tail of valid predictions (future, story, world. . .).

Feature-based knowledge distillation

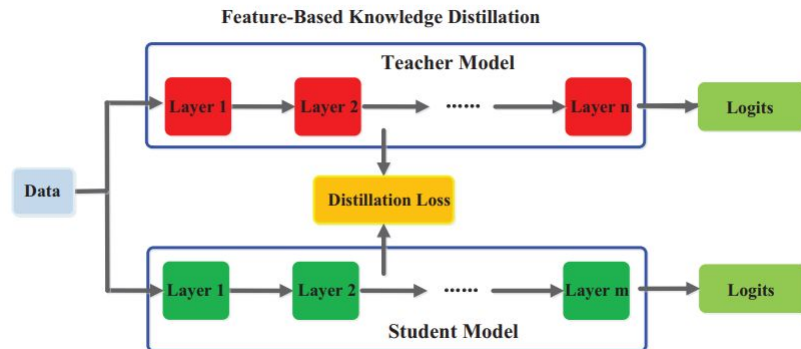
- Learning feature maps of the intermediate layers from teacher to student models
 - L2-norm distance, L1-norm distance, cosine loss etc.
 - Due to the significant differences between sizes of hint and guided layers, how to properly match feature representations of teacher and student also needs to be explored.

$$L_{FeaD}(f_t(x), f_s(x)) = \mathcal{L}_F(\Phi_t(f_t(x)), \Phi_s(f_s(x))) \quad , \quad (4)$$

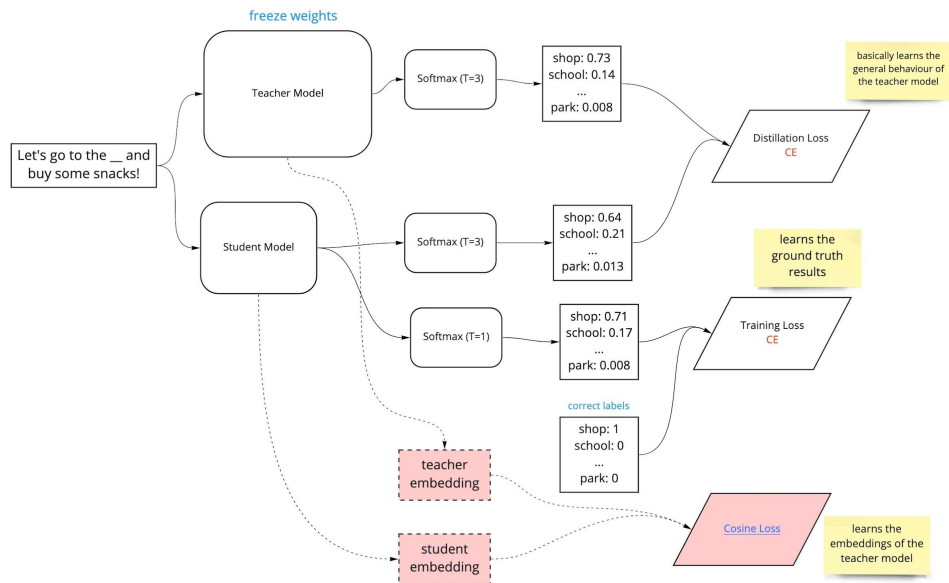
$f_t(x)$ and $f_s(x)$ are the feature maps o

The transformation functions, $\Phi_t(f_t(x))$

$$\Phi_s(f_s(x)), i$$



Case study: distillBert



- 3 losses (for both response & feature)
 - Distillation loss
 - Training loss
 - Cosine loss
$$\mathcal{L}_{cos} = 1 - \cos(h_T, h_S)$$
- Architecture
 - the number of layers is reduced by a factor of 2.
 - Initialize by teacher every other layer
- DistillBERT model retains almost 97% of the original BERT-base model's language understanding when evaluated on GLUE benchmarks. In addition to this, it is 40% smaller and 60% faster at inference.
- General-purpose pre-training distillation rather than a task-specific distillation

Case study: distillBert & distillation variant from Internal Representations

- (1) KL divergence loss across the self attention probabilities of all the transformer heads
- (2) the cosine similarity loss between the [CLS] activation vectors for the given layers

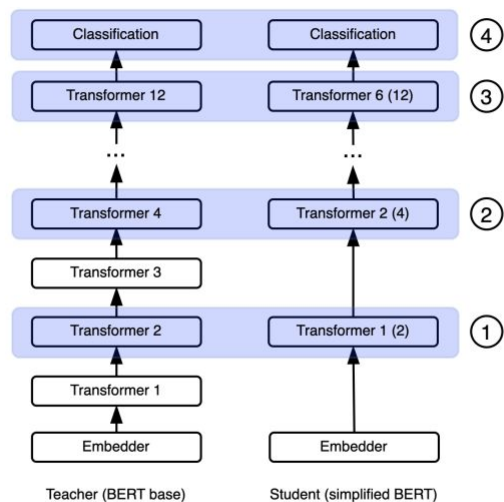


Figure 1: Knowledge distillation from internal representations. We show the internal layers that the teacher (left) distills into the student (right).

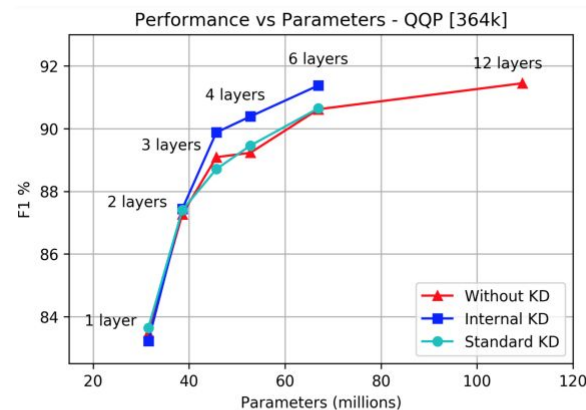


Figure 2: Performance vs. parameters trade-off. The points along the lines denote the number of layers used in BERT, which is reflected by the number of parameters in the x-axis.

Variants of Knowledge Distillation

- Offline distillation
 - Two steps: Pretrain teacher then distill student; one way; capacity gap
- Online distillation
 - Both teacher and student are updated simultaneously
 - E.g., multi-branch architecture, in which each branch indicates a student model and different branches share the same backbone network.
 - E.g., Any one network can be the student model and other models can be the teacher during the training process.
- Self distillation
 - the same networks are used for the teacher and the student model

- Teacher-Student Architecture

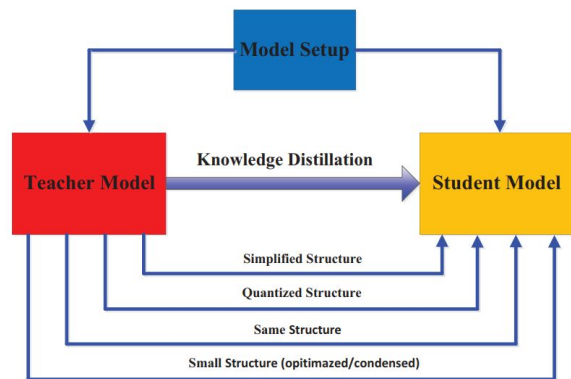


Fig. 9 Relationship of the teacher and student models.

- Adversarial Distillation
- Multi-Teacher Distillation
- Data-Free Distillation

Case study: search relevance ranking at Microsoft Ads

- Point-wise relevance score by against human labels
- Used as externality in ranking
 - $p\text{Defect} = 1 - \text{Relevance}$
 - $\text{RankScore} = \text{Bid} * p\text{Click} - w * p\text{Defect}$

The screenshot shows a Microsoft Ads search results page for the query "adidas tennis shoes women". The page includes a search bar at the top with the query, navigation tabs (All, Shopping, Images, Videos, Maps, News, My saves), and a Microsoft logo. Below the navigation, there are filters for "adidas running shoes women" and "adidas tennis shoes women". The main content area displays several product listings from different retailers, including Tennis Shoe Adidas Womens from tennisexpress.com, Stockx.com, and Women's Adidas Shoes at JCP. Each listing includes a product image, title, price, and retailer information. On the right side, there is a section titled "See adidas tennis shoes women" with a grid of product images and a link to "See more shopping results for adidas tennis shoes women".

adidas tennis shoes women

All Shopping Images Videos Maps News | My saves

Microsoft Show results from Microsoft >

adidas shoes > women > adidas tennis shoes women

adidas running shoes women adidas tennis shoes women

Tennis Shoe Adidas Womens - tennisexpress.com
<https://www.tennisexpress.com/womens>
Ad Tennis Shoe Adidas Womens. Free Shipping and Free Returns*
Ratings: Price 4/5 - Customer Service 4/5 - Return Policy 4/5
Shop Now

Tennis Shoes
All the latest releases from top brands. Men, Women & Junior!

Tennis Racquets
We carry a giant selection and have a racquet for every player

Tennis Strings
We offer the most extensive selection of tennis strings

Tennis Bags
Tennis Bags, Backpacks and Racquet Covers. Shop Now!

Shop by Brand
Nike, Babolat, too many to list! Order Today, Ships Today!

Tennis Apparel
Look good while you play. Top brand selection of apparel.

See results only from tennisexpress.com

Stockx.com
<https://stockx.com>
Ad Buy and Sell 100% Authentic Deadstock Sneakers. Adidas Yeezy, Jordans & More. Yeezy 350, Yeezy 500, Adidas NMD, Adidas Ultra Boost, Jordan 1, Jordan 3 + More.

Women's Adidas Shoes at JCP® - Free Shipping Every Day
<https://www.jcpennney.com/adidasshoes>
Ad Shop Women's Adidas Shoes at JCP®. Save Big on Brands You Love Today at JCPennney®. Earn Rewards Points - Save With Coupons - Free Shipping to Stores

See adidas tennis shoes women

Ads 1

Adidas Women's ... \$47.99 Amazon.com 4.5 stars 34

Adidas Women's ... \$159.95 Amazon.com Free Shipping

Adidas Women's ... \$143.98 Amazon.com 4.5 stars 44

Adidas Women's ... \$70.00 Macy's 18% price drop

Adidas Women's ... \$33.97 Amazon.com Free Shipping

Adidas Women's ... \$64.35 Amazon.com 4.5 stars 3

Adidas Women's VI ... \$42.99 Amazon.com 18% price drop

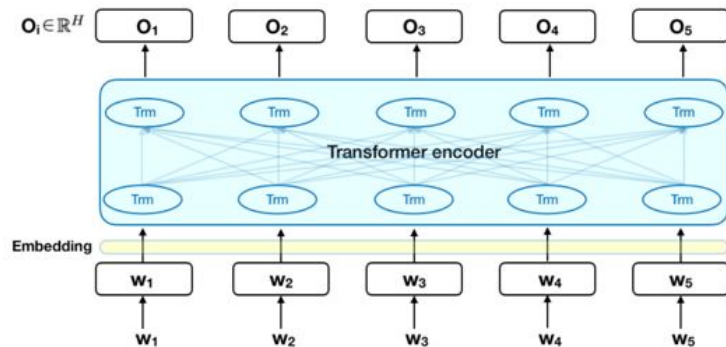
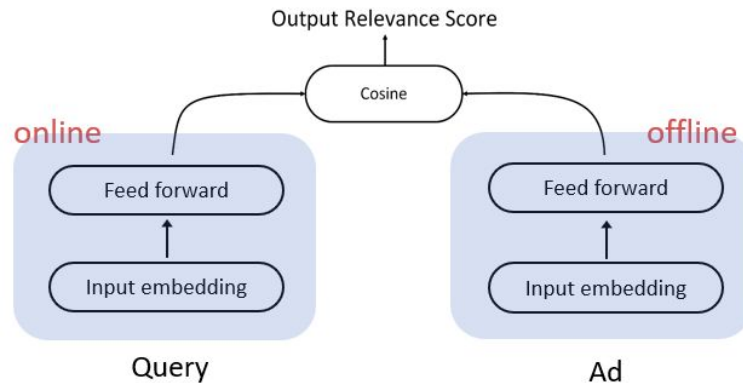
Adidas Women's ... \$100.00 Amazon.com Free Shipping

See more shopping results for adidas tennis shoes women

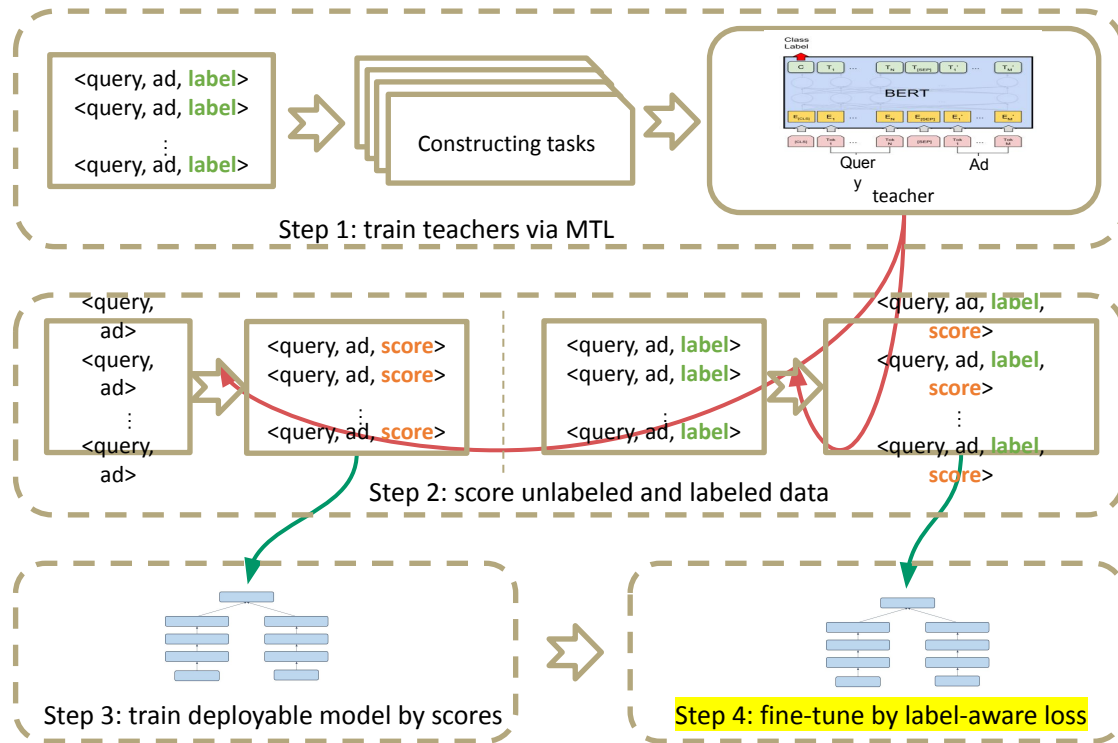
See your ad here >

Choice of student model

- Two tower
 - CDSSM, TwinBert
 - Doc embedding is offline computed
- BERT-like
 - cannot support fast compute, latency prohibitive.



Knowledge distillation for search Relevance



Teacher Training

- Narrow the gap between pre-trained models and target tasks

Inference Data

- Score both labeled / unlabeled data

- Train a deployable student model
- Using scored unlabeled data

Student Fine-tuning

- Fine-tune student model
- Using scored labeled data

Recipe of AdsBERT Distillation



Initialization

Pre-trained BERT
340M params



Pretrain

MLM/NSP
400M Ads data



MTL Finetune

8 ad tasks
40M samples



Inference

Vast amount
Proper distribution



Distillation

CDSSM keep **70%**
AUC gain

Case study: TwinBERT

- Two tower Bert
- Pooling & crossing layer

Table 2: ROC-AUC of TwinBERT models comparing with C-DSSM, BERT₃ and BERT₁₂ on two test sets

Model	AUC ₁	AUC ₂
C-DSSM	0.8713	0.8571
BERT ₃	0.8995	0.9107
TwinBERT _{cos}	0.8883	0.8743
TwinBERT _{res}	0.9010	0.9113
BERT ₁₂	0.9011	0.9137

Table 3: Density differences of all 4 labels by comparing top 5 results from TwinBERT_{cos} and C-DSSM

bad	fair	good	excellent
-7.4%	-2.6%	1.9%	18.8%

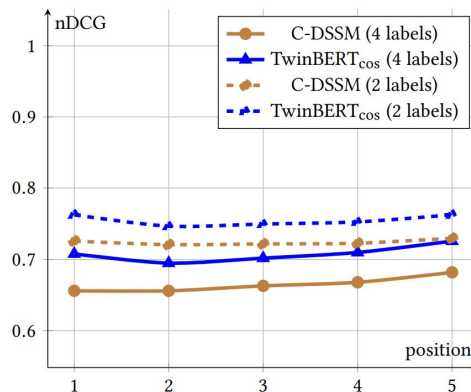


Figure 3: nDCGs of TwinBERT_{cos} and C-DSSM

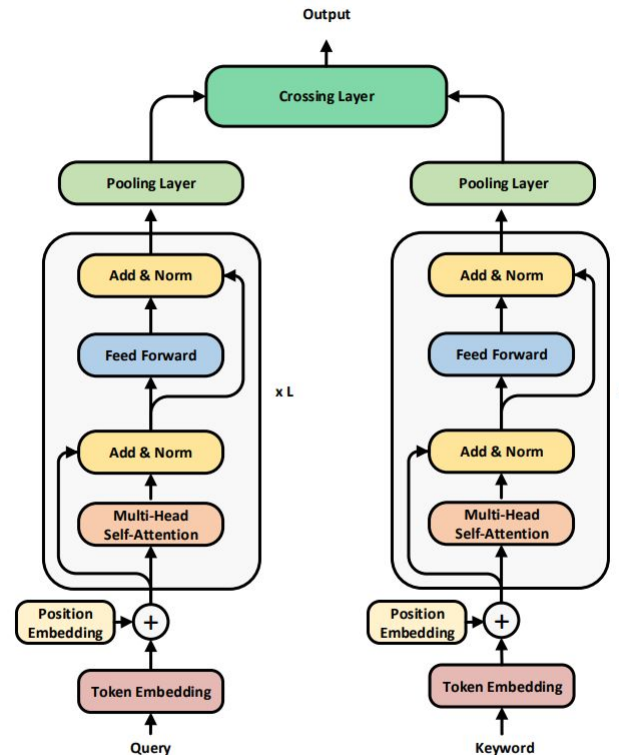


Figure 1: TwinBERT Architecture

Colab

- QADSM task in [xGLUE](#) dataset, which is extracted from real Bing Ads traffic.