1 Response to all Reviewers

Thank you for the valuable suggestions. We will cite missing relevant works and add the results as you suggest.

Comparison to MobileBERT. We distill a base-sized teacher, which achieves similar performance as the specially designed teacher of MobileBERT, into a same-sized student (25.3M parameters) using the same training data as MobileBERT. We compare with MobileBERT w/o OPT (without operational optimizations) and compute the speedup over $\text{BERT}_{\text{BASE}}$ according to their reported latency. As shown in Table 1, MiniLM achieves competitive results with faster speed. Moreover, our method can be applied for different teachers and has much fewer restrictions of students.

Model	#Params	Speedup over $\text{BERT}_{\text{BASE}}$	SQuAD 2.0	MNLI-m	QNLI	SST-2	MRPC
MobileBERT w/o OPT	25.3M	1.8×	80.2	84.4	91.5	**92.5**	87.0
L12-H384-E128 MiniLM	25.3M	2.7×	**80.6**	**85.2**	**91.7**	92.1	**89.5**

Table 1: Results of 12-layer MiniLM with 384 hidden size and 128 embedding size.

Visual analysis of attention weight transfer and Why distilling the last layer works better? We visualize the attention distributions of the teacher ($\text{BERT}_{\text{BASE}}$) and the 6x384 students using last layer and layer-wise distillation. We find that attention distributions for each layer of the layer-wise distilled student are very similar to its corresponding layers of the teacher. For the last-layer distilled student, attention distributions of the last layer mimic its teacher's last layer very well, while the bottom four layers are more similar to the teacher's bottom four layers. The fifth student layer is similar to teacher's middle or top layers. Last layer distillation also learns features of teacher's lower layers. Moreover, layer-wise transfer sets a tight restriction for each student layer. Transferring the last layer gives more flexibility for the bottom layers to learn the knowledge. Experiments (Table 8 of supplementary material) also demonstrate that relaxing restrictions of layer mapping improves performance. We will add the visual analysis in the revised paper.

2 Response to Reviewer #1

Inconsistent results with TinyBERT paper. We focus on task-agnostic compression of pre-trained Transformers. As stated in the caption of Table 2, we compare task-agnostic distilled models without task-specific distillation (TD) and data augmentation (DA). Besides task-agnostic distillation, TinyBERT further uses TD and DA to achieve improvements for specific tasks, and reports the overall results. For a fair comparison, we fine-tune the latest version of their public task-agnostic model (using the same fine-tuning code and range of hyper-parameters) to report the results.

Using $\text{BERT}_{\text{LARGE}}$ as teacher. Thanks for the suggestion. We are exploring it and will add the results in the future.

3 Response to Reviewer #2

Which TinyBERT model we used? Besides the model with 14.5M parameters, TinyBERT also reports the results of 6x768 (6 layers, 768 hidden size, 66M parameters) model in Table 10 and 11 of their paper, and releases both the two models. Since most of previous works, such as DistilBERT and BERT-PKD, distill $\text{BERT}_{\text{BASE}}$ into a 6x768 student, we also adopt the same setting and compare with 6x768 TinyBERT.

Analysis to what extent representations of the lower-level layers differ between the student and teacher model. Thanks for the suggestion. As stated above, visual analysis suggests that attention distributions of student's bottom four layers are similar to teacher's bottom four layers. We will try the methods you suggest and add the discussion.

4 Response to Reviewer #3

Why not query-query or key-key relation but value-value relation? We have tried to add query-query or key-key relation, which improves the performance if we do not transfer attention distributions. But if we introduce attention distributions, it will not bring improvements. Attention distributions capture the relation between queries and keys. Knowledge of queries and keys can be transferred via distilling attention distributions. Transferring query or key relation has a similar effect as transferring attention distributions. Besides queries and keys, the remaining important vectors in self-attention module are values, so we introduce the value relation to achieve a deeper mimicry.

5 Response to Reviewer #4

Why scaled dot-product between values performs better? 1) Since self-attention module is vital in Transformer (attention is all you need), the key idea of our method is to deeply mimic the self-attention. Queries, keys, and values are the most basic and important vectors in self-attention, so we transfer attention distributions (relation between queries and keys) and value relation to achieve a deeper mimicry. 2) Using dot-product converts different dimensional vectors into relation matrices with the same size. Compared with transferring value vectors, it avoids introducing additional parameters (randomly initialized) to transform student vectors into the same size as its teacher. The additional transformation transforms vectors into another vector space and restricts teacher from directly transferring knowledge.

How the student is initialized? Our students are randomly initialized. We will make it clear in the revised paper.

Analysis of value relation. Thanks for the suggestion. We will add the analysis in the revised paper.