SimTDE: Simple Transformer Distillation for Sentence Embeddings

Jian Xie* †
jianxie@Knights.ucf.edu
University of Central Florida
Orlando, Florida, USA

Xin He*
xih@amazon.com
Amazon Alexa
Cambridge, Massachusetts, USA

Jiyang Wang
jiyangw@amazon.com
Amazon Alexa
Cambridge, Massachusetts, USA

Zimeng Qiu
zimengqi@amazon.com
Amazon Alexa
Bellevue, Washington, USA

Ali Kebarighotbi
alikeba@amazon.com
Amazon Alexa
Cambridge, Massachusetts, USA

Farhad Ghassemi
gfarhad@amazon.com
Amazon Alexa
Bellevue, Washington, USA

ABSTRACT

In this paper we introduce SimTDE, a simple knowledge distillation framework to compress sentence embeddings transformer models with minimal performance loss and significant size and latency reduction. SimTDE effectively distills large and small transformers via a compact token embedding block and a shallow encoding block, connected with a projection layer, relaxing dimension match requirement. SimTDE simplifies distillation loss to focus only on token embedding and sentence embedding. We evaluate on standard semantic textual similarity (STS) tasks and entity resolution (ER) tasks. It achieves 99.94% of the state-of-the-art (SOTA) SimCSE-Bert-Base performance with 3 times size reduction and 96.99% SOTA performance with 12 times size reduction on STS tasks. It also achieves 99.57% of teacher's performance on multi-lingual ER data with a tiny transformer student model of 1.4M parameters and 5.7MB size. Moreover, compared to other distilled transformers SimTDE is 2 times faster at inference given similar size and still 1.17 times faster than a model 33% smaller (e.g. MiniLM). The easy-to-adopt framework, strong accuracy and low latency of SimTDE can widely enable runtime deployment of SOTA sentence embeddings.

CCS CONCEPTS

• **Information systems** → **Language models**; **Similarity measures**.

KEYWORDS

Knowledge Distillation, Sentence Embeddings, Semantic Text Similarity, Entity Resolution

ACM Reference Format:
Jian Xie, Xin He, Jiyang Wang, Zimeng Qiu, Ali Kebarighotbi, and Farhad Ghassemi. 2023. SimTDE: Simple Transformer Distillation for Sentence Embeddings. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592063

*Both authors contributed equally to this research.
†This work was performed as part of an internship at Amazon Alexa.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3592063

1 INTRODUCTION

Sentence embeddings capture the lexical and semantic meaning of given text and are fundamental for Information Retrieval [13]. It allows quantitative assessment of whether text pairs share similar meaning, e.g. semantic textual similarity (STS) task, and allows linking between references of the same canonical entity, e.g. entity resolution (ER) task [7]. ER is a crucial component of the spoken language understanding pipeline for voice assistants (VAs) such as Amazon Alexa and Apple Siri, where ER resolves entity mentions in user utterances to actionable entities stored in catalogs [16].

The state-of-the-art (SOTA) of sentence embeddings learning has been advanced by transformer [25] model architecture [6] [17]. However, transformer's high computational complexity and memory consumption poses challenges to deploy SOTA sentence embeddings in runtime. Knowledge distillation (KD) [11] has been shown as a promising method to compress transformer model [19] [21] [12] [26]. KD transfers knowledge from a large teacher network to a small student network [11], such that the student model mimics the teacher model obtaining competitive or even superior performance.

In this paper, we introduce SimTDE, a simple yet effective transformer model KD framework for sentence embeddings. Existing KD methods often start with a task-agnostic distillation design to benchmark with BERT [5] on multiple NLP tasks and have evolved into multi-steps distillation with complicated distillation objectives which is computational costly, time consuming and constrains the student model architecture design. Specifically, DistillBERT [19] requires student hidden size to match with the teacher and needs soft target probabilities in loss objectives; BERT-PKD [21] requires layer-to-layer distillation and also soft target probabilities; TinyBERT[12] requires layer-to-layer distillation plus hidden states and self-attention distillation and relies on a two-stage-distillation: general then task-specific, miniLM [26] requires self-attention distillation and distillation assistant intermediary models.

SimTDE presents a single stage distillation with loss objective composed of only token embedding loss and output embedding loss. For the student model we use a compact token embedding block and a shallow encoding block connected with a projection layer. This design offers significant latency benefits, allowing 2

times faster than deep and narrow student architectures used in other study [26] under similar model size. The projection layer allows the token embedding and encoding blocks to have different dimensions. It can effectively further compress distilled transformer model where token embedding block takes up a major size portion. In the encoding block, SimTDE keeps the same dimension as the teacher with a reduced number of transformer layers. We initialize encoder with teacher's weights without additional hidden layer and self-attention distillation.

We summarize our main contributions as follows: (1) We propose a simple yet novel KD framwork, SimTDE, to effectively compress large and small transformer-based sentence embedding models. (2) We conduct extensive experiments and demonstrate SimTDE can achieve 99.94% of SOTA SimCSE-bert-base performance with $X3$ size reduction and 96.99% SOTA performance with $X12$ size reduction on STS tasks. Additionally, it achieves $>$ 99.5% of teacher's performance on multi-lingual ER data set with a tiny transformer student model of $1.4M$ parameters and $5.7Mb$ size. (3) We compare SimTDE inference time with a collection of transformer models and demonstrate at least $X3$ faster than teacher and $X2$ faster than other distilled student of the same size.

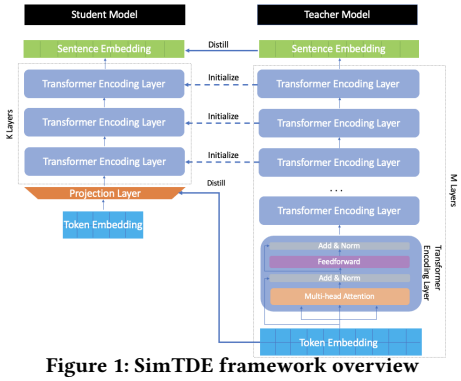

Figure 1: SimTDE framework overview

2 RELATED WORK

Universal sentence embeddings learning has been extensively studied in literature such as InferSent [4], Universal Sentence Encoder [3] and Sentence-BERT [17]. Recently, researchers have adopted new techniques for further improvement, such as data augmentation and contrastive learning. ConSERT [28] uses a combination of four data augmentation strategies: adversarial attack, token shuffling, cut-off, and dropout. SimCSE [6] proposes a simple yet effective approach with contrastive objectives and dropout noises and advances state-of-the-art performance.

Model compression techniques [9] aim to reduce model size and accelerate inference while preserving performance on target tasks. Common approaches of model compression includes quantization which uses fewer bits to represent parameter weights [8], weight pruning [10] which reduce or dilute network connections and knowledge distillation that extract desired knowledge from teacher model and distill it into a student model of smaller size [18, 22, 23]. Knowledge distillation has proven to be a promising method for model compression. [11] first proposed using the soft target distribution to train the student model and impart the knowledge of teachers to students. DistilBERT [19] uses soft-label cross-entropy loss and cosine hidden-state embedding loss between teacher and student. It requires the student to have the same hidden size as the teacher. BERT-PKD [21] further incorporates loss from output of intermediate transformer layers. TinyBERT [12] adds additional self-attention distillation. It introduces learnable projection matrices between teacher and student layers to remove the limit on model compression but requires layer-to-layer distillation and needs a general and a task-specific distillation stage. MiniLM [26] further improves with deep self-attention distillation and introduces a intermediary teacher assistant model.

3 METHODOLOGY

SimTDE's framework is illustrated in **Figure 1**. We first propose a compact token embedding block. Specifically, we define $d_S^{TE} << d_T^{TE}$ where d_S^{TE} and d_T^{TE} are the token embedding dimension of the student and teacher model. The total reduction effect is factored by the vocabulary size v ($v >> d_{S/T}^{TE}$). For a large transformer model, this easily reduces size with minimal performance impact. For a compact transformer model, this further reduces model size significantly as the token embedding layer consumes a major size portion, e.g. token embedding accounts for 55% of MiniLM-384-L6 parameters and 91% of BertTiny-128-L2.

For the encoding block, we propose to use a shallow and wide block which demonstrates better performance and significant latency improvement (See Section 4) compared to a narrow and deep design [26]. Our encoding layers dimension matches with the teacher's instead of the student token embedding as in [12] [26], and we reduce the number of encoding layer to K. We introduce a projection layer to align the compact token embedding output with the teacher encoding dimension. This does not need learnable projection matrices to distill every hidden layer [12] which is computationally intensive. With the aligned dimension, we further simplify the framework to reuse the selected last K teacher encoder layers' weights to initialize the student encoder without additional layer-to-layer distillation as in [21]. This allows a simple loss objective composed of only the token embedding and output embedding distillation. We introduce a hyper-parameter α to adjust the weight of each loss component ($\alpha = 0.5$ in our experiments), detailed as follows:

$$TokenEmb_T = Emb_T(I_T), \quad TokenEmb_S = Emb_S(I_S) \tag{1}$$

$$TokenEmb_S^{proj} = Proj(TokenEmb_S) \tag{2}$$

$$L_{TE} = MSE(TokenEmb_S^{proj}, TokenEmb_T), \tag{3}$$

$$L_{SE} = MSE(SentenceEmb_S, SentenceEmb_T) \tag{4}$$

$$L_{KD} = \alpha L_{TE} + (1-\alpha) L_{SE} \tag{5}$$

where S and T are the subscription labels student and teacher networks respectively; I is the tokenized input; Emb is the token embedding layer; $TokenEmb$ is the generated token embedding ($TokenEmb_T \in \mathbb{R}^{d_T^{TE} \times v}$, $TokenEmb_S \in \mathbb{R}^{d_S^{TE} \times v}$); $SentenceEmb$ is the generated sentence embedding; $Proj$ is the added projection layer; $TokenEmb_S^{proj} \in \mathbb{R}^{d_T^{TE} \times v}$ is the projected student token embedding which has the same dimension as the teacher; L_{KD} is the full distillation loss composed of the token embedding loss L_{TE} and the sentence embedding loss L_{SE}. It is worth noting that we relax the distillation from using ground truth label. We fully rely on the intermediary and final output of the teacher model in

a semi-supervised manner. Therefore, no labeled training data is needed which can reduce data annotation cost.

4 EXPERIMENT

We conduct our experiments on standard semantic textual similarity (STS) task and Entity Resolution (ER) task. Our STS tasks follow a zero shot setup where STS data sets are only used in testing and not involved in model training. Full model is constructed with a dual-encoder design [17], which holds a sentence embedding model in siamese and triplet network structures and derive embeddings for each sentence to a common embedding space for similarity comparison. This is widely used for runtime sentence-pair modeling and has inference speed advantage over cross-encoder.

All models use AdamW [14] as the optimizer. In all tasks, unless noted otherwise, we create final representations using mean pooling over all tokens. We train our base models (e.g. small model without KD) on a server with 4 * V100 (16GB) GPUs and distill from large models on a server with 8 * A100 (40GB) GPUs. All main experiments have the same fixed random seed.

4.1 Semantic Textual Similarity (STS) task

STS is a standard natural language processing task to quantitatively assess the semantic similarity between text pairs. Following [6, 17, 20], we use Spearman rank correlation to measure the correlation quality between calculated similarity and human labels to assess the performance on 7 STS tasks from STS Benchmark [2] and SICK-Relatedness [15]. Spearman correlation ranges from -1 and 1 and increases when predicted similarities ranks align with groundtruth.

We use large (supervised-SimCSE-Bert-Base) and compact (MiniLM) transformer as our teacher models and distill student models to different sizes using SimTDE. We implemented supervised-SimCSE-Bert-Base according to [6] and trained model with QQP, QNLI, MRPC, NLI datasets: SNLI [1], MNLI [27] and NLI for SimCSE [6].

Next, we compare our SimTDE students model with SOTA sentence embedding models including Sentence BERT models [17], SimCSE [6] and recent TENC-mutual [13], and we also compare with different KD methods. Comparisons are performed in the aspects of accuracy and model size, measured by the number of parameters. Lastly, we perform inference time analysis of models with different size and distilled with different methods.

4.2 Entity Resolution (ER) Task

ER in voice assistants resolves entity mentions in user utterance, query, with canonical entities from catalogs, during which query and each catalog entity are paired up for relevancy prediction. The performance is measured by Recall@1(R@1) which calculates the percentage that the predicted top-1 relevant entity of a given query matches the groundtruth. We construct the data set from radio station voice search in 5 European languages: Italian, German, English, French and Spanish, in the format of *(query, entity, binary label)*. Train, dev, test data size are respectively 11M, 1M and 1M.

In the ER tasks, we use SimTDE to push already compact transformer model to extreme to enable application under limited computational resources and tiny footprint budget (e.g. edge devices). We use 17MB BERT-Tiny [24] model with 128 embedding dimension and 2 transformer encoding layers as teacher and distill student models to size below 10MB while having strong performance.

4.3 Results

Accuracy: **Table 1** shows that via proposed SimTDE the student model achieves 99.94% of SOTA STS performance, exceeding widely used SentenceBert embedding by 7.5%, while having 31% size reduction. Also, SimTDE is capable of further distilling compact transformer model and achieves 96.99% of SOTA STS performance, exceeding SentenceBert embedding by 4.27% [17], while having 9% of the size. Specifically, we denote model in the convention of $method-token\_embedding\_dim-encoder\_layer\_num(encoder\_layer\_dim)$. The SimTDE-384-L3(768) student model is distilled from supervised-SimCSE-Bert-Base by reducing the token embedding size from 768 to 384 and reducing the encoding layer number from 12 to 3, and initialize with the last 3 encoder layers' weights from teacher. The SimTDE-128-L3(384) student model is distilled from MiniLM by reducing the token embedding size from 384 to 128 and reducing the encoding layer from 6 to 3 and with similar initialization strategy.

In **Table 2**, we compare SimTDE with other KD methods: 1) Layer Reduction KD (LRN-KD): distill a small model with the same hidden size as the teacher and reduce the encoder layers inspired by BERT-PKD [21], but we did not use target soft probabilities to keep the semi-supervised setting. The LRN-KD-768-L2(768) student model has a token embedding dimension of 768 and 2 encoding layer of 768D (last 2 layers from teacher) 2) Learnable Matrix KD (LMX-KD): build a small BERT model with smaller hidden size and add learnable linear projection matrices to perform layer-to-layer distillation inspired by Tiny-Bert [12]. We only distill the hidden states not the self-attention in the encoding block and the training time per epoch is already $X5$ longer. The LMX-KD-384-L12(384) student model has a token embedding dimension of 384 and 12 encoding layer of 384D (distilled each layer from teacher). SimTDE demonstrates superior performance than both.

Besides distilling large transformer model, we also distill from already compact transformer. To our knowledge this has been rarely explored in other sentence embedding distillation literature which often directly fine tunes the compact model on target tasks. Our SimTDE allows for different distillation strategy in token embedding and encoding blocks, focusing on dimension size reduction and depth reduction respectively. When the encoding block is already shallow we can still effectively reduce the overall size by compacting token embedding. On STS task, **Table 2** demonstrates that SimTDE is able to distill MiniLM to 41.85% of the size while retaining 99.00% of the performance. On ER task, **Table 3** shows that SimTDE is able to retain 99.57% of the BERT-Tiny-128-L2(128) teacher model's performance while having 31.82% of its size and only takes up $5.7Mb$ storage. SimTDE-32-L2(128) reduces the token embedding size from 128 to 32, retains 2 encoding layer structure as the teacher and uses teacher's weights for initialization. These extremely small models should enable wider runtime application of transformer sentence embedding models even on edge devices with low footprint budget and computational resources.

Inference Speed: To assess the latency benefit, we compare the inference time for a full pass over STS-B (1379 pairs) on CPU using $batch\_size = 1$. The statistics is an average from 3 runs. **Table 4**

Table 1: SimTDE performance comparison with SOTA sentence embeddings models. SBERT models' performances are cited from [6], TENC-mutual's performance is cited from [13] and SimCSE result is from self-implementation according to [6]

Model	STS12	STS13	STS14	STS15	STS16	SICKR	STSB	Average	Params
SBERT-base	70.97	76.53	73.19	79.09	74.3	72.91	77.03	74.89	109.5M
SBERT-base-flow	69.78	77.27	74.35	82.01	77.46	76.21	79.12	76.60	109.5M
SBERT-base-whitening	69.65	77.57	74.66	82.27	78.39	76.91	79.52	77.00	109.5M
TENC-mutual	75.09	**85.10**	**77.90**	85.08	**83.05**	72.76	**83.90**	80.41	109.5M
SimCSE-bert-base	**75.47**	82.40	76.78	**85.36**	80.71	**80.22**	82.69	**80.52**	109.5M
SimTDE-384-L3(768):Base	75.44	83.06	77.24	85.35	80.60	79.66	81.93	**80.47**	**34.1M**
SimTDE-128-L3(384):Small	73.44	77.82	74.30	84.18	79.35	76.49	81.04	**78.09**	**9.5M**

Table 2: SimTDE performance comparison with other distillation methods and ablation study on SimTDE:Base

Model	STS12	STS13	STS14	STS15	STS16	SICKR	STSB	Average	%SimCSE (%Tcher)	Params
Tchr:SimCSE-768-L12(768)	75.47	82.40	76.78	85.36	80.71	80.22	82.69	80.52	100.00%	109.5M
LRN-KD-768-L2(768)	74.00	81.13	75.94	84.60	79.75	78.59	80.78	79.53	98.77%	38.4M
LMX-KD-384-L12(384)	75.11	80.03	75.39	83.76	79.43	79.09	80.97	79.11	98.25%	**33.4M**
SimTDE-384-L3(768):Base	**75.44**	**83.06**	**77.24**	**85.35**	**80.60**	**79.66**	**81.93**	**80.47**	**99.94%**	34.1M
- Encoding layer reduction	75.96	82.10	77.26	85.39	80.36	79.96	82.36	80.51	99.99%	97.9M
+ Hidden states distillation	74.63	81.95	76.35	84.84	79.82	78.98	81.06	79.66	98.94%	34.1M
- Token embedding loss	75.02	82.42	77.11	84.95	80.46	79.49	81.87	80.19	99.59%	34.1M
Tchr:MiniLM-384-L6(384)	72.37	80.60	75.60	85.39	78.99	77.15	82.03	78.87	97.96% (100%)	22.7M
SimTDE-128-L3(384):Small	73.44	77.82	74.30	84.18	79.35	76.49	81.04	78.09	96.99% (99.00%)	**9.5M**

Table 3: SimTDE performance on ER multi-lingual data in relative terms w.r.t to the BertTiny teacher performance

Model	German	English	Spanish	French	Italian	Average	Params	TokEmbParams	Size
Tchr: BertTiny-128-L2(128)	-	-	-	-	-	-	4.4M	4M	17.6 MB
SimTDE-64-L2(128)	**+0.32%**	-1.26%	**+0.31%**	**+1.09%**	**-0.05%**	**+0.09%**	2.4M	2M	9.7 MB
SimTDE-32-L2(128)	-0.17%	**-0.89%**	-0.12%	-0.39%	-0.53%	-0.43%	**1.4M**	**1M**	**5.7MB**

Table 4: Inference time comparison

Model	Params	Time (seconds)
SimCSE-BERT-base-768-L12(768)	109.5M	55.3 (X1)
LMX-KD: BERT-base-384-L12(384)	33.4M	37.3 (X1.5)
MiniLM: BERT-base-384-L6(384)	22.7M	22.0 (X2.5)
SimTDE-384-L3(768):Base	34.1M	18.7 (X3)
SimTDE-128-L3(384):Small	9.5M	14.0 (X4)

illustrates that SimTDE:Base is $X3$ faster than the teacher model; $X2$ faster than LMX-KD model which adopts a deep and narrow student structure (i.e 12 encoding layers of 384D), despite of similar size; 25% faster than MiniLM which also adopts a deep and narrow student structure (i.e 6 encoding layers of 384D) and is 33% smaller. Our results indicate that the number of transformer layer is more latency costly than its dimension when running on CPU which is most common in runtime deployment. This also explains that although SimTDE:Small is $X3$ smaller in size than SimTDE:Base, having the same number of encoding layers, its inference is only $X1.3$ faster.

Ablation Study: As shown in **Table 2**, we first remove the token embedding layer number reduction and only use a compact token embedding with a projection layer. It demonstrates that a compact token embedding merely drops the performance of a large transform by 0.01% while the size reduces 11%, which is very effective. Next, we replace the encoder layer initialization with teacher's weight in SimTDE by layer-to-layer distillation (last 3 layers from teacher). However, this method results 1% worse performance. We suspect the model performs better with retrained encoding layers instead of pegged distillation to specific teacher encoder layer. It is possible the results can improve if we perform more advanced layer mapping. Nevertheless, this requires additional loss objective and computation compared to our proposal. We also investigate the effects of token embedding loss component and observe consistent performance improvement in every single STS task with L_{TE} by up to 0.77% indicating the necessity of this component. Lastly, we assess the effects of various forms of output loss component calculation, such as MSE loss and KL divergence loss on the sentence embedding, and MSE and MAE loss on the cosine similarity embeddings pairs. MSE loss on the output sentence embedding mildly outperforms the others and become our choice.

5 CONCLUSION

We propose SimTDE to compress large and compact transformer models for sentence embeddings. This easy-to-adopt framework with strong accuracy and low latency can widely enable runtime deployment of SOTA sentence embeddings, even on edge devices with limited footprint budget and computational resources. We demonstrate that SimTDE generates high-quality sentence embeddings comparable to SOTA models with a fraction of their sizes. On STS tasks, it achieves 99.95% of SOTA performance with $\frac{1}{3}$ size and 96.99% of performance with $\frac{1}{12}$ size. On ER tasks, SimTDE compresses a compact transformer model to 1.4M parameters and 5.7Mb while retaining 99.57% of teacher's performance. In inference, SimTDE:Base is $X3$ faster than teacher and $X2$ faster than other distilled transformer of similar size; SimTDE:Small is $X4$ times faster than teacher. Our current approach aims to maximally transfer knowledge from teacher. Next, it will be interesting to explore fine-tuning techniques (e.g. data augmentation) to further advance the student's performance. Another direction is to compound different compression techniques for additional size and latency reduction.

REFERENCES

[1] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. The SNLI corpus. (2015).

[2] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semantic textual similarity-multilingual and cross-lingual focused evaluation. In *Proceedings of the 2017 SEMVAL International Workshop on Semantic Evaluation (2017). https://doi. org/10.18653/v1/s17-2001.*

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, Eduardo Blanco and Wei Lu (Eds.). Association for Computational Linguistics, 169–174. https://doi.org/10.18653/v1/d18-2029

[4] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, 670–680. https://doi.org/10.18653/v1/d17-1070

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[6] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 6894–6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[7] Lise Getoor and Ashwin Machanavajjhala. 2012. Entity resolution: theory, practice & open challenges. *Proceedings of the VLDB Endowment* 5, 12 (2012), 2018–2019.

[8] Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. 2014. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115* (2014).

[9] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).

[10] Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both Weights and Connections for Efficient Neural Network. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 1135–1143. https://proceedings.neurips.cc/paper/2015/hash/ae0eb3eed39d2bcef4622b2499a05fe6-Abstract.html

[11] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* 2, 7 (2015).

[12] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020 (Findings of ACL, Vol. EMNLP 2020)*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 4163–4174. https://doi.org/10.18653/v1/2020.findings-emnlp.372

[13] Fangyu Liu, Yunlong Jiao, Jordan Massiah, Emine Yilmaz, and Serhii Havrylov. 2022. Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=AmUhwTOHgm

[14] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[15] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), 216–223. http://www.lrec-conf.org/proceedings/lrec2014/summaries/363.html

[16] Ross McGowan, Jinru Su, Vince DiCocco, Thejaswi Muniyappa, and Grant Strimel. 2021. SmallER: Scaling neural entity resolution for edge devices. In *Interspeech 2021*. https://www.amazon.science/publications/smaller-scaling-neural-entity-resolution-for-edge-devices

[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/v1/D19-1410

[18] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for Thin Deep Nets. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6550

[19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[20] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316* (2021).

[21] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient Knowledge Distillation for BERT Model Compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 4322–4331. https://doi.org/10.18653/v1/D19-1441

[22] Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual Neural Machine Translation with Knowledge Distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=S1gUsoR9YX

[23] Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136* (2019).

[24] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: The impact of student initialization on knowledge distillation. *arXiv preprint arXiv:1908.08962* 13 (2019).

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[26] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *ArXiv* abs/2002.10957 (2020).

[27] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (New Orleans, Louisiana). Association for Computational Linguistics, 1112–1122. http://aclweb.org/anthology/N18-1101

[28] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 5065–5075. https://doi.org/10.18653/v1/2021.acl-long.393