InhibiDistilbert: Knowledge Distillation for a ReLU and Addition-based Transformer Language Model

The inhibitor transformer uses an attention mechanism based on Manhattan distance and ReLU activation for computational efficiency. It achieves comparable GLUE benchmark performance to conventional dot-product transformers when trained via knowledge distillation.

Tony Zhang and Rickard Brännvall
RISE Research Institutes of Sweden

This work explores optimizing transformer-based language models by integrating model compression techniques with inhibitor attention, a novel computationally efficient alternative attention mechanism.

Inhibitor attention employs Manhattan distances and ReLU activations instead of the matrix multiplications and softmax activation of the conventional scaled dot-product attention. This shift offers potential computational and energy savings while maintaining model effectiveness. We propose further adjustments to improve the inhibitor mechanism's efficiency when trained by Knowledge Distillation (KD) and evaluate its performance on the DistilBERT architecture.

Our KD experiments indicate that the modified inhibitor transformer model achieves competitive performance on standard NLP benchmarks, including General Language Understanding Evaluation (GLUE) and sentiment analysis tasks.

Models	GLUE	IMDb
Conv. DistilBERT	77.0	92.82
Inhibi.DistilBERT	74.5	92.81

Table 1: Experiment comparing a pre-trained conventional DistilBert with the Inhibitor trained by task-agnostic KD (experiment E1).

E1: Task-agnostic KD experiment

- **Layerwise Training**: Train only Q, K and V sequentlially bottom-up with MSE loss to align self-attention, whilte other layers remain frozen.
- **Full-Layer Training**: All layers were unfrozen and trained together using MSE loss applied to hidden states to refine the representations.

E2: Task-specific KD experiment

- **Soft Probability Distillation Loss**: The distillation loss function uses the teacher model's soft probabilities to encourage the student model to replicate the teacher's predictions..
- **Hidden State Loss**: To help guide the inner layers of the student model toward better alignment with the teacher's representations.

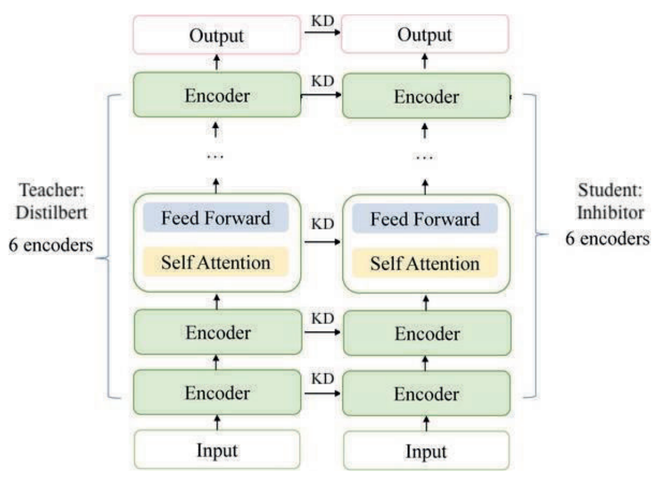

Diagram: Knowledge distillation (KD) set-up.

Method

The dot-product attention score of the conventional Transformer is replaced with the Manhattan distance:

$$S = \frac{QK^T}{\sqrt{d}} \quad \longrightarrow \quad Z_{ij} = \sum_k \frac{\gamma}{\sqrt{d}} |Q_{ik} - K_{jk}| \tag{1}$$

where Q, K, V are the query, key, and value matrices and d is the size of the latent dimension.

The attention head output is then similarly replaced

$$H = \text{softmax}(S)V \quad \longrightarrow \quad H'_{il} = \eta \sum_j \left((V^+_{jl} - \bar{Z}_{ij})^+ + (V^-_{jl} + \bar{Z}_{ij})^- \right) \tag{2}$$

written with in the notation $(x)^+ = \max(x,0)$ and $(x)^- = \min(x,0)$ for the positive and negative ReLU functions.

Results

The overall performance the two NLP tasks in Table 1 for the Task-agnostic training set-up (E1) is comparable to the conventional DistilBERT. The details in Table 2 show competitive accuracy across most GLUE tasks with the exception for the CoLA task. Task-specific KD (E2) lags significantly behind.

A performance drop may be expected for both experiments as we used the original Distilbert model both as a teacher model and as a benchmark baseline. Furthermore, we note that the training data for the GLUE tasks is very sparse

Future work

- Quantization aware training. Special hardware.
- KD from full sized BERT.
- Generative LM and Vision Transformers
- Test on modern and harder NLP tasks.

Models	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
Conv. DistilBERT (E0)	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3
Inhibi.DistilBERT (E1)	40.0	79.2	86.8	85.4	89.5	59.2	90.2	83.5	56.3
Inhibi.DistilBERT (E2)	47.5	72.2	77.0	80.0	63.4	47.3	91.0	83.5	56.3

Table 2: Comparison on GLUE between conventional DistilBERT (E0) and inhibitor trained by Task-agnostic KD (E1) and Task-specific KD (E2). Task scores are averaged over three runs.

ABOUT RISE

RISE Research Institutes of Sweden is an independent, State-owned research institute with over 3000 employees, which offers unique expertise and over 130 testbeds and demonstration environments for future technologies, products and services.

RISE Research Institutes of Sweden
www.ri.se / info@ri.se

CONTACT
Rickard Brännvall
rickard.brannvall@ri.se