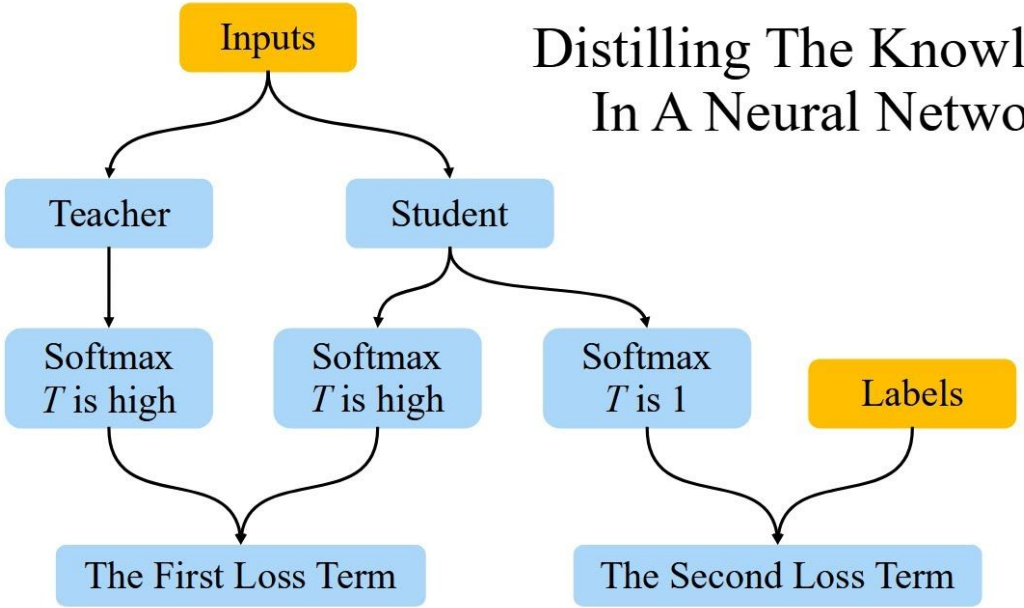
INF721 Final Project - Emulating MiniLM:Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers

Olavo Barros
Departamento de Informatica
Universidade Federal de Viçosa
Viçosa, Brazil
olavo.barros@ufv.br

Summary:

- Knowledge distillation
- BERT
- Implementation
  - Self-Attention Distribution Transfer
  - Self-Attention Value-Relation Transfer
- Experimental Setup
- Results and Discussion

Knowledge distillation



https://www.youtube.com/watch?app=desktop&v=qPwGfAfBmFI

BERT

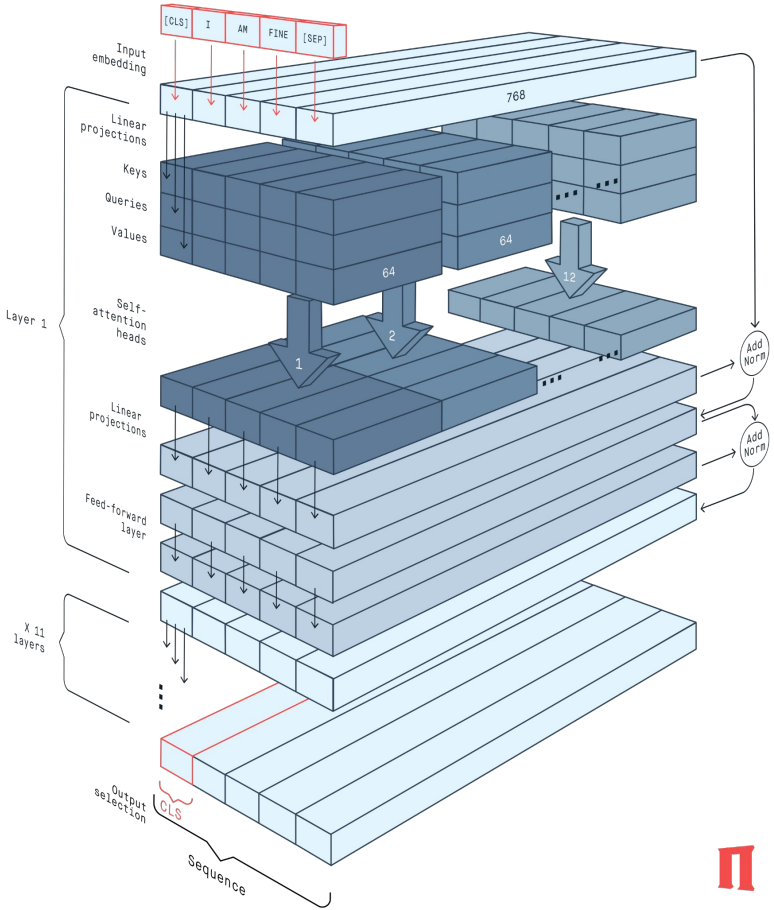
1. Input Embeddings (Token, Segment, and Positional Embeddings)

2. Multi-Head Self-Attention Mechanism

3. Feedforward Neural Networks and Layer Normalization

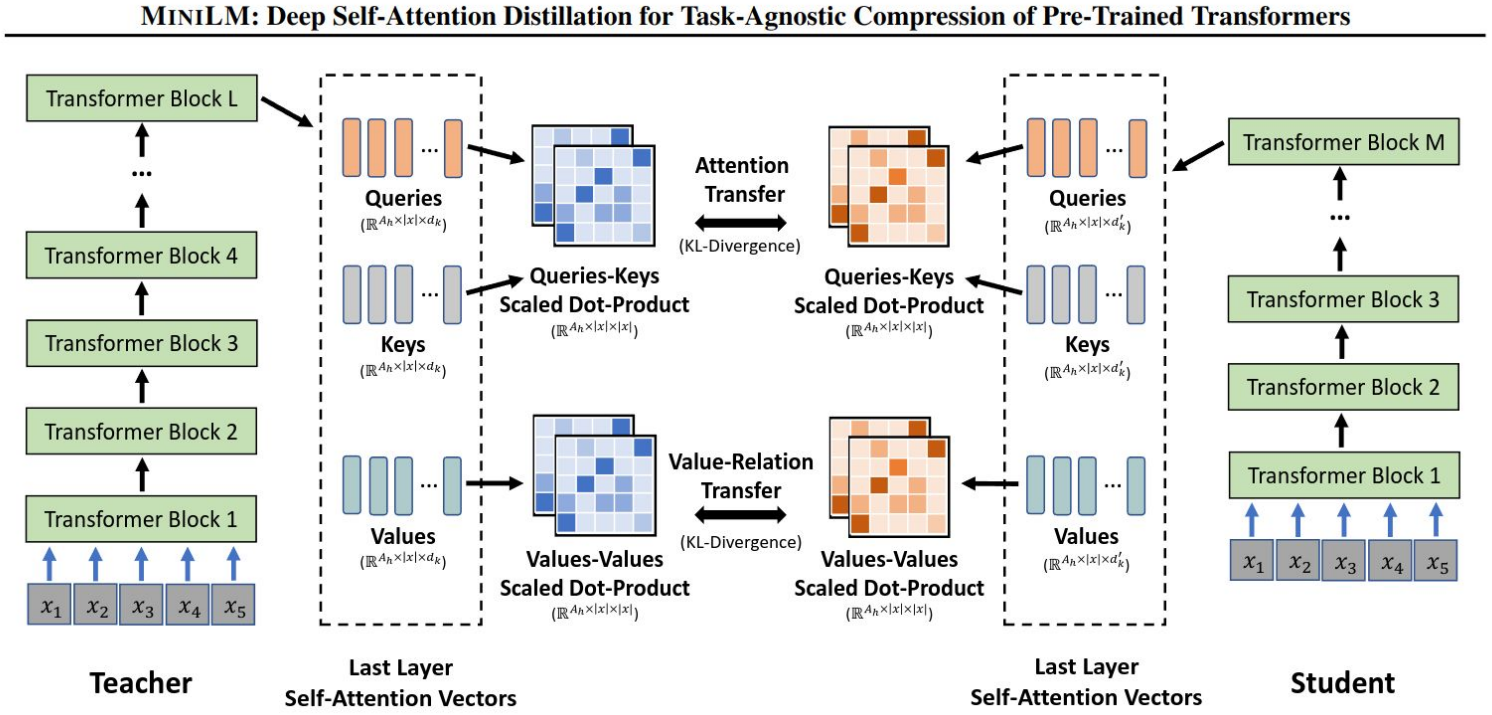
4. Transformer Encoder Layers (Stacked Architecture)

5. Output Representations for Tokens and Sequences



https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/blocks/bert-encoder

Implementation



Wang, Wenhui, et al. "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers." Advances in Neural Information Processing Systems 33 (2020): 5776-5788.

Implementation

$$\mathcal{L} = (\mathcal{L}_{AT} + \mathcal{L}_{VR}) + \mathcal{L}_{OUT}$$

The training loss is computed by summing the attention distribution transfer loss and value-relation transfer loss

Introducing the relation between values enables the student to deeply mimic the teacher's self-attention behavior.

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right)$$

Self-Attention Distribution Transfer

$$\mathcal{L}_{AT} = \frac{1}{A_h|x|} \sum_{a=1}^{A_h} \sum_{t=1}^{|x|} D_{KL} \left(\mathbb{A}^T_{L,a,t} \parallel \mathbb{A}^S_{M,M,a,t} \right)$$

1. $|\mathbf{x}|$: Represents the **sequence length**, i.e., the number of tokens in the input sequence.
2. $\mathbf{Ah}$: Represents the **number of attention heads** in the attention mechanism of the model.
3. $\mathbf{L \text{ and } M}$: Represent the **number of layers** in the teacher and student models, respectively.
4. $\mathbf{A}$: Represents the **attention distribution** of the **last Transformer layer** in the **teacher model and student model**.

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right)$$

Self-Attention Value-Relation Transfer

$$\mathcal{L}_{VR} = \frac{1}{A_h |x|} \sum_{a=1}^{A_h} \sum_{t=1}^{|x|} D_{KL} \left(VR^T_{L,a,t} \parallel VR^S_{M,a,t} \right)$$

The value relation is computed via **the multi-head scaled dot-product between values**. The KL-divergence between the value relation of the teacher and student is used as the training objective

$$VR^T_{L,a} = \text{softmax} \left(\frac{V^T_{L,a} (V^T_{L,a})^T}{\sqrt{d_k}} \right)$$

$$VR^S_{M,a} = \text{softmax} \left(\frac{V^S_{M,a} (V^S_{M,a})^T}{\sqrt{d'_k}} \right)$$

Experimental Setup

Model:

Description	Teacher	Student
Teacher Model Architecture	BERTBASE (Devlin et al., 2018)	My implementation
Model Type	12-layer Transformer	6-layer Transformer
Hidden Size	768	768
Attention Heads	12	12
Total Parameters	109M	65M

Experimental Setup

Hyperparameters:

Description	MiniLM	My Implementation
Vocabulary Size	30,522	30,522
Maximum Sequence Length	512	128
Adam Hyperparameters	β1=0.9, β2=0.999	β1=0.9, β2=0.999
Batch Size (Student)	1024 (for 6-layer student model)	32
Learning Rate (Student)	5e-4 (peak learning rate)	5e-4 (peak learning rate)
Training Steps (Student)	400,000	210,500
Linear Warmup Steps	4,000	2,000
Dropout Rate	0.1	0.1

UFV
Universidade Federal de Viçosa

Experimental Setup

Dataset SST-2:

1. **Dataset Overview**: SST-2 is a sentiment analysis dataset for binary classification (positive or negative sentiment).

2. **Text Data**: Contains movie reviews labeled as positive (1) or negative (0).

3. **Number of Samples**: 67,349 samples—53,142 for training, 1,872 for validation, 1,821 for testing.

4. **Labels**: Two sentiment labels: "positive" (1) and "negative" (0).

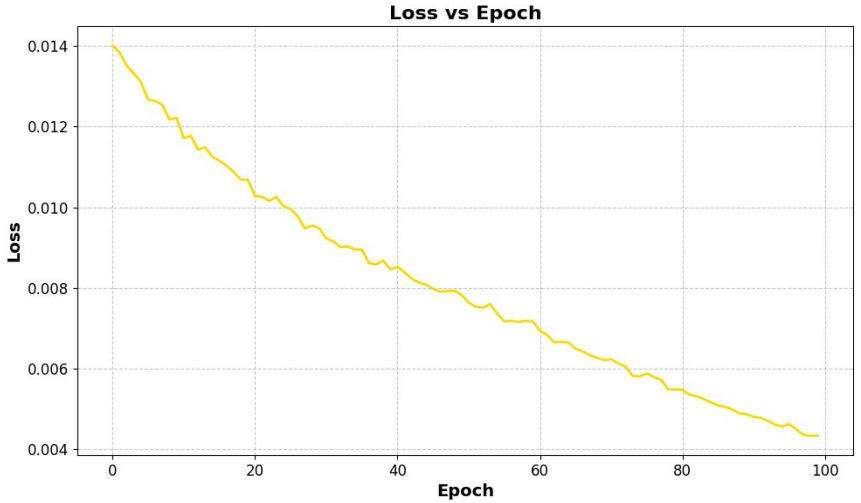
5. **Input Length**: Sentences with varying lengths, typically capped at 128 or 512 tokens.

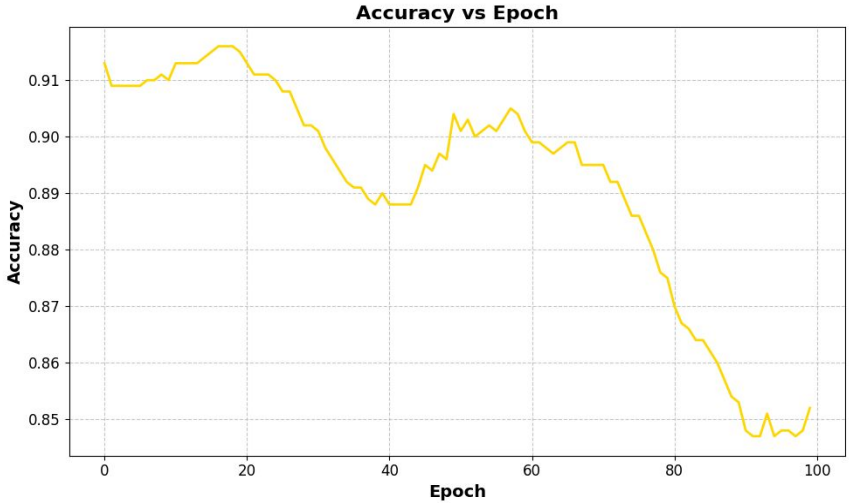
BERT-Base checkpoint (PyTorch, AMP, SST-2, seqLen128)

Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." Proceedings of the 2013 conference on empirical methods in natural language processing. 2013.

Results and Discussion

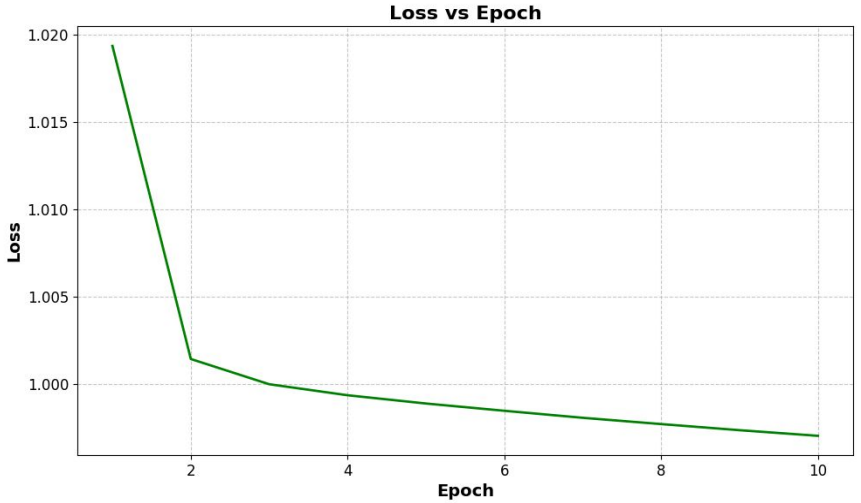
- Uses the first 1000 samples of the dataset;
- Uses the transforms' implementation for the student;
- Uses the original loss function proposed in the paper.

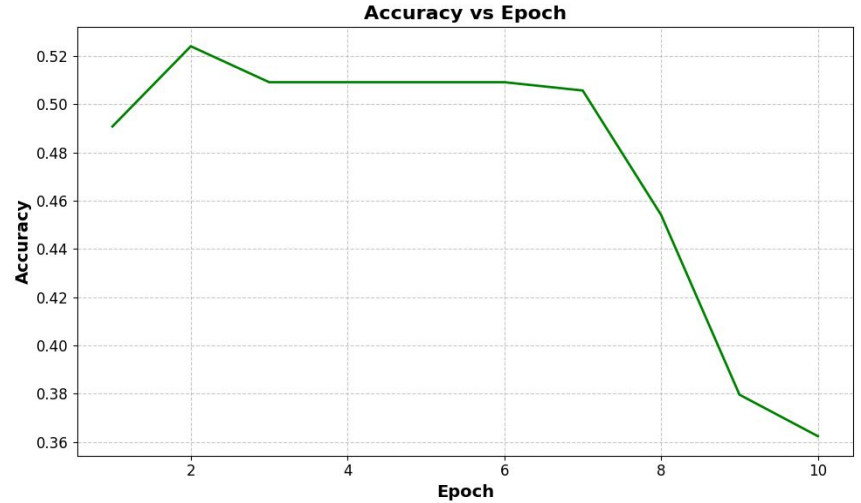




Results and Discussion

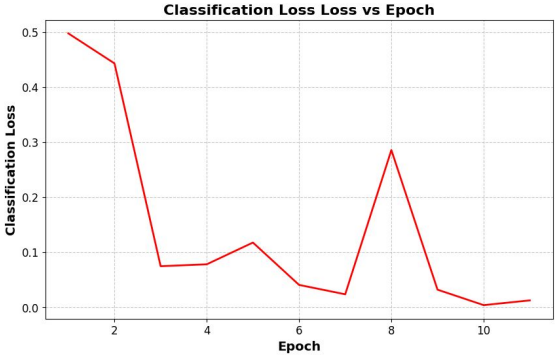
- Uses the full dataset;
- Uses my implementation for the student;
- Uses the original loss function proposed in the paper.

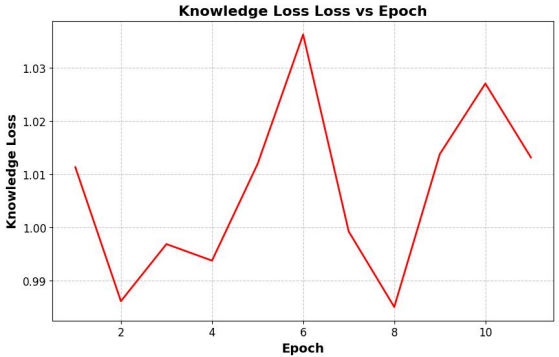


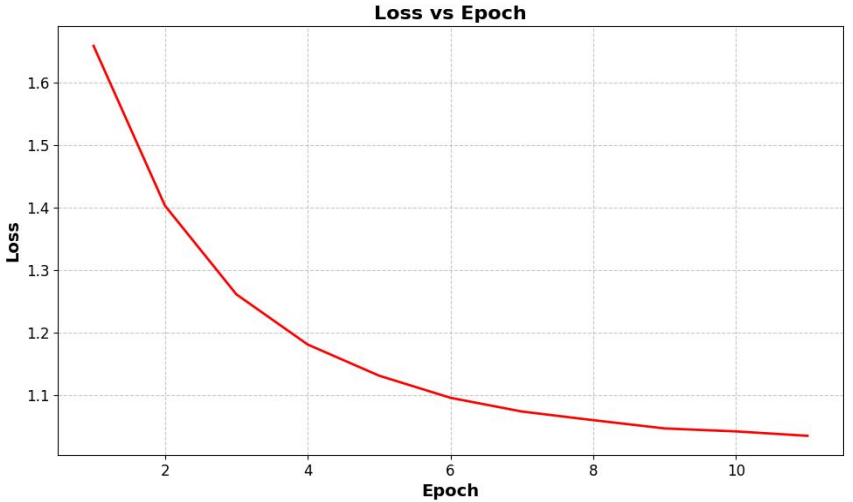


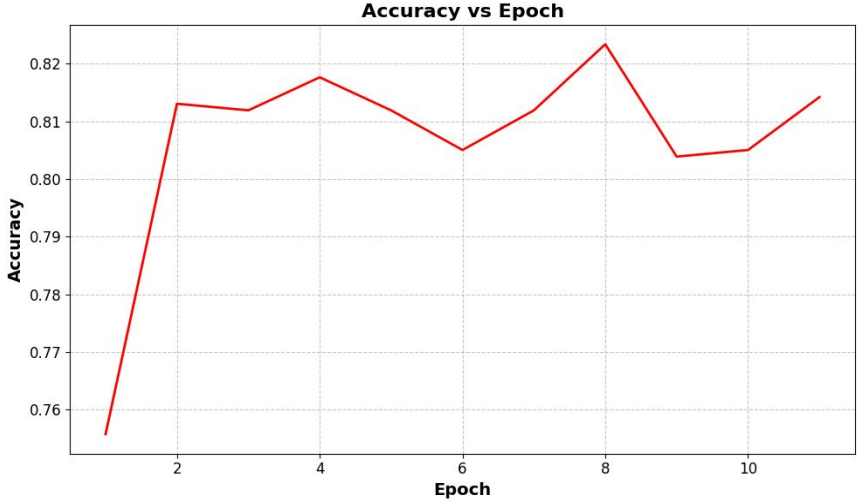
Results and Discussion

$$\mathcal{L} = (\mathcal{L}_{AT} + \mathcal{L}_{VR}) + \mathcal{L}_{OUT}$$









Conclusion

Model	#Param	SQuAD2	MNLI-m	SST-2	QNLI	CoLA	RTE	MRPC	QQP	Average
$\text{BERT}_{\text{BASE}}$	109M	76.8	84.5	93.2	91.7	58.9	68.6	87.3	91.3	81.5
DistillBERT	66M	70.7	79.0	90.7	85.3	43.6	59.9	87.5	84.9	75.2
TinyBERT	66M	73.1	83.5	91.6	90.5	42.8	**72.2**	**88.4**	90.6	79.1
MINILM	66M	**76.4**	**84.0**	**92.0**	**91.0**	**49.2**	71.5	**88.4**	**91.0**	**80.4**

Wang, Wenhui, et al. "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers." Advances in Neural Information Processing Systems 33 (2020): 5776-5788.

https://github.com/Olavo-B/Emulating-MiniLM.git

olavo.barros@ufv.br