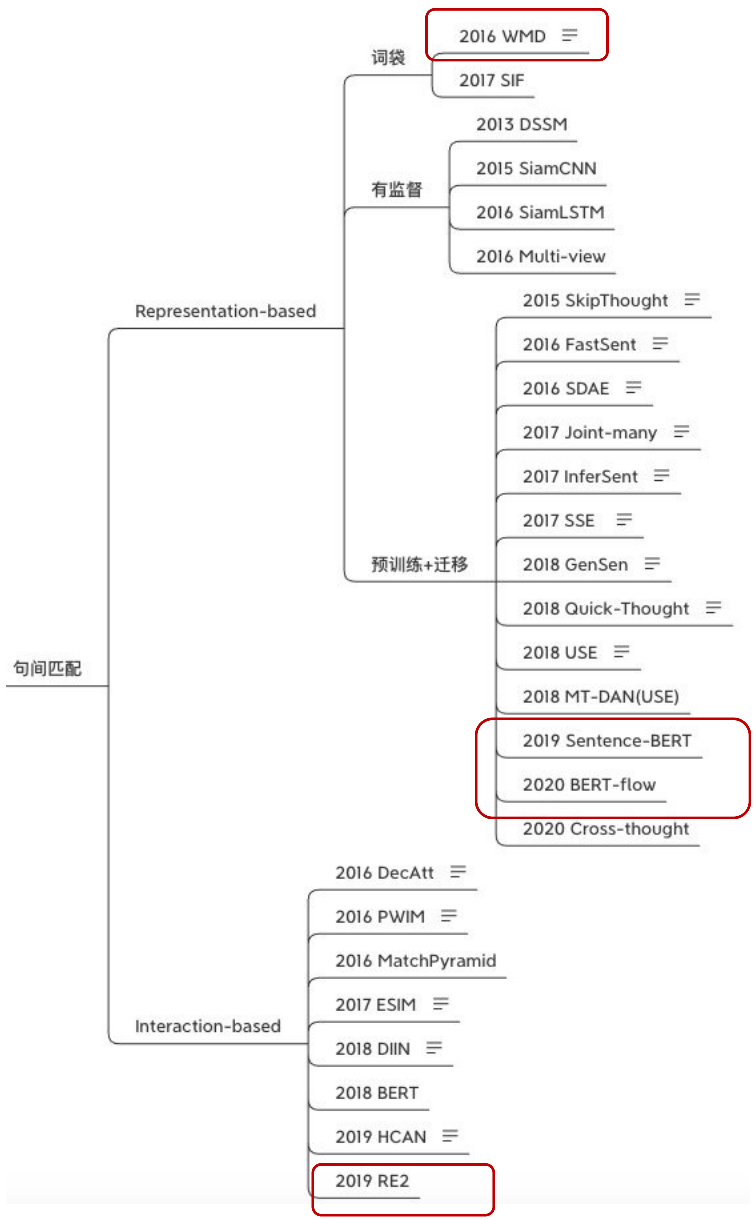
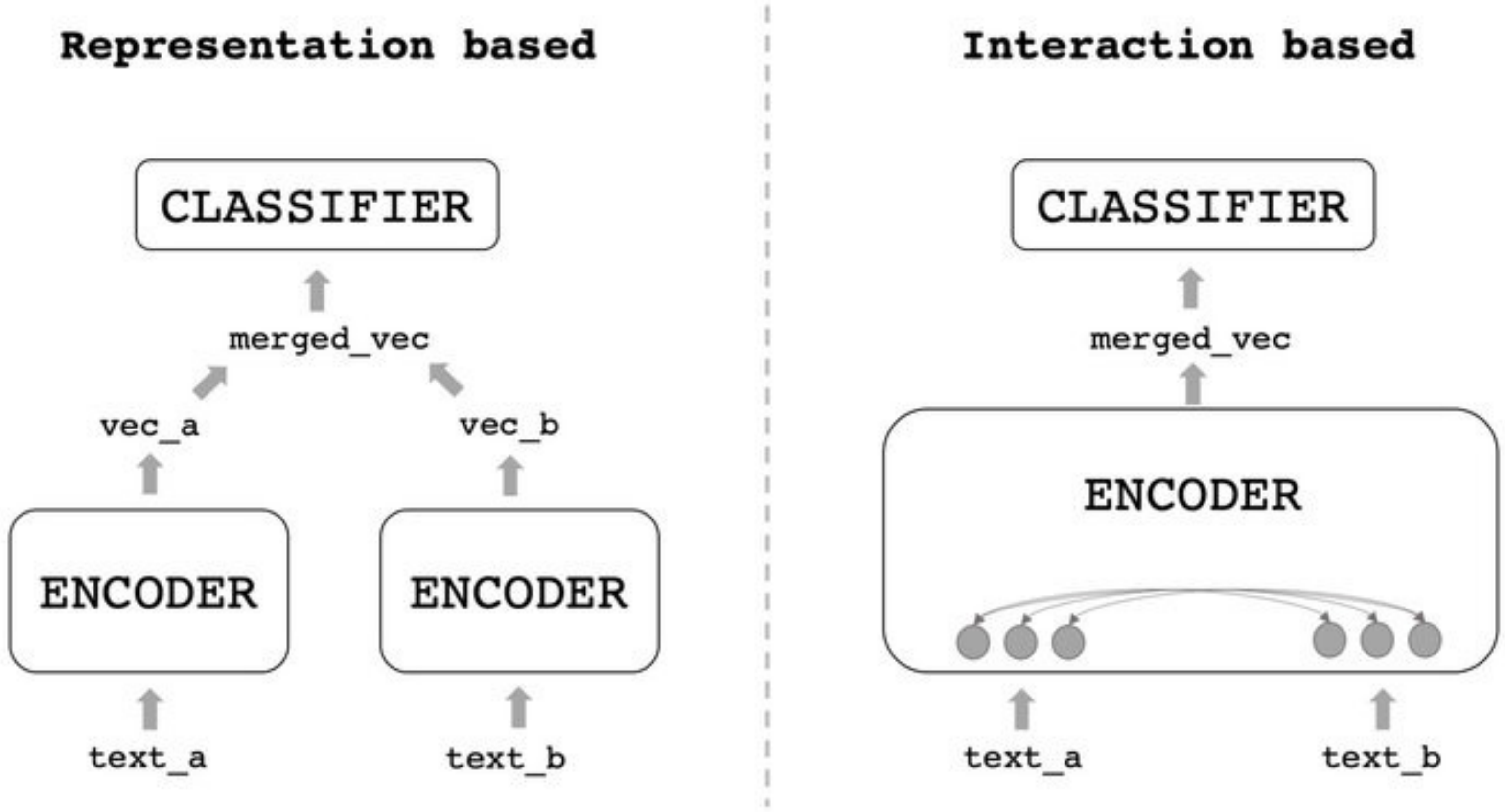
Dual-View Distilled BERT for Sentence Embedding

——SIGIR 2021

Xingyi Cheng
Ant Group
fanyin.cxy@alibaba-inc.com

——ZJD 20210701


句间匹配
Representation-based
词袋
2016 WMD
2017 SIF
有监督
2013 DSSM
2015 SiamCNN
2016 SiamLSTM
2016 Multi-view
预训练+迁移
2015 SkipThought
2016 FastSent
2016 SDAE
2017 Joint-many
2017 InferSent
2017 SSE
2018 GenSen
2018 Quick-Thought
2018 USE
2018 MT-DAN(USE)
2019 Sentence-BERT
2020 BERT-flow
2020 Cross-thought
Interaction-based
2016 DecAtt
2016 PWIM
2016 MatchPyramid
2017 ESIM
2018 DIIN
2018 BERT
2019 HCAN
2019 RE2


Representation based
CLASSIFIER
merged_vec
vec_a
vec_b
ENCODER
ENCODER
text_a
text_b
Interaction based
CLASSIFIER
merged_vec
ENCODER
text_a
text_b

Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

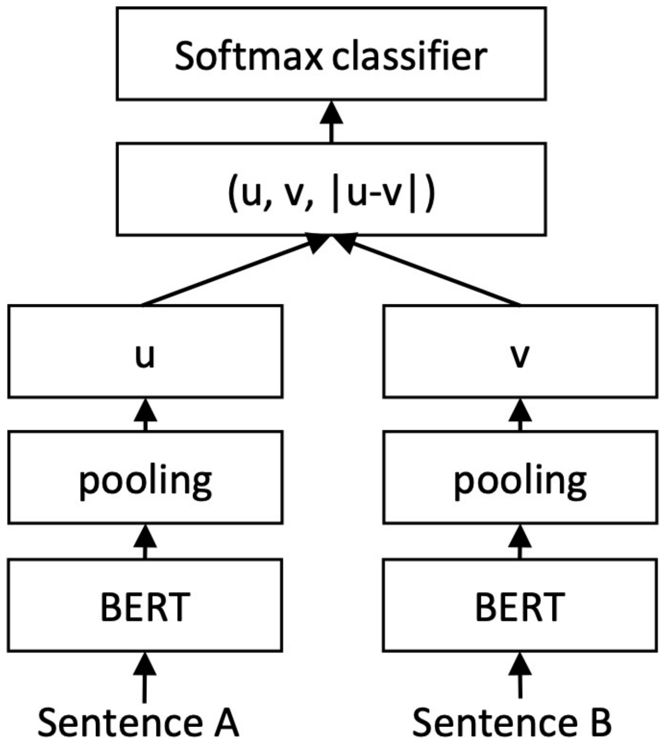


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

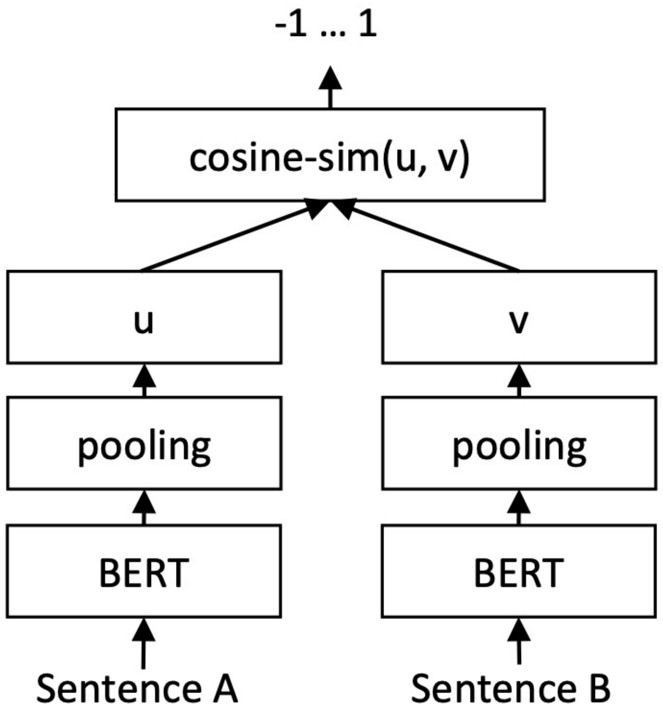


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

Dual-View Distilled BERT for Sentence Embedding

——SIGIR 2021

Xingyi Cheng
Ant Group
fanyin.cxy@alibaba-inc.com

Train the sentence matching model from two views:

(1) Siamese View, they start with the siamese BERT-networks as a backbone to derive sentence embeddings, to be able to capture semantics similarity efficiently by calculating distances on the two fixed-size vectors.

(2) Interaction View, the standard pre-trained models with cross-sentence interactions are utilized, acting as multiple teachers that generate predictions about the training set provided to the siamese networks to learn.

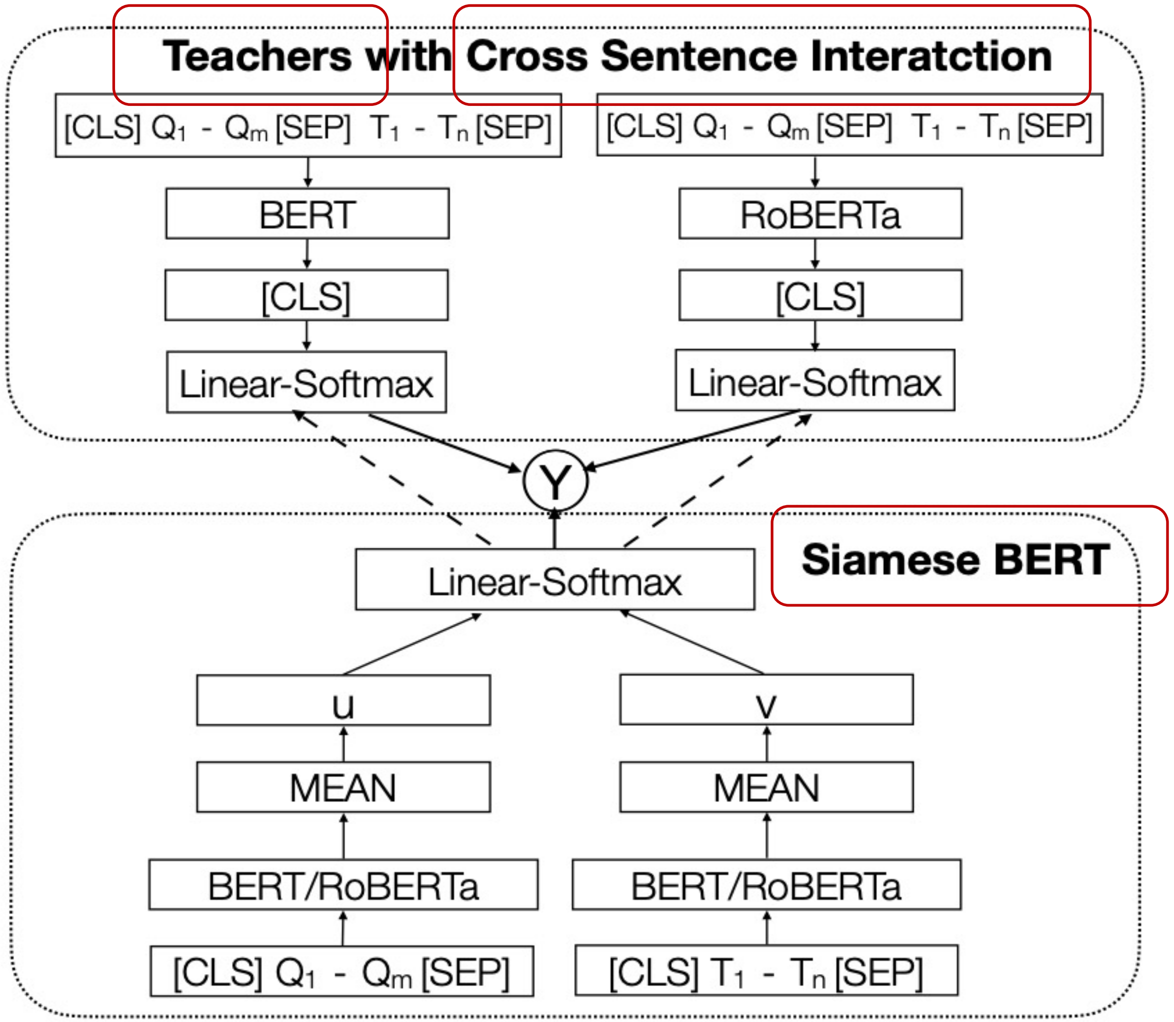


Figure 1: Overview of Dual View Distilled BERT. Dash lines indicate distillation.

Siamese BERT-Networks

dataset $D_l \longrightarrow y \in Y$

$$\mathbf{Y} = \{entailment, contradiction, neutral\}$$

对于分类任务，如NLI，将u、v和|u-v|连接起来，然后是一个全连接层将其映射到一个固定的概率分布中，如下：

hidden size into a probability distribution.

$$p(y|\mathbf{u},\mathbf{v};\theta) = \mathtt{softmax}(W[\mathbf{u},\mathbf{v},|\mathbf{u}-\mathbf{v}|]),$$

where $\boxed{\theta}$ represents all learnable parameters from BERT, shared for $\mathbf{u},\mathbf{v}$. And $\boxed{W \in \mathbb{R}^{3d\times n}}$ is the parameter of the fully-connected layer. $\boxed{d}$ is the dimension of the sentence embeddings. We optimize the standard cross-entropy loss.

For any sentence-pairs, the siamese BERT converts the two sentences into sequential vectors individually, and then pool these two vectors into two sentence embeddings $\mathbf{u}$ and $\mathbf{v}$. SBERT (Reimers and Gurevych, 2019)

Cross Sentence Interaction

使用多种不同的预训练语言模型作为教师模型，指导Siamese BERT-Networks学习，通过引入句子之间的词交互信息，以丰富词级别的交互特征。

Each model first pre-trains with labeled data, then re-labeling the data and adds it to a new training set.

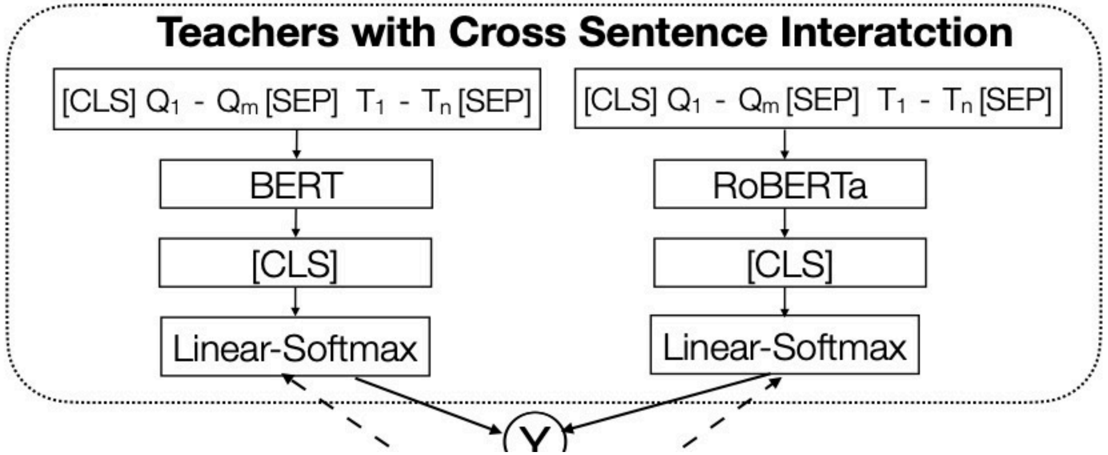


Fig 1 (top), we concatenate the sentence-pair $Q = \{Q_i\}_{i=1,\ldots,m}$ and $T = \{T_i\}_{i=1,\ldots,N}$ into a text sequence $[\texttt{[CLS]}Q\texttt{[SEP]}T\texttt{[SEP]}]$.

第k个预训练模型的[CLS]标记

$$q(y|\mathbf{z}_k^c;\phi_k) = softmax(O\mathbf{z}_k^c)),$$

$$\mathcal{L}(\theta, W) = \sum_{k=1}^{K} D(q(y|\boxed{\mathbf{z}_k^c;}\phi_k), p(y|\mathbf{u}, \mathbf{v}; \theta)),$$

模型的参数

Teacher Annealing

We leverage teacher annealing (Clark et al., 2019) strategy, which mixes the teacher prediction with the gold label during training.

孪生BERT-Networks模型和其他K个BERT模型的目标函数，如下：

$$\mathcal{L}(\theta) = \sum_{\tau \in \mathcal{T}} \sum_{x_\tau^i, y_\tau^i \in \mathcal{D}_\tau} \ell(f_\tau(x_\tau^i, \theta_\tau), f_\tau(x_\tau^i, \theta))$$

其中，λ从0到1线性增加。一开始，λ=0时，意味着模型完全基于教师的软目标进行训练。随着模型的逐渐收敛，模型更多地学习硬目标。

BAM! Born-Again Multi-Task Networks for Natural Language Understanding

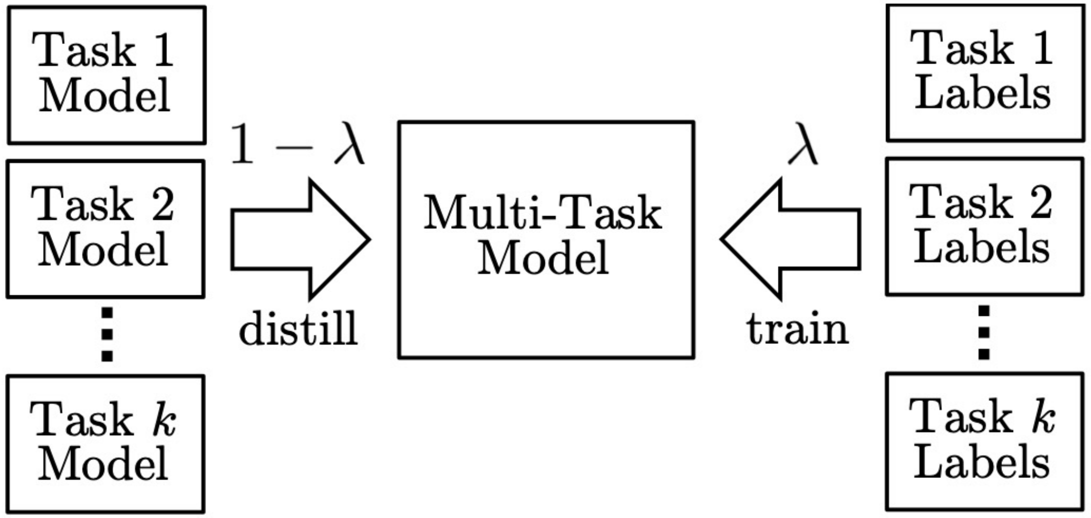


Figure 1: An overview of our method. λ is increased linearly from 0 to 1 over the course of training.

3.1 Dataset

The NLI dataset consists of SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018), annotated with the labels contradiction, entailment, and neutral. STS (Agirre et al., 2012) assesses the matching degree to which two sentences are semantically equivalent to each other, which are human-annotated with a level of equivalence from 1 to 5. We follow the previous works (Conneau et al., 2017; Cer et al., 2018) to merge the training and test datasets in both NLI data as pre-training datasets of 940k sentence pairs. STS 2012-2016 datasets have no training data but 26k test data, so the datasets are used to evaluate the pre-trained DvBERT on NLI. STS-B is a collection of 8.6k sentence pairs and contains training, development, and test sets drawn from heterogeneous sources.

	STS12	STS13	STS14	STS15	STS16	STS-B	Avg.
BERT Avg. embedding	38.78	57.98	57.98	63.15	61.06	46.35	54.22
BERT [CLS] embedding	20.16	30.01	20.09	36.88	38.08	16.50	26.95
SBERT-base	70.4	71.77	70.66	78.67	74.11	76.28	73.64
SRoBERTa-base	71.70	73.43	71.47	**80.79**	75.99	77.02	75.06
DvBERT-base	70.52	73.17	71.18	79.88	75.08	77.96	74.63
DvRoBERTa-base	**72.42**	**73.44**	**72.21**	80.43	**76.52**	**78.32**	**75.56**
SBERT-large	71.68	72.79	72.20	80.32	76.45	78.00	75.24
SRoBERTa-large	72.14	76.69	74.12	79.81	75.97	78.60	76.22
DvBERT-large	72.95	72.26	71.87	79.27	76.16	78.28	75.13
DvRoBERTa-large	**74.99**	**76.16**	**73.34**	**81.93**	**78.77**	**79.61**	**77.47**

Table 1: Spearman correlation of STS tasks without fine-tuning on task-specific data.

	Base models	Large models
BERT-NLI	87.33 ± 0.23	89.09 ± 0.36
RoBERTa-NLI	**89.77** ± 0.47	**91.12** ± 0.17
SBERT	84.57 ± 0.2	84.72 ± 1.01
SRoBERTa	84.89 ± 0.34	86.13 ± 0.94
DvBERT	84.67 ± 0.23	85.31 ± 0.21
DvRoBERTa	**85.31** ± 0.37	**86.23** ± 0.67
SBERT-NLI	85.01 ± 0.17	85.91 ± 0.58
SRoBERTa-NLI	85.40 ± 0.2	86.15 ± 0.35
DvBERT-NLI	85.15 ± 0.24	86.21 ± 0.13
DvRoBERTa-NLI	**86.05** ± 0.22	**86.98** ± 0.46

Table 2: Spearman correlation of STS tasks. The average of 10 runs with different random seeds is reported. "-NLI" indicates the model is pre-trained on NLI data.

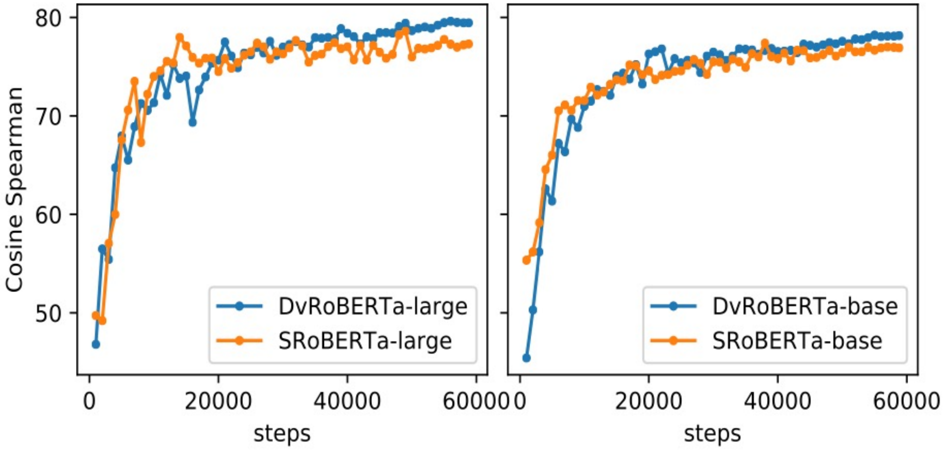


Figure 2: Spearman correlation for SRoBERTa and DvRoBERTa.

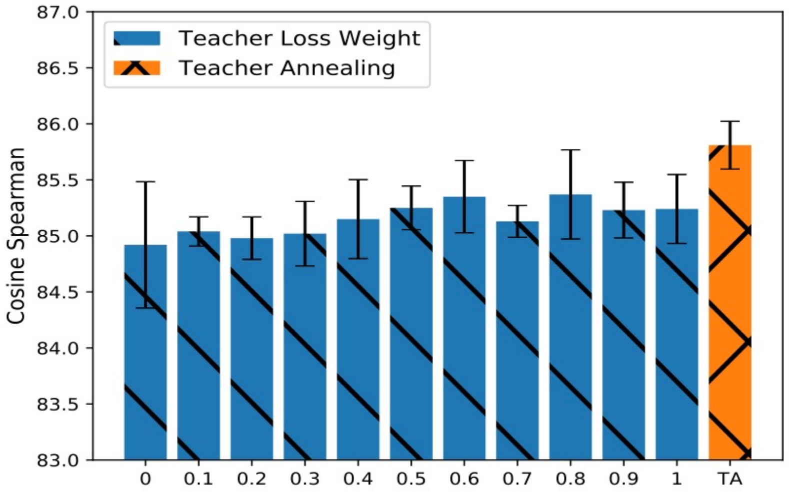


Figure 3: Comparison of teacher loss Weighting and teacher annealing