An Improved SSL Model for Adversarial NLI

Mehmet Serhan Çiftlikçi
Department of Computer Engineering
Bogazici University
Istanbul, Turkey, 34342
mehmet.ciftlikci@boun.edu.tr

Göktuğ Öcal
Department of Computer Engineering
Bogazici University
Istanbul, Turkey, 34342
goktug.ocal@boun.edu.tr

Abstract

We present a robust Natural Language Inference (NLI) model pipeline by fine-tuning the current state-of-the-art Transformer model DeBERTaV3 (He et al., 2021) in two stages. In the first stage, our model is fine-tuned with a combination of high-quality paraphrase datasets and an appropriate Metric Learning (Bellet et al., 2015) loss function. In the second stage, our model is fine-tuned with NLI datasets and Cross Entropy Loss. Our approach brings the best of both worlds together to utilize DeBERTaV3's potential for the downstream task fully, which is the evaluation our model's performance on ANLI (Nie et al., 2019), an adversarial NLI benchmark dataset with challenging examples to encourage building robust models. With this approach, our model performs competitively to the current state-of-art performance for the ANLI dataset while being simpler and more customizable. The source code can be found on GitHub[1].

1 Introduction

Since large amounts of data being generated daily, deep learning becomes necessary to have powerful algorithms while processing such data. Although deep learning has enjoyed most success by tackling supervised learning problems where labeled data is required, as the amount of data increases rapidly, data labeling becomes practically infeasible. A method called "Self-Supervised Learning" (SSL) (Jing and Tian, 2020) has emerged to alleviate this issue, where models aim to learn inherent information representations directly getting supervision from the large amounts of data. Supervision means that the derived signals from the data itself, frequently by exploiting the underlying structure of data. Any unseen or concealed component of the input can be predicted using SSL. As such, Transformer-based or, broadly, attention-based models are proposed for building high-quality representations to improve the performance of downstream tasks. Soon after, they became the state-of-art models for a wide range of NLP and vision tasks and were used as the first choice of practitioners for building deep models (Braşoveanu and Andonie, 2020).

Conversely, the rise of Transformer (Devlin et al., 2018) models has also raised doubts about whether these models truly internalize the learning behavior and create meaningful representations that can be used while generalizing to unseen data or not (Niven and Kao, 2019). An increasing amount of study has been conducted to understand how these models respond to different types of adversarial perturbations or to train more robust models by open-sourcing challenging benchmark datasets.

Under this context, in this study, our research is mainly focused on NLI, which is considered as a fundamental subtask in NLP, deciding whether a hypothesis can be deducted or not for a given premise. To solve this task, we mainly focused on ANLI, a crowdsourced NLI dataset created

[1] https://github.com/sciftlikci/improved-adversarial-nli-model

with an iterative human-and-model-in-the-loop procedure to collect primarily adversarial examples. As presented in the paper, despite well-known BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) are performing well on other NLI datasets, they quickly fail on ANLI, showing that there is still a long way to reach the human level of understanding. The current state-of-the-art model for ANLI is InfoBERT (B. Wang et al., 2020), where an information-theoretic perspective was applied to RoBERTa to suppress noise while building token representations, coupled with the alignment of alignment local patterns with global sentence representations to make them more useful. Although the idea behind the paper is effective, we believe it can be easily surpassed by simply fine-tuning DeBERTaV3 (He et al., 2021), a DeBERTa (A. Wang et al., 2019) model where major improvements have been made to increase the overall embedding quality. To clarify the reason for our model selection, DeBERTa model family was the first one to pass human-level performance in the SuperGLUE benchmark. Hence, our intuition is that fine-tuning a DeBERTaV3 model with the selected NLI datasets under an appropriate Metric Learning loss function will display better performance for the downstream task. To this extent, our contribution is adapting a successful language model to an adversarial NLI task to reach the current state-of-the-art performance for the ANLI dataset, while presenting a simpler and easily customizable approach.

2 Related Work

Sentence embedding techniques refer to representing inherent syntactic and semantic information of an arbitrary text on sentence level via dense vectors. Since sentences are invaluable parts of humans' everyday natural language usage, efficiently encoding them will allow us to achieve higher performances for downstream tasks in general. Hence, it is a well-researched area with numerous studies trying to yield the "best" sentence embeddings. Early work is mostly inspired by the methods to learn embeddings on word level. Le and Mikolov, 2014 have introduced Doc2Vec, where paragraph vectors are merged with word vectors gathered from a paragraph and used to infer the following word in a given window. Skip-Thought (Kiros et al., 2015), which is an encoder-decoder model, tried to model surrounding sentences by looking at the sentence in the center of a context window. In more recent researches, supervised learning is also leveraged such as InferSent (Conneau et al., 2017) and they performed better than the unsupervised approaches. Based upon this results, Universal Sentence Encoder (Vaswani et al., 2017) aimed to connect supervised and unsupervised worlds by training a Transformer model relying on self-attention mechanism, which is proven to be particularly useful for various NLP tasks. Current state-of-art sentence embedding methods is mostly based on fine-tuning a novel pre-trained Transformer model on sentence-based datasets. One successful study with this approach is called SBERT (Reimers and Gurevych, 2019) and it outperformed other state-of-art methods in both computational speed and accuracy on downstream NLI tasks. This work has inspired many studies that have displayed great performances by following this recipe while obtaining sentence embeddings.

Parallel to the rise of innovative sentence embedding methods based on SSL, the domain of Metric Learning is also transformed with novel approaches to fine tune these pretrained models better for downstream tasks. Contrastive loss (CL) (Chen et al., 2020) is a commonly-used Metric Learning method seeking to group similar samples together and different samples apart by utilizing a similarity metric to compare the distance between them in the feature space. (Liao, 2021) fine tuned a model by interpolating the BERT fine-tuning process between Cross Entropy Loss and supervised CL and that approach outperformed SBERT on STS-B (Reimers and Gurevych, 2019) task. Triplet loss (TL) (Dong and Shen, 2018) is a Metric Learning loss function that takes a reference sample and simultaneously reduces the distance between a similar input and increases the distance between a different sample from it. Finally, Multiple Negative Ranking Loss (MNRL) (Henderson et al., 2017) is proposed to produce high-quality sentence embedding specifically for paraphrase detection task. This loss function can be considered as a slight modification to the TL function where only pairs of similar texts are collected as a dataset but during fine-tuning, for each pair, all other samples in the batch are considered as negative samples, hence the relative distances between the inputs are increased while the distance between current pair is reduced. Ha et al., 2021 stated that MNRL and CL as two of the commonly used loss functions for Transformer-based model fine-tuning studies.

Table 1: Sample data from ANLI

premise	hypothesis	label	reason
Roberto Javier Mora Garcia (c. 1962 – 16 March 2004) was a Mexican journalist and editorial director of "El Manana", a newspaper based in Nuevo Laredo, Tamaulipas, Mexico. He worked for a number of media outlets in Mexico, including the "El Norte" and "El Diario de Monterrey", prior to his assassination.	Another individual laid waste to Roberto Javier Mora Garcia.	neutral	The context states that Roberto Javier Mora Garcia was assassinated, so another person had to have "laid waste to him." The system most likely had a hard time figuring this out due to it not recognizing the phrase "laid waste."
A melee weapon is any weapon used in direct hand-to-hand combat; by contrast with ranged weapons which act at a distance. The term "melee" originates in the 1640s from the French word "melee", which refers to hand-to-hand combat, a close quarters battle, a brawl, a confused fight, etc. Melee weapons can be broadly divided into three categories	Melee weapons are good for ranged and hand-to-hand combat.	entailment	Melee weapons are good for hand to hand combat, but NOT ranged.
If you can dream it, you can achieve it—unless you're a goose trying to play a very human game of rugby. In the video above, one bold bird took a chance when it ran onto a rugby field mid-play. Things got dicey when it got into a tussle with another player, but it shook it off and kept right on running. After the play ended, the players escorted the feisty goose off the pitch. It was a risky move, but the crowd chanting its name was well worth it.	The crowd believed they knew the name of the goose running on the field.	contradiction	Because the crowd was chanting its name, the crowd must have believed they knew the goose's name. The word "believe" may have made the system think this was an ambiguous statement.

3 Method

3.1 Task Description

Natural Language Inference is considered as an important problem for studies on Natural Language Understanding (NLU) and it is seen to be a suitable task for assessing overall understanding ability of NLP models (Williams et al., 2020). ANLI is one of the most recent and challenging benchmark dataset on the NLI task and the dataset illustrated that current state-of-the-art models, such as GPT-3, disastrously fail if challenging examples are introduced for this task.

The data collection procedure of ANLI is based on studies where training collaborative machine learning agents in couple of rounds, which define difficulty levels, via a gamification technique (Nie et al., 2019). In each round, human agents generated examples trying to fool the current level's model from Wikipedia texts with explanations. Then, train/dev/test splits are created and a newer model is trained for the next round with current round's accumulated data. Some examples from ANLI is presented in Table 1.

3.2 Our Approach

Our work aims to associate SSL and Metric Learning thoughtfully to produce high-quality sentence embeddings and show state-of-art performance on ANLI dataset, while presenting a simple and easily-customizable pipeline. To accomplish this goal, we claimed that a two-stage fine tuning procedure with a the DeBERTaV3 model is more than enough.

In the first stage, a DeBERTaV3 model is fine-tuned to produce semantically useful word embeddings with positive or negative sentence pairs with an appropriate Metric Learning loss function (Figure 1). In this stage, following Metric Learning loss functions have been used; CL, Online CL (OCL) (Reimers and Gurevych, 2019) and MNRL. Paraphrase datasets collected for this stage and reformatted based on the applied loss function. Then pooling layers are attached to the model for fine-tuning purposes. After fine tuning, pooling layers are removed and the model is extracted from the established network to start the next fine-tuning process.

In the second stage, the input sentence pair is given to the fine-tuned DeBERTaV3 models separately and pooling layers follows these models. Sentence embeddings gathered from the pooling layer outputs are concatenated and fed through a classification layer. The established network is fine-tuned with Cross Entropy Loss and the trained network with DeBERTaV3 transformers is used to evaluate ANLI dataset for NLI task (Figure 2)

In the presented method, basically, aim is to fine-tune pretrained DeBERTaV3 model with Metric Learning Loss and Cross Entropy Loss and generating sentence representations for ANLI test set to

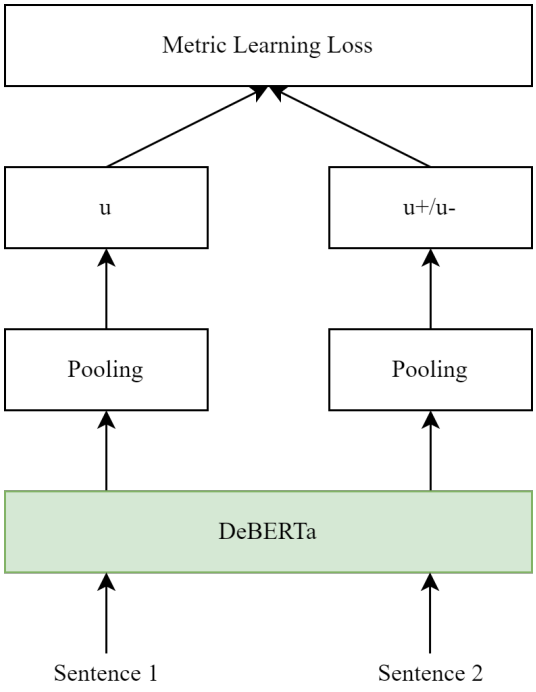

Figure 1: Stage 1 fine-tuning with Metric Learning structure

increase accuracy performance for the NLI task. The used loss functions are invesitagated in Section 3.3.

3.3 Loss Functions

3.3.1 Contrastive Loss

Assume a given a set of input vectors $I = \left\{\vec{X_1}, ..., \vec{X_P}\right\}$, where $\vec{X_i} \in \Re^D$, $\forall i = 1, ..., n$ and a parametric function $G_W : \Re^D \rightarrow \Re^d$ with $d \ll D$. Let $\vec{X_1}, \vec{X_2} \in I$ be a pair of input vectors given to the model and Y be a representation of the model where $Y = 0$ if $\vec{X_1}$ and $\vec{X_2}$ are admitted as entailment, and $Y = 1$ if they are admitted as contradictory. Euclidean distance between outputs of G_W is the distance function to be learned D_W between $\vec{X_1}$ and $\vec{X_2}$. The exact loss function is,

$$L(W, Y, \vec{X_1}, \vec{X_2}) = (1 - Y)\frac{1}{2}(D_W)^2 + (Y)\frac{1}{2}\left\{max(0, m - D_W)\right\}^2 \tag{1}$$

where $m > 0$ is a margin. The margin defines a radius around $G_W(\vec{X})$.

3.3.2 Online Contrastive Loss

OCL is a modified version of Contrastive Loss, where it computed for hard positive (positives that are far apart) and hard negative pairs (negatives that are close).

3.3.3 Multiple Negatives Ranking Loss

Let there is a set is used to approximate $P(y|x)$, which has K possible responses, one positive response and $K - 1$ random negatives selected from other samples in a training batch. There are K premises $x = (x_1, ..., x_K)$ and their corresponding hypothesis $y = (y_1, ..., y_K)$ in a batch of size K. Every hypothesis y_j is admitted as a negative for x_i if $i \neq j$. The $K - 1$ negative examples for each x are different at each backpropagation step due to shuffling. The objective function that is being tried to minimize for each batch is,

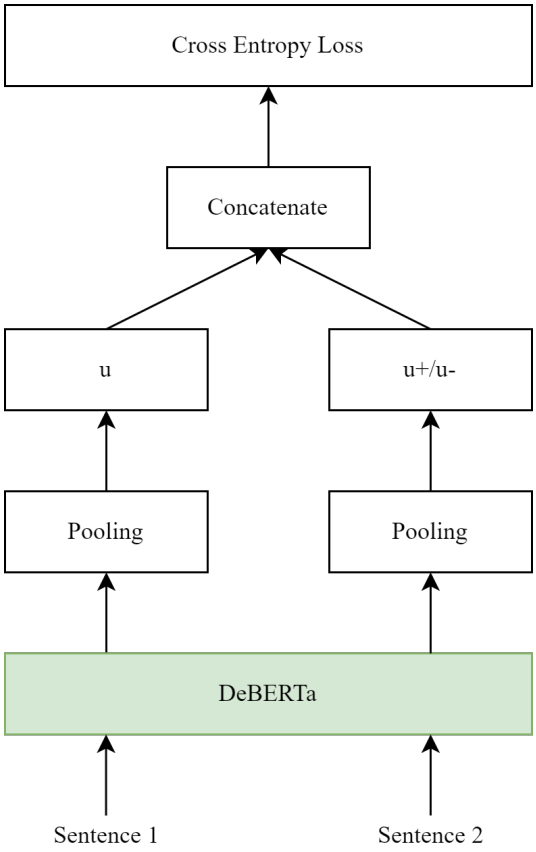

Figure 2: Stage 2 fine-tuning of DeBERTaV3 models

$$J(x, y, \theta) = -\frac{1}{K} \sum_{i=1}^{K} \left[S(x_i, y_i) - log \sum_{j=1}^{K} e^{S(x_i, y_j)} \right] \tag{2}$$

3.3.4 Cross Entropy Loss

Cross Entropy Loss function is the classification objective function. The sentence embeddings u and $u + /u-$, provided by DeBERTaV3 model, concatenate and multiplied with weights W_t.

$$\sigma = -\sum_{c=1}^{M} y_c log(\hat{y}_c) \tag{3}$$

where $y_c = W_t(u, u\pm)$ for class number c. If $M > 2$, it means multi-class classification and loss calculation is done for each class label separately and each result are summed. In the stage 2, the network optimizes the Cross Entropy Loss. Structure is shown in Figure 2.

4 Experiments

Data

For the experiments, we planned to gather and try various different datasets for both of the stages of fine-tuning. The datasets are selected by considering similar researches on NLI domain in the literature and introduced by stages where they are used.

While fine-tuning with Metric Learning Loss in the stage 1, paraphrase datasets under **GLUE** benchmark are aggregated with other paraphrase datasets which are considered to be useful for the process (Table 2). **PAWS** (Zhang et al., 2019) is a dataset that contains paraphrase and non-paraphrase pairs whose subsets are based ond Wikipedia and Quora. **SICK** (Marelli et al., 2014) is a collection of around 10,000 English sentence pairings that showcase a variety of lexical, syntactic, and semantic issues. **RTE** (A. Wang et al., 2018) dataset is a combination of different datasets such as RTE1, RTE2, RTE3, and RTE5. **WNLI** (Levesque et al., 2012) is a reading comprehension task, and sentence pairs of the dataset were created by substituting the ambiguous pronoun with each alternative referent to turn the problem into sentence pair classification. **MRPC** (Dolan and Brockett, 2005) is a corpus of sentence pairs gathered automatically from internet news sources and annotated by humans to determine the semantic equivalency of the sentences in the pair. **STS-B** (Cer et al., 2017) is a set of sentence pairs derived from news headlines, video and image captions, and data on natural language inference. **ART** (Bhagavatula et al., 2019) is a dataset which is created for Abductive Natural Language Inference task. **QUORA** (Quora, n.d.) dataset is made up of pairs of questions, and the goal is to determine if the two questions are semantically identical. Finally, **SWAG** (Zellers et al., 2018) is a dataset which is prepared in order to predict which event in a video is most likely to happen next, where ach question is a video caption and with four answer choices.

Table 2: Summary Statistics for Stage 1 Datasets

DATASET	LABEL	SAMPLE SIZE	TOTAL SAMPLE SIZE	LABEL RATIO	AVERAGE TEXT LENGTH
ART		73578	73578	1	12.888
MRPC	0	1901	5801	0.328	21.892
	1	3900		0.672	
PAWS	0	36497	65401	0.558	21.353
	1	28904		0.442	
QUORA		149263	149263	1	9.854
RTE	0	1395	2767	0.504	26.099
	1	1372		0.496	
SICK	0	2821	9840	0.287	9.643
	1	5595		0.569	
	2	1424		0.145	
STSB		7249	7249	1	10.242
SWAG		73546	73546	1	11.175
WNLI	0	363	706	0.514	13.944
	1	343		0.486	

In the stage 2, fine-tuning of the new pipeline is conducted with the aggreagating of ANLI's train split, **MNLI** (Williams et al., 2017), **ConjNLI** (Saha et al., 2020) and **EQUATE** (Ravichander et al., 2019). **MNLI**, which is a crowd-sourced NLI dataset consisting of sentence pairs coming from a variety of sources, including transcribed conversation, fiction, and official reports. **ConjNLI**, a stress-test for NLI over conjunctive sentences in which the premise differs from the hypothesis due to conjuncts being corrupted. **EQUATE**, which is made up of five NLI test sets with amounts, three of which use language from real-world sources such as news articles and social media for quantitative reasoning.

Table 3: Summary Statistics for Stage 2 Datasets

DATASET	PART	LABEL	SAMPLE SIZE	TOTAL SAMPLE SIZE	LABEL RATIO	AVERAGE TEXT LENGTH
ANLI	Train	0	52111	162865	0.32	31.863
		1	68789		0.422	
		2	41965		0.258	
	Validation	0	1070	3200	0.334	32.380
		1	1068		0.334	
		2	1062		0.332	
	Test	0	1070	3200	0.334	32.318
		1	1068		0.334	
		2	1062		0.332	
MNLI		0	137841	412349	0.334	14.893
		1	137152		0.333	
		2	137356		0.333	
ConjNLI		0	536	1623	0.330	17.598
		1	748		0.461	
		2	339		0.209	
EQUATE		0	1068	2106	0.507	14.888
		1	658		0.312	
		2	380		0.180	

Experiments Setup

For stage 1, **sentence-transformers** [2] library provided by the researchers of SBERT (Reimers and Gurevych, 2019) and for stage 2, **HuggingFace Trainer API** [3] were utilized. For essential text preprocessing, **Texthero** [4] and **contractions** [5] were used. Experiments were run on Google Colab with a single Tesla P100-PCIE-16GB GPU. Experiment tracking was conducted via the Weights and Biases platform. During fine tuning, AdamW optimizer with learning rate 2e-5 and a linear optimizer scheduler for 10% of the train data as warm-up step were deployed. Finally, models are fine-tuned with batch size of 32 and 64 depending on the memory usage of fine-tuning processes. Executions have took 4 to 16 GPU hours.

Fine-Tuning and Evaluation

Summary results for all experiments are presented in the Table 4. To start with, Experiment A is conducted by considering only Stage 2 fine-tuning to both understand the difference of contributions between stages and to establish a baseline, which resulted with a dev accuracy of 39.28%.

In experiments B and C, the proposed method is fully applied. For Experiment B, initially, MNRL is utilized. However, since MNRL is a memory-intensive loss function, to run an experiment, the batch size is reduced down to 4. Considering the strong positive correlation between the batch size and the performance of this function, it is decided that it would be pointless to apply it with this batch size.

[2]https://github.com/UKPLab/sentence-transformers

[3]https://github.com/huggingface/transformers

[4]https://github.com/jbesomi/texthero

[5]https://github.com/kootenpv/contractions

Table 4: Summary of Experiments

		Loss		Datasets			ANLI Accuracy	
	Framework	**Stage 1**	**Stage 2**	**Stage 1**	**Stage 2**	**No. Epochs**	**Dev**	**Test**
A	PyTorch	-	Cross Entropy Loss	-	ANLI train + MNLI	1	39.28%	38.11%
B	PyTorch	MNR	Cross Entropy Loss	All para-phrase datasets	ANLI train + MNLI	1	OOM*	
C	PyTorch	CL	Cross Entropy Loss	All para-phrase datasets	ANLI train + MNLI	1	38.69%	37.55%
D	HuggingFace Trainer	-	Cross Entropy Loss	-	ANLI train + MNLI	3	55.21%	55.09%
E	HuggingFace Trainer	CL	Cross Entropy Loss	All para-phrase datasets	ANLI train + MNLI	3	55.59%	56.18%
F	HuggingFace Trainer	-	Cross Entropy Loss	-	ANLI train + MNLI + ConjNLI + EQUATE	5	58.59%	52.37%
G	HuggingFace Trainer	OCL	Cross Entropy Loss	All para-phrase datasets	ANLI train + MNLI + ConjNLI + EQUATE	5	57.15%	**57.37%**
State-of-the-art				**InfoBERT** (RoBERTa)			**58.3%**	**58.3%**

* Out of memory

Hence, it is skipped until a feasible approach is established. In Experiment C, the proposed method is used successfully with CL in stage 1, and the reached dev accuracy was 38.69%.

After getting initial results, suspicions were raised about the correctness of the implemented custom PyTorch model since the expectations were an increase in the performance with the proposed method. However, there was a slight drop in the performance. At that point, instead of debugging the implementation, we switched to HuggingFace Trainer API due to time constraints. Then, with the new implementation, number of epochs is increased to 3 and experiments were repeated. The results of these experiments were much more in line with the performance of DeBERTaV3 presented in the literature, where ANLI Dev accuracy increased significantly up to 55.59%. That experiment showed us the current framework worked successfully and we achieved a comparable performance to state-of-the-art model for ANLI.

For further experiments, the latest used framework is kept since the results were reliable. In experiment F, two new datasets were added to the stage 2 training which were ConjNLI and EQUATE. Experiment F was conducted by skipping stage 1 and carried out with only stage 2 training via Cross Entropy Loss for 5 epochs. The prediction accuracy for ANLI Dev climbed to 58.59% while the test accuracy was 52.37%, the best performance we got so far. In the latest experiment, extended dataset for stage 2 were kept. This time stage 1 was performed with OCL for training. Same training settings were replied and training was done in 5 epochs. The accuracy were slightly decreased to 57.15% but the test accuracy reached to 57.27%.

This final result showed us the training done in two stages, with OCL in stage 1 and extended datasets in stage 2, perform better when the test accuracy get into account. The two staged training is more than enough for DeBERTaV3 model to achieve a performance level close to the state-of-art.

5 Conclusion

SSL is eagerly motivating researchers in recent years, due to its ability to achieve state-of-art performance with downstream tasks after the application of the well-known fine-tuning recipes. SSL is alleviating the problem related with the labeling large amounts of data by leveraging input data itself as supervision, which is found out to be suitable for almost all types of downstream tasks. Those approach is also beneficial in tasks of NLP such as NLI and currently, Transformer-based models are leading the applications of SSL on NLP.

We showed that by applying a simple yet efficient fine-tuning recipe, we achieved comparable results to state-of-the-art performance on ANLI. Our approach was based on two stage fine-tuning process by carefully selecting datasets and loss functions for each stage. In the first stage, with positive and negative sentence pairs and a Metric Learning loss function, a DeBERTaV3 model was fine-tuned. Then, the model is extracted and transferred to the second stage where the fine-tuning is conducted with Cross Entropy Loss for given NLI input pairs. During the experiments, different frameworks have been used for the experiments.

During the study, a turning point is occurred due to the framework usage. Initial experiments were conducted by a custom PyTorch framework, but ANLI Dev accuracy was below than expected. Another difference from expectations happened when MNRL is selected for the first stage. The memory-intensive nature of MNRL forced us to consider small batch sizes, where achieving high performance became practically impossible.

After switching to HuggingFace Trainer, Dev accuracy values attained from new experiments were much more in line with the results presented in the literature. To increase robustness in our approach, two datasets have been added and the results got even better. The accuracy for ANLI Dev in single stage (only stage 2) fine tuning approach with cross entropy loss function and extended datasets, is 58.59%. However the accuracy of ANLI Test in same experiment was quite different, which is 52.37%. We suspect that an overfitting is occured due to training with an insufficient number of epochs. An interesting result and the best test accuracy was emerged while adding stage 1 with OCL to previous single stage experiment. In that experiment, ANLI Dev accuracy slightly drop down to 57.15%. But, the ANLI Test accuracy climbed up to 57.37%. That result showed us using multi-stage fine tuning process enabled DeBERTaV3 model to reach state-of-art performances in Adverserial NLI dataset.

To sum up, our work has provided a simple and easily-customizable pipeline for NLI task while experiments have achieved comparable performance to the state-of-art performance. All the accuracy performances shown in Table 4.

References

Bellet, A., Habrard, A., & Sebban, M. (2015). Metric learning. *Synthesis lectures on artificial intelligence and machine learning*, *9*(1), 1–151.

Bhagavatula, C., Bras, R. L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S. W.-t., & Choi, Y. (2019). Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*.

Braşoveanu, A. M., & Andonie, R. (2020). Visualizing transformers for nlp: A brief survey. *2020 24th International Conference Information Visualisation (IV)*, 270–279.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *International conference on machine learning*, 1597–1607.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dolan, B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. *Third International Workshop on Paraphrasing (IWP2005)*.

Dong, X., & Shen, J. (2018). Triplet loss in siamese network for object tracking. *Proceedings of the European conference on computer vision (ECCV)*, 459–474.

Ha, T.-T., Nguyen, V.-N., Nguyen, K.-H., Nguyen, K.-A., & Than, Q.-K. (2021). Utilizing sbert for finding similar questions in community question answering. *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*, 1–6.

He, P., Gao, J., & Chen, W. (2021). Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.-H., Lukács, L., Guo, R., Kumar, S., Miklos, B., & Kurzweil, R. (2017). Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *43*(11), 4037–4058.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. *Advances in neural information processing systems*, *28*.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *International conference on machine learning*, 1188–1196.

Levesque, H., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Liao, D. (2021). Sentence embeddings using supervised contrastive learning. *arXiv preprint arXiv:2106.04791*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., & Zamparelli, R. (2014). A sick cure for the evaluation of compositional distributional semantic models. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 216–223.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. (2019). Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Niven, T., & Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*.

Ravichander, A., Naik, A., Rose, C., & Hovy, E. (2019). Equate: A benchmark evaluation framework for quantitative reasoning in natural language inference. *arXiv preprint arXiv:1901.03735*.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Saha, S., Nie, Y., & Bansal, M. (2020). Conjnli: Natural language inference over conjunctive sentences. *arXiv preprint arXiv:2010.10418*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, *32*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wang, B., Wang, S., Cheng, Y., Gan, Z., Jia, R., Li, B., & Liu, J. (2020). Infobert: Improving robustness of language models from an information theoretic perspective. *arXiv preprint arXiv:2010.02329*.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Williams, A., Thrush, T., & Kiela, D. (2020). Anlizing the adversarial natural language inference dataset. *arXiv preprint arXiv:2010.12729*.

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Zhang, Y., Baldridge, J., & He, L. (2019). Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.