NATURAL LANGUAGE INFERENCE USING DEBERTA AND BILSTM

STEVEN MOUSSA
EUGENE LIAN

GROUP 70

The University of Manchester

NATURAL LANGUAGE INFERENCE (NLI) LEVERAGING DEBERTA AND BILSTM

ABSTRACT

We present two deep learning (DL) solutions to address the Natural Language Inference Task (NLI), with only one underpinned by transformer architectures. We leveraged and enhanced the Bidirectional Long Short-Term Memory Networks (**BiLSTM**) and **DeBERTaV3** models.

INTRODUCTION

NLI determines the relationship between a **premise** and **hypothesis**. which is useful in several real-life applications such as Banking, Retail, Search Engines, etc.

We tested different models for the transformer solution and for the embeddings in the BiLSTM. We found that DeBERTa works best as an **embedding** for BiLSTM (**87.41%** acc) and a **transformer model** (**92.34%** acc).

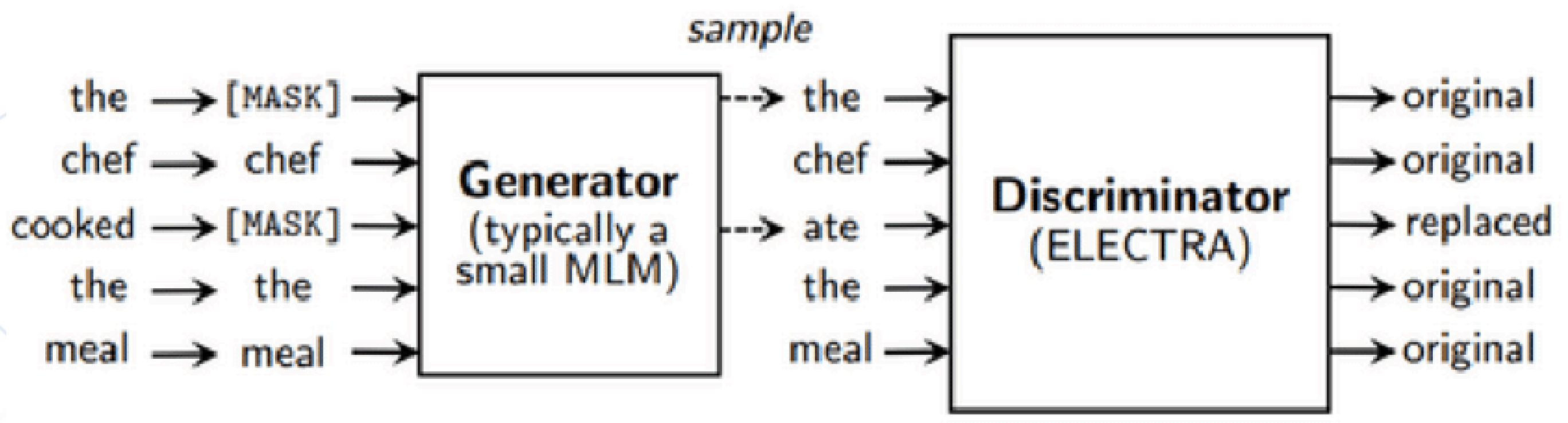

Figure 1: ELECTRA Pre-training (Clark, K. et al., 2020)

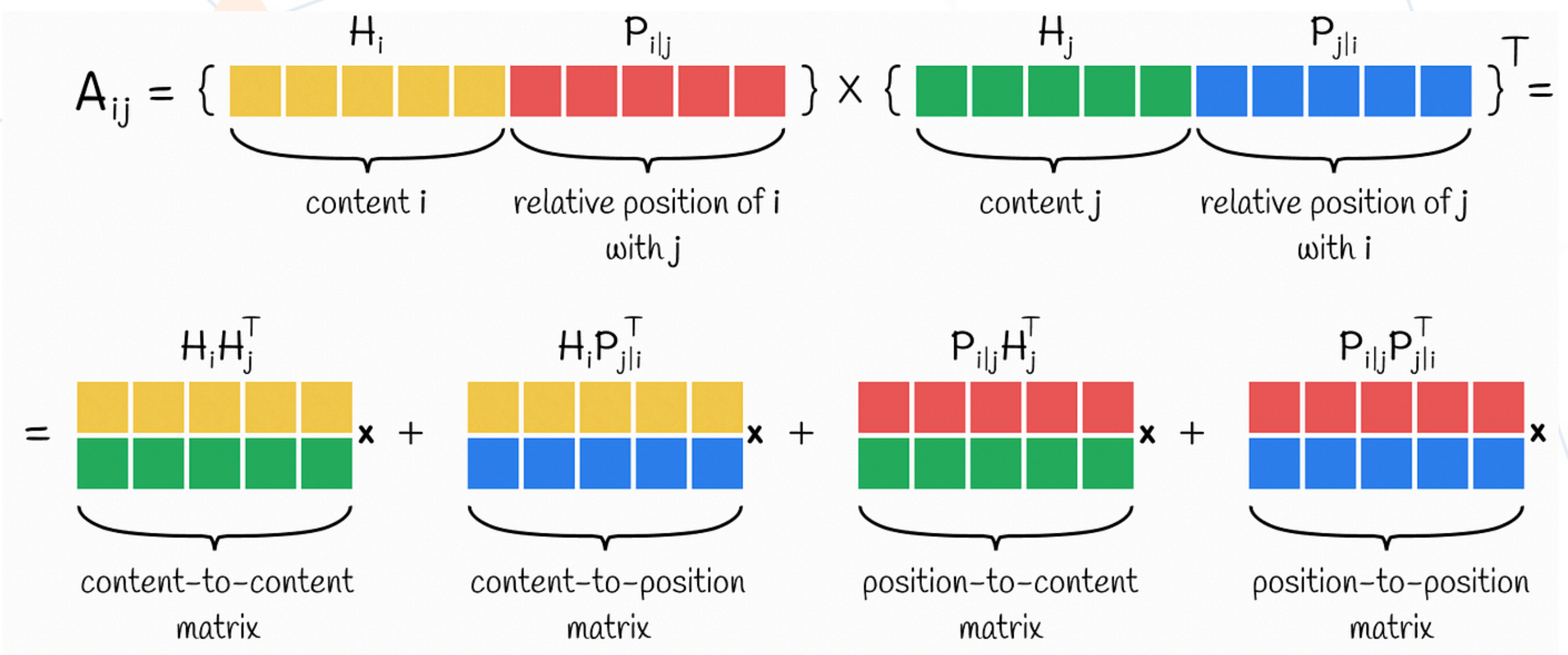
A_{ij} =
H_i
$P_{i|j}$
content i
relative position of i with j
H_j
$P_{j|i}$
content j
relative position of j with i
$H_iH_j^T$
$H_iP_{j|i}^T$
$P_{i|j}H_j^T$
$P_{i|j}P_{j|i}^T$
content-to-content matrix
content-to-position matrix
position-to-content matrix
position-to-position matrix

Figure 2: Attention (Chowdhury, T., 2023)

METHODS AND ARCHITECTURE

DeBERTaV3 with ELECTRA-Style pre-training

DeBERTa improves upon BERT and RoBERta by using:

- Disentangled attention
- An enhanced mask encoder
- Scale invariant fine-tuning (SIFT): Improves robustness by applying peturbations to normalised word embeds.

It achieved state of the art results in 2021, being applied to a range of tasks such as NLI and Question Answering.
We perform fine-tuning on the NLI dataset:

- Global Average Pooling (after last hidden state)
- Dropout layer
- Final Classifier with sigmoid activation

BiLSTM with DeBERTa embeddings

Our BiLSTM model had the following components: DeBERTaV3-base as embeddings and BiLSTM.

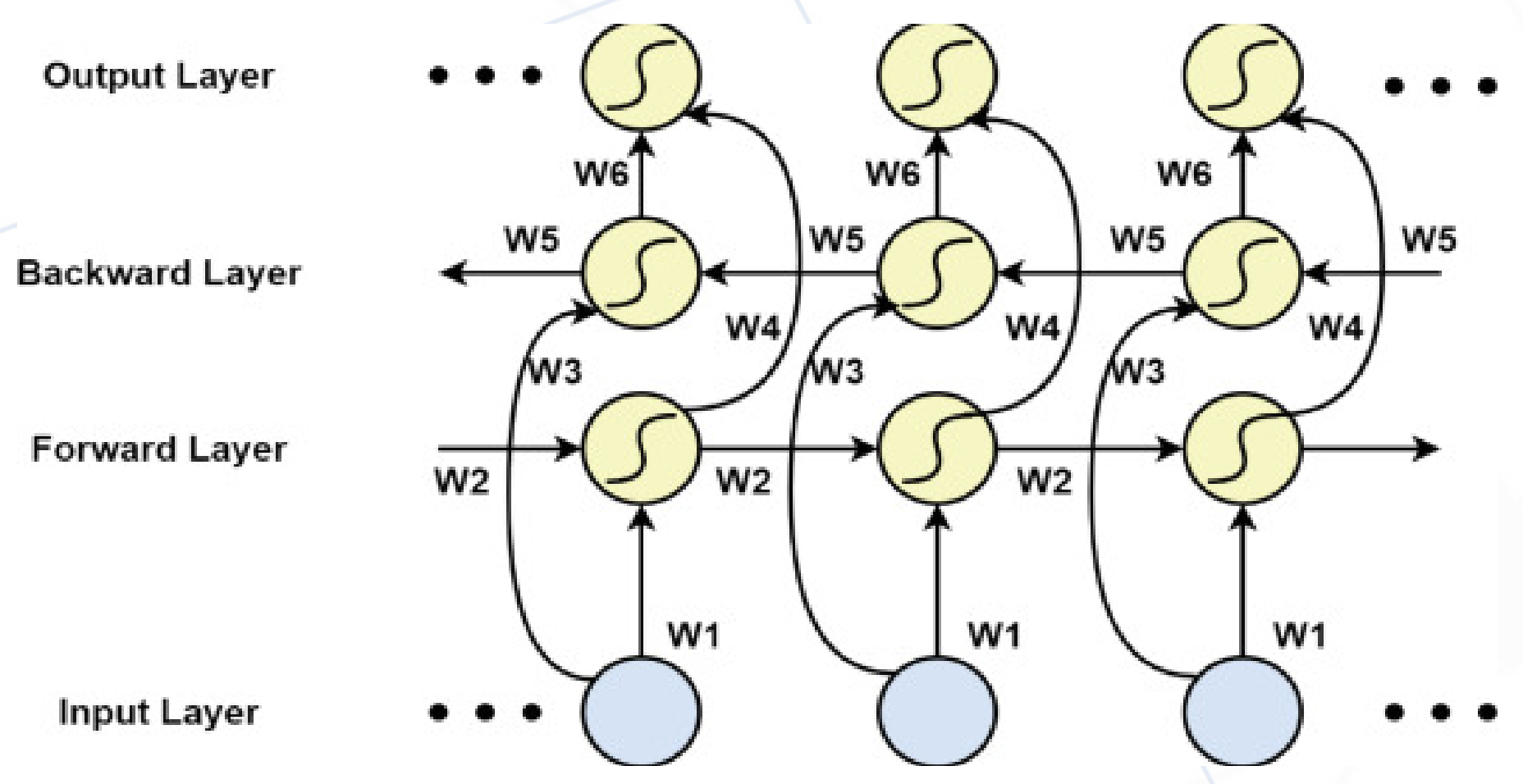

Figure 3: BiLSTM Structure
(Science Direct, n.d.)

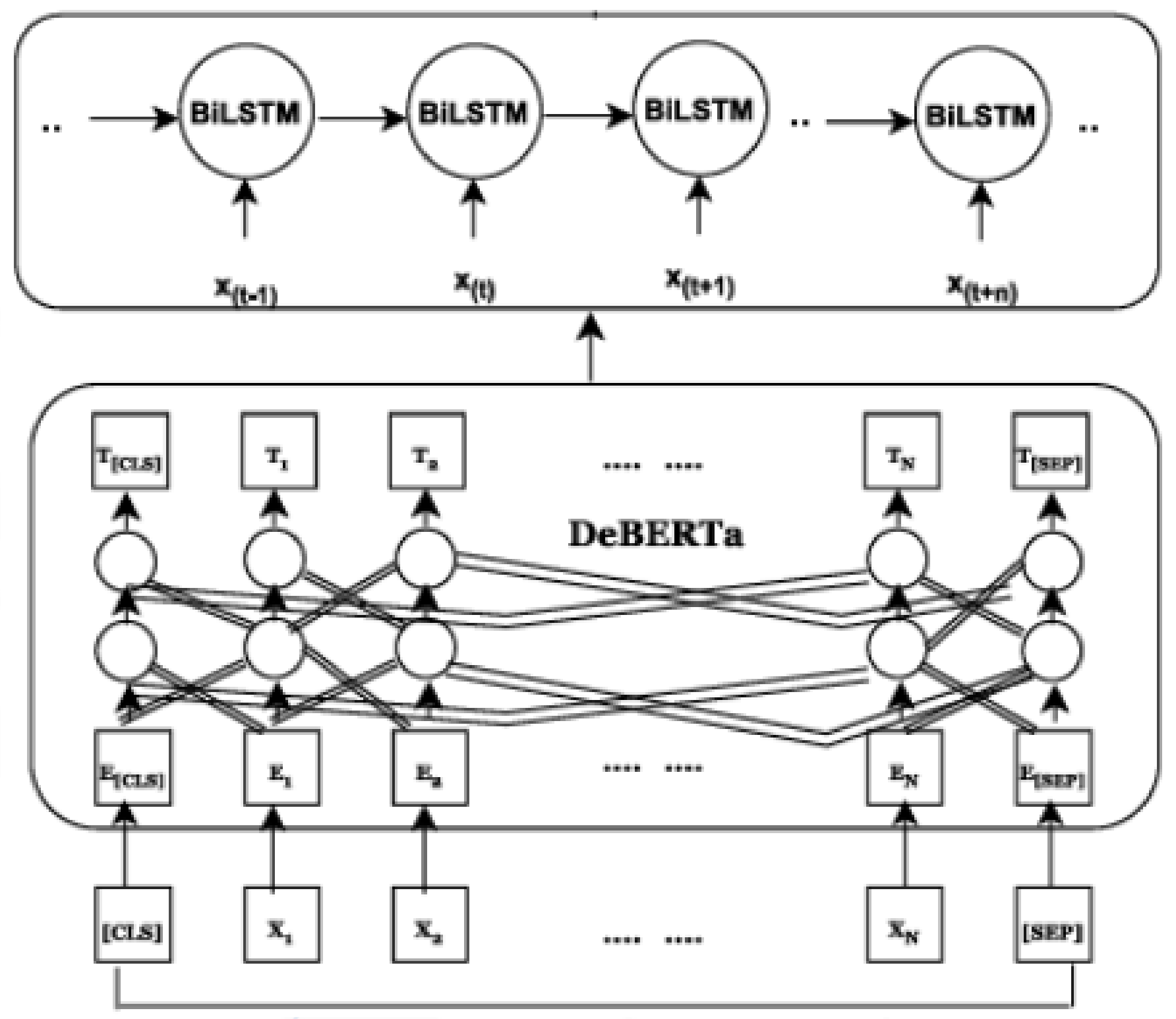

Figure 4: BiLSTM with DeBERTa embeddings
(Alqahtani, S.M., et al. 2023)

Our model had the following structure:

- Frozen DeBERTa model as embeddings (Figure 3)
- Batch Normalisation
- Fully Connected Layer
- Dropout
- Final classifier with sigmoid activation

RESULTS

Model	Accuracy	F1-Score	Precision	Recall	MCC	ROC AUC
Not pre-trained	60.88%	60.89%	60.88%	60.88%	36.88%	60.88%
gloVe.6B.100d	70.26%	70.19%	70.26%	70.26%	38.26%	67.38%
gloVe.27B.200d	69.67%	69.68%	69.67%	69.67%	40.30%	69.67%
BERT-base	70.06%	77.03%	79.07%	75.13%	42.35%	70.49%
DeBERTa-base	**87.41%**	**87.26%**	**91.38%**	**83.50%**	**75.17%**	**87.54%**

Table 1: Metrics across different embeddings

Model	Accuracy	F1-Score	Precision	Recall	MCC	ROC AUC
BERT-base	82.20%	83.61%	79.68%	87.95%	64.61%	82.01%
AlBERT-base	82.82%	83.85%	81.48%	86.37%	65.67%	82.70%
RoBERTa-base	87.90%	88.10%	89.48%	86.77%	75.84%	87.94%
DeBERTaV3-base	**92.34%**	**92.65%**	**91.79%**	**93.53%**	**84.67%**	**92.30%**

Table 2: Metrics across different transformer models

Models are evaluated on the **development set** containing 6K premise-hypothesis pairs.

CONCLUSION

- **DeBERTa** has great potential to be leveraged on for DL approaches, regardless of underlying architecture.
- Improvements when **freezing DeBERTaV3-base as an embedding layer** was surprising when compared to:
    - Training embeddings from scratch (+**26.26**% acc)
    - Pre-trained glove embeddings (+**16.88**% acc).
- Our model has **applicability** to a range of tasks, as shown by DeBERTaV3's performance on the GLUE benchmark.

FURTHER STEPS

- **Population based training** for more optimal hyperparameter tuning
- **Synthetic data generation** as used by state of the art (UnitedSynT5 for Few-Shot NLI)
- **Ensemble learning** to improve model accuracy and robustness

References