data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language （META）

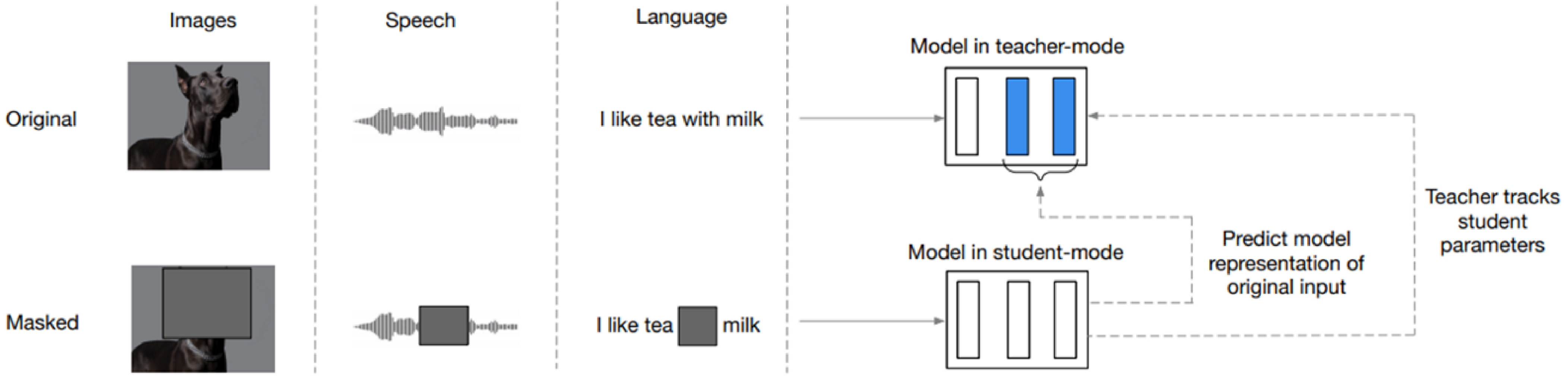
Motivation

- There are currently big differences in the way self-supervised learning algorithms learn from images, speech, text, and other modalities.

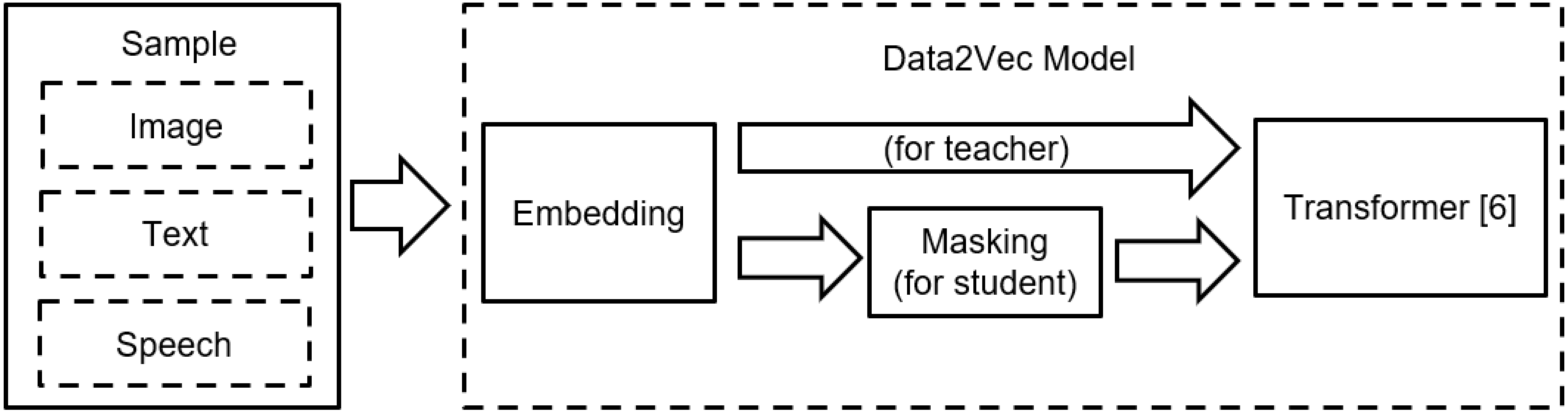
  data2vec, the first high-performance self-supervised algorithm that works for multiple modalities ✅

Method Overview

The teacher mode generates representation from a given sample (i.e. image, speech, text). A masked version of the same sample is passed to the student mode. Learning happens by minimizing the objective function between the student's prediction of a target that is constructed by teachers' parameters.



Method



	Embedding Method	Masking Method
Image	ViT Strategy [1][2]	Block-wise masking strategy [1]
Text	1D CNN [3]	Mask tokens [5]
Speech	Pre-processed to obtain sub-word units [4][5]	Mask spans of latent speech representation [3]

- The target is then constructed by using the top K (closer to the output) blocks of the transformer.

$$y_t = \frac{1}{K}\sum_{l=L-K+1}^{L} \hat{a}_t^l$$

- Objective Function

$$\mathcal{L}(y_t, f_t(x)) = \begin{cases} \frac{1}{2}(y_t - f_t(x))^2/\beta & |y_t - f_t(x)| \leq \beta \\ (|y_t - f_t(x)| - \frac{1}{2}\beta) & \text{otherwise} \end{cases}$$

Results

1. Image (metric: accuracy. Higher value, better performance):

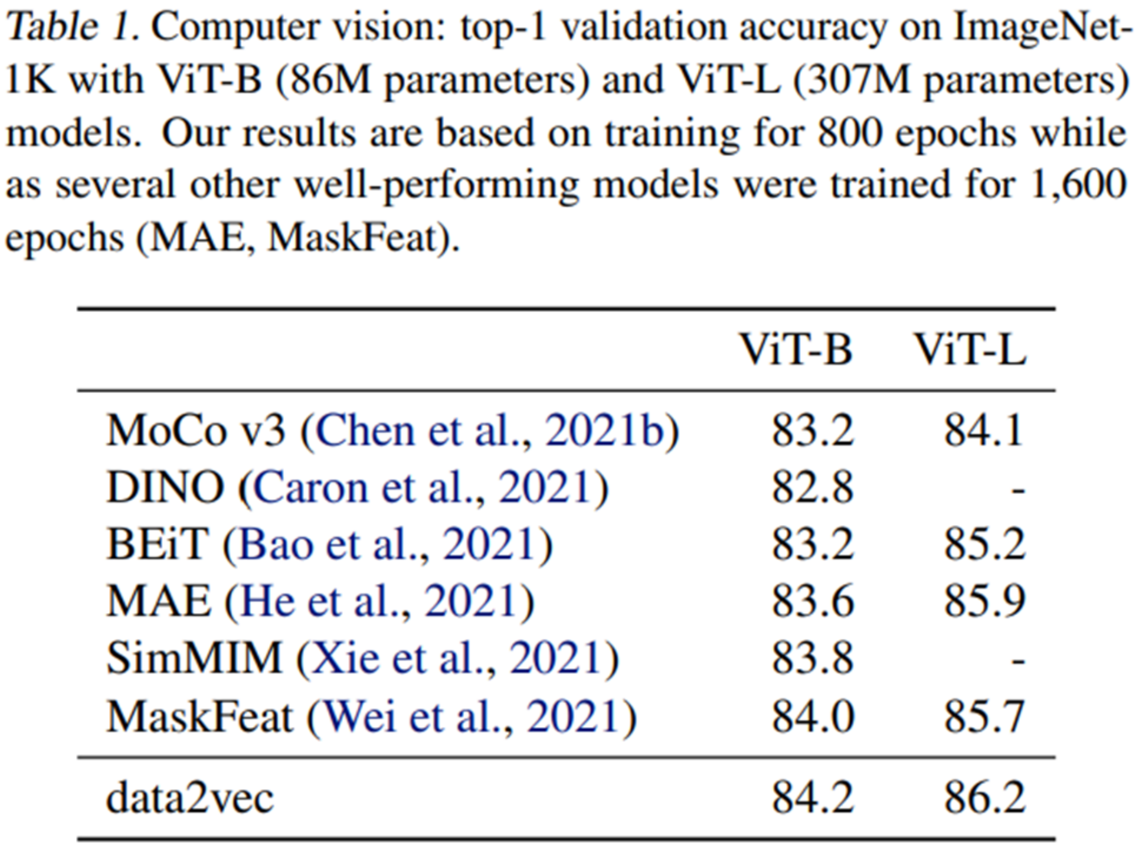


Table 1. Computer vision: top-1 validation accuracy on ImageNet-1K with ViT-B (86M parameters) and ViT-L (307M parameters) models. Our results are based on training for 800 epochs while as several other well-performing models were trained for 1,600 epochs (MAE, MaskFeat).

	ViT-B	ViT-L
MoCo v3 (Chen et al., 2021b)	83.2	84.1
DINO (Caron et al., 2021)	82.8	-
BEiT (Bao et al., 2021)	83.2	85.2
MAE (He et al., 2021)	83.6	85.9
SimMIM (Xie et al., 2021)	83.8	-
MaskFeat (Wei et al., 2021)	84.0	85.7
data2vec	84.2	86.2

2. Speech (metric: word error rate. Lower value, better performance)

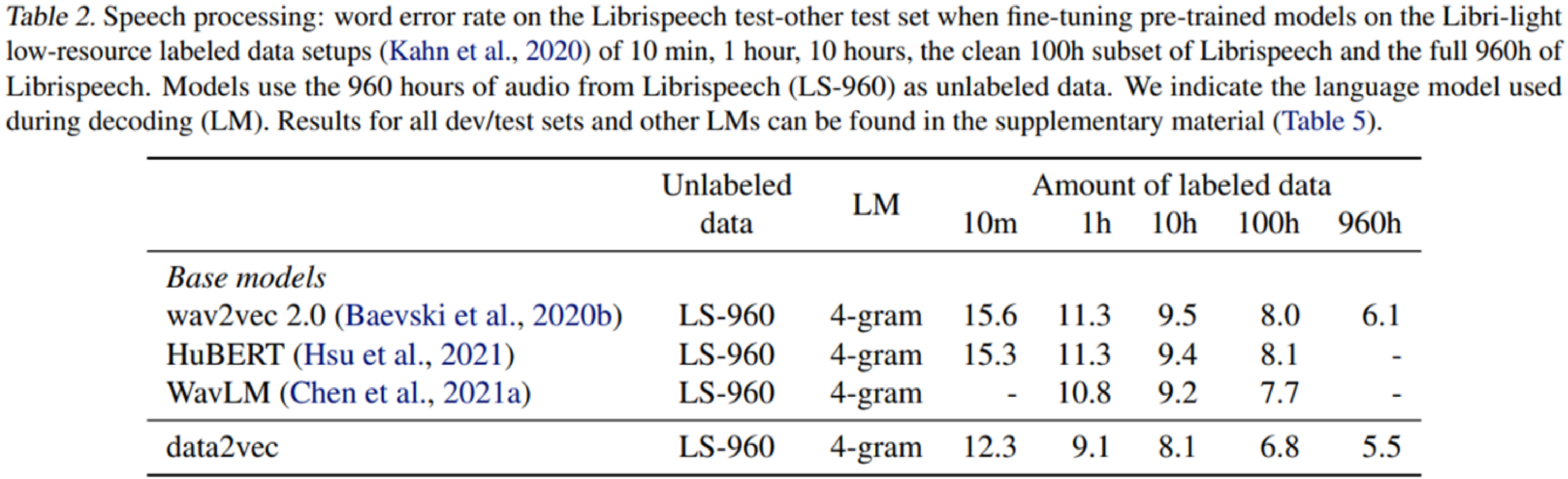


Table 2. Speech processing: word error rate on the Librispeech test-other test set when fine-tuning pre-trained models on the Libri-light low-resource labeled data setups (Kahn et al., 2020) of 10 min, 1 hour, 10 hours, the clean 100h subset of Librispeech and the full 960h of Librispeech. Models use the 960 hours of audio from Librispeech (LS-960) as unlabeled data. We indicate the language model used during decoding (LM). Results for all dev/test sets and other LMs can be found in the supplementary material (Table 5).

	Unlabeled data	LM	Amount of labeled data				
			10m	1h	10h	100h	960h
Base models							
wav2vec 2.0 (Baevski et al., 2020b)	LS-960	4-gram	15.6	11.3	9.5	8.0	6.1
HuBERT (Hsu et al., 2021)	LS-960	4-gram	15.3	11.3	9.4	8.1	-
WavLM (Chen et al., 2021a)	LS-960	4-gram	-	10.8	9.2	7.7	-
data2vec	LS-960	4-gram	12.3	9.1	8.1	6.8	5.5

3. Texts (metric: GLUE score. Higher value, better performance)

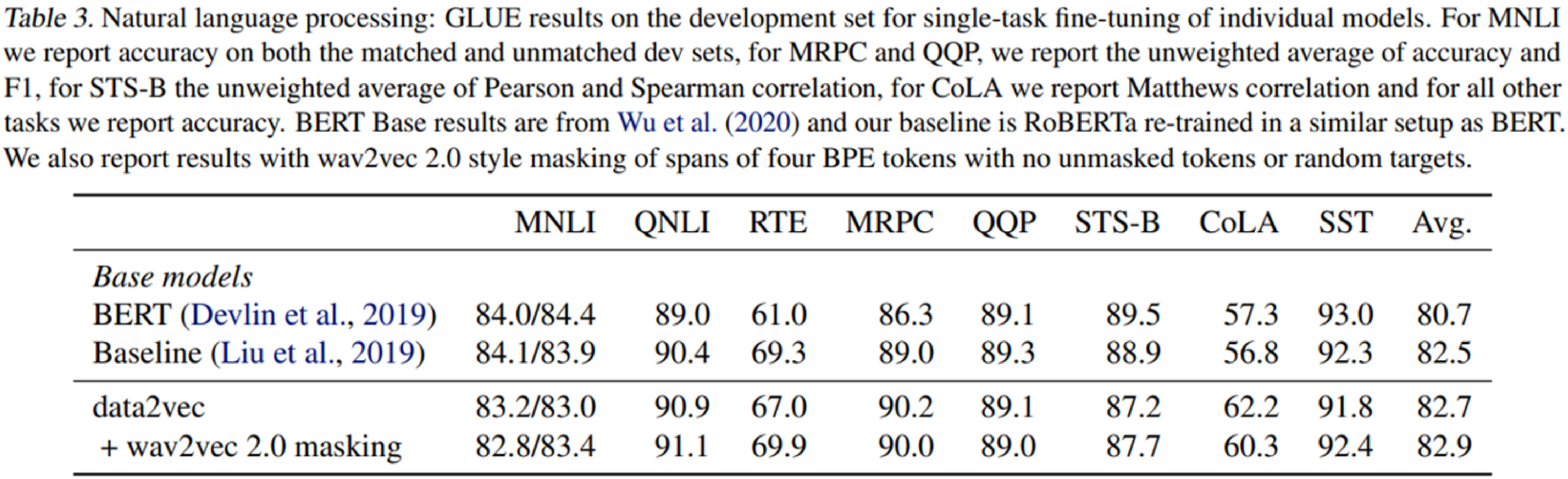


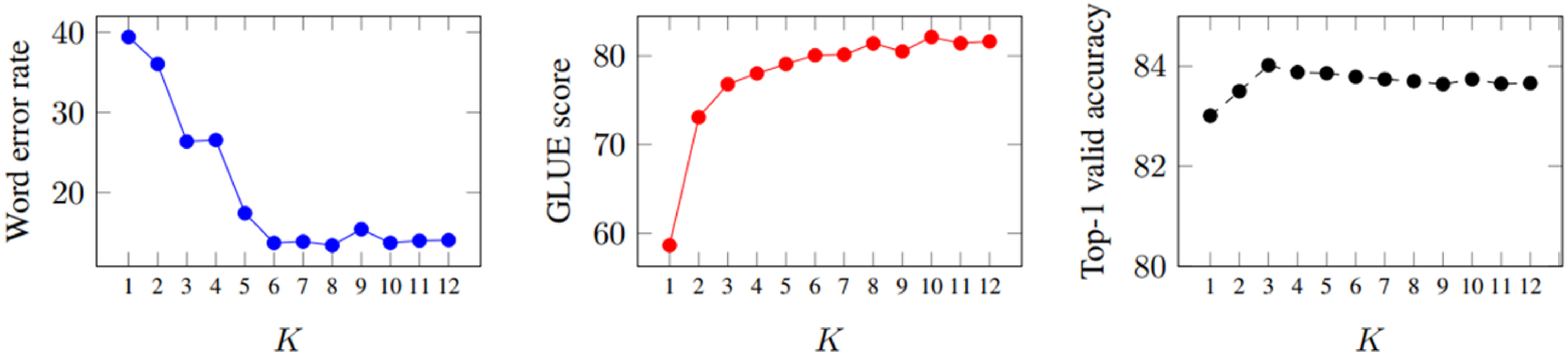
Table 3. Natural language processing: GLUE results on the development set for single-task fine-tuning of individual models. For MNLI we report accuracy on both the matched and unmatched dev sets, for MRPC and QQP, we report the unweighted average of accuracy and F1, for STS-B the unweighted average of Pearson and Spearman correlation, for CoLA we report Matthews correlation and for all other tasks we report accuracy. BERT Base results are from Wu et al. (2020) and our baseline is RoBERTa re-trained in a similar setup as BERT. We also report results with wav2vec 2.0 style masking of spans of four BPE tokens with no unmasked tokens or random targets.

	MNLI	QNLI	RTE	MRPC	QQP	STS-B	CoLA	SST	Avg.
Base models									
BERT (Devlin et al., 2019)	84.0/84.4	89.0	61.0	86.3	89.1	89.5	57.3	93.0	80.7
Baseline (Liu et al., 2019)	84.1/83.9	90.4	69.3	89.0	89.3	88.9	56.8	92.3	82.5
data2vec	83.2/83.0	90.9	67.0	90.2	89.1	87.2	62.2	91.8	82.7
+ wav2vec 2.0 masking	82.8/83.4	91.1	69.9	90.0	89.0	87.7	60.3	92.4	82.9

Ablation Study

1. Top K blocks

The paper argues that using the average of top K blocks in the teacher mode is better than using just the top one



(a) Speech

(b) NLP

(c) Vision

2. Target Feature Type

Rather than just use the top K blocks, the authors also tried using different parts of the teacher mode and found that using the FFN is the best.

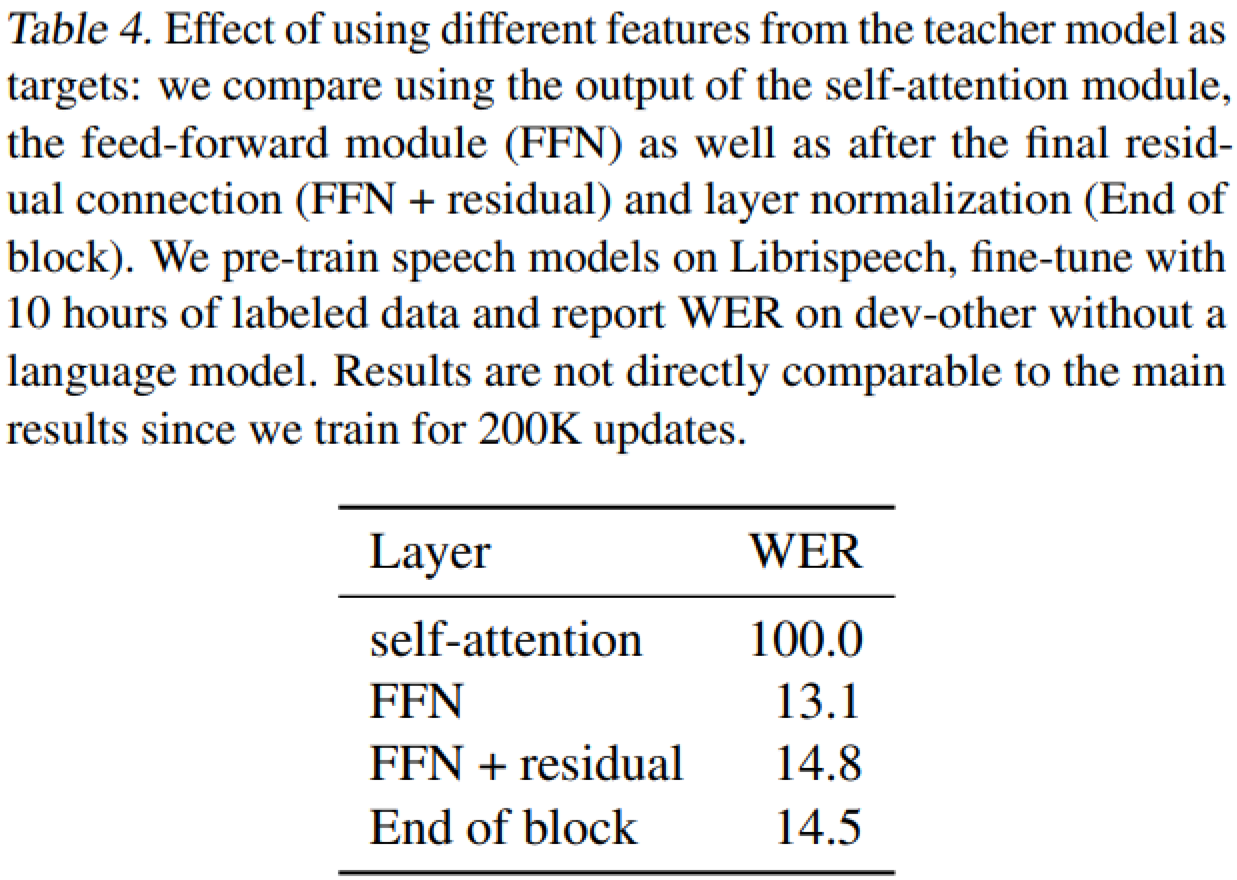


Table 4. Effect of using different features from the teacher model as targets: we compare using the output of the self-attention module, the feed-forward module (FFN) as well as after the final residual connection (FFN + residual) and layer normalization (End of block). We pre-train speech models on Librispeech, fine-tune with 10 hours of labeled data and report WER on dev-other without a language model. Results are not directly comparable to the main results since we train for 200K updates.

Layer	WER
self-attention	100.0
FFN	13.1
FFN + residual	14.8
End of block	14.5

Conclusion

- The paper introduces a new general self-supervised learning framework and achieves SOTA performance for three modalities.

Personal Remarks:

- It’d be more interesting to see how this method performs for unstructured modality, e.g. graphs
- Transformer, as a flexible architecture not constraint to a specific modality, plays an important role in the success of this method.
- This work serves as a key step for unifying inputs from different modalities.