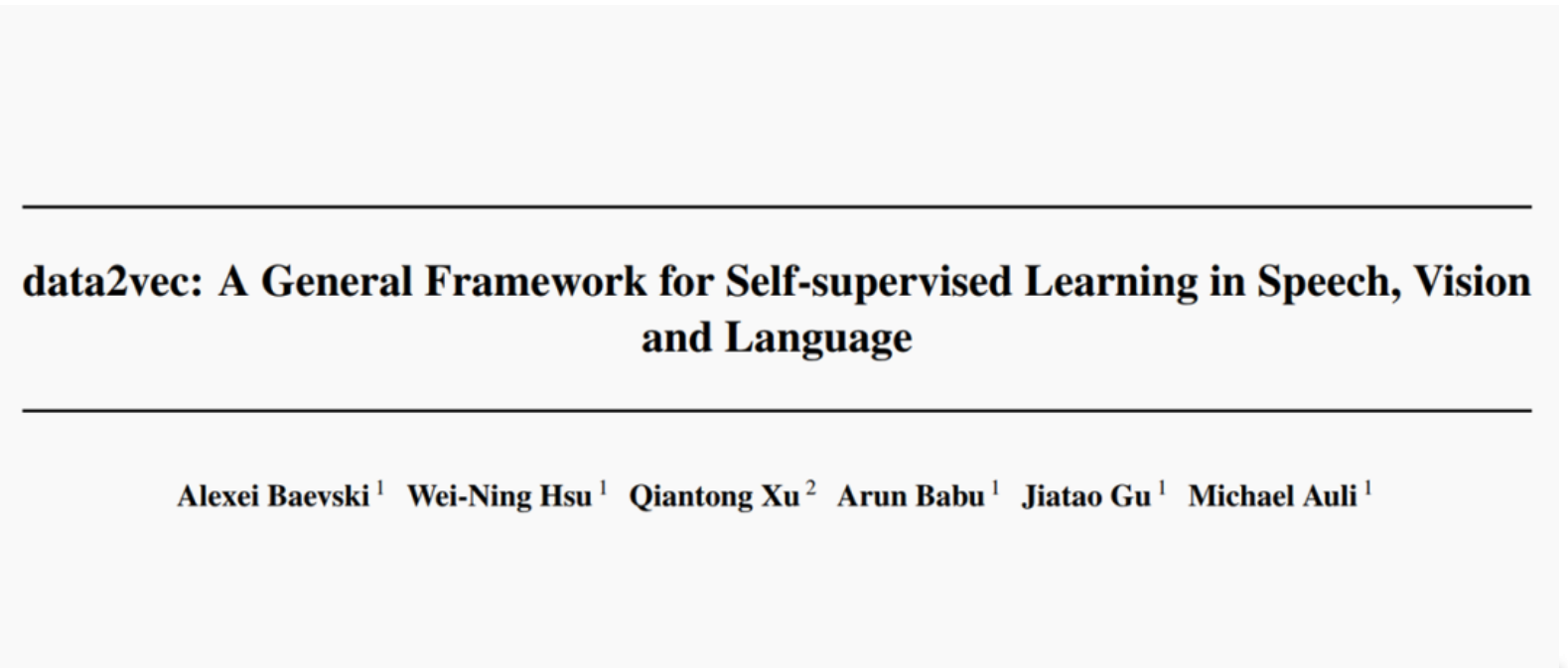
Meta AI 发布 data2vec！统一模态的新里程碑！

原创 ZenMoore 夕小瑶的卖萌屋 2022-01-22 22:05



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

Alexei Baevski[1] Wei-Ning Hsu[1] Qiantong Xu[2] Arun Babu[1] Jiatao Gu[1] Michael Auli[1]

文 | ZenMoore
编 | 小轶

如果让大家举一个最成功的自监督模型的例子，尤其对于各位 NLPer，肯定毫不犹豫地祭出我大 BERT. 想当年 BERT 打了一个名叫 MLM (Masked Language Model) 的响指，直接成了 NLP 灭霸。

视觉界、语音界闻声而来，纷纷开启了 BERT 的视觉化、语音化的改造。

视觉界，以 patch 或者像素类比 NLP 的 token；语音界，虽然不能直接找到 token 的替代，但是可以专门做 quantification 硬造 token.

但是，思考这样一个问题：为什么这些图像或者语音模态的自监督，非要一股 NLP 味儿呢?

要知道，虽然确实有生物学的研究表明，人类在进行视觉上的学习时，会使用与语言学习相似的机制，但是，**这种 learning biases 并不一定完全可以泛化到其他模态**。

所以有没有什么办法，能够**把不同模态的自监督表示学习统一起来**，不再是仿照 MLM 做 MIM (Masked Image Modelling)、MAM (Masked Audio Modelling)?

昨天，Meta AI （原 Facebook）发布最新自监督学习框架 Data2Vec，立即在整个 AI 圈疯狂刷屏。这份工作或许预示着——多模态的新时代，即将到来。

本文就为大家简单解读一下，这份 AI 圈的今日头条，究竟做了些什么。

论文标题:
Data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language

论文作者:
Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, Michael Auli

Meta AI, SambaNova

论文链接:
https://ai.facebook.com/research/data2vec-a-general-framework-for-self-supervised-learning-in-speech-vision-and-language

模型算法

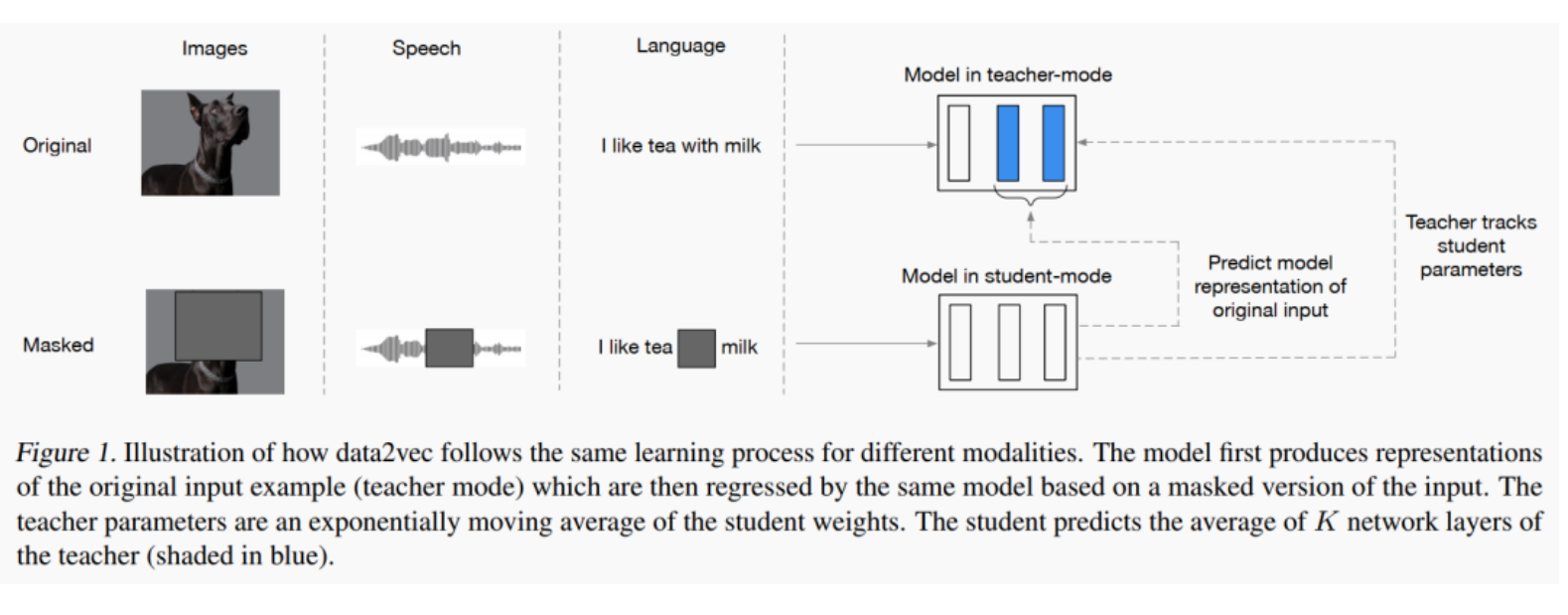


Figure 1. Illustration of how data2vec follows the same learning process for different modalities. The model first produces representations of the original input example (teacher mode) which are then regressed by the same model based on a masked version of the input. The teacher parameters are an exponentially moving average of the student weights. The student predicts the average of K network layers of the teacher (shaded in blue).

编码、掩码

首先，对于三个不同的模态：文本、图像、语音，采用不同的编码方式以及掩码方式。

模态特定的编码方式 ：

1. 文本模态：token embedding
2. 图像模态：参考 ViT[1, 2]，以 image patch 为单位，经过一个线性变换(linear transformation)
3. 语音模态：使用多层一维卷积对 waveform 进行编码[3]。

模态特定的掩码方式：

1. 文本模态：对 token 掩码
2. 图像模态：block-wise masking strategy [2]
3. 语音模态：对语音模态来说，相邻的音频片段相关性非常大，所以需要对 span of latent speech representation 进行掩码 [3]

> 掩码符为训练后得到的 MASK embedding token，而不是简单的 MASK token，原因且看下文。

Student：模型训练

之后，在 student-mode 中，根据 masked input 对掩码位置的**表示**进行预测。需要注意的是，这里模型预测的并不是掩码位置(如 text token, pixel/patch, speech span)，而是**掩码位置经过模型编码后的表示**。因为这个表示经过了 Attention/FFN 等一系列模块的处理，自然是**模态无关的**，不仅如此，它还是**连续的(continuous)**，编码了丰富的**上下文语义(contextualized)**。

> 如果把输入空间比作物理世界，表示空间比作精神空间。那么，作者相当于直接在“精神空间”中想象被遮住的部分(mask),颇有一种“梦里看花”的感觉。上次见到这“梦一般”的算法，还是 Hinton 老爷子的 Sleep-Wake[4].

具体地，训练目标为如下的 smooth L1 loss：

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

其中，y_t 为使用 teacher model 构建的 training target；$f_t(x)$ 为 student model 在时刻 t 的输出；β 是超参，用来调整 L1 损失的平滑度。

Teacher：数据构建

最后，还有一个问题，既然变成了对表示的掩码而非对原输入的掩码，那么训练数据怎么来呢?

这就是 teacher-mode 的妙用。与 student-mode 不同的是，teacher-mode 的输入不再是 masked input，而是 original input, 这样，掩码位置对于 teacher 来说就是可见的，自然能够得到掩码位置对应的表示，而这个表示，就是 student-mode 的 **training target**.

当然，为了保证“师生”两个模型的一致性，两者的**参数是共享的**。另外，又为了在训练初期让 Teacher 的参数更新更快一些，作者采用了一个指数滑动平均(EMA)：$\Delta \leftarrow \tau\Delta + (1-\tau)\theta$.

其中，Δ 是 Teacher 的参数，θ 是 Student 的参数，τ 类似于学习率，也是一个带有 scheduler 的参数。

具体地，training target 这么构建(按步骤)：

1. 找到 student-mode 输入中被 mask 掉的 time-step$t$
2. 计算 teacher network 最后 K 层 transformer block 的输出：$(a_t^l)_{l\in[[L-K+1,L]]}$
3. 归一化 :$\hat{a}_t^l$
4. 平均：$y_t = \frac{1}{K}\sum_{l=L-K+1}^{L}\hat{a}_t^l$, 即 training target.

> 对于第三步的归一化：语音模态采用 instance normalization 文本和图像模态采用 parameter-less layer normalization

Representation Collapse

在实验中，作者还遇到了 Representation Collapse 的问题：模型对于所有掩码片段输出非常相似的 representation.

这个已经有好多解决办法啦~ 对于本文，有以下几种情况：

1. 学习率太大或者其 warmup 太短：通过调参解决
2. 指数滑动平均太慢了：还是调参
3. 对于相邻 target 相关性强的模态或者掩码片段较长的模态 (比如语音模态)：设置 variance 罚项[5]，或者归一化[6]，归一化的效果更好一些。
4. 而对于 targets 相关性不那么强的模态例如 NLP/CV 等，momentum tracking 就足够。

与同类工作的对比

与其他 NLP 自监督算法的对比：

和 BERT 不同，本文预测的并不是离散 token，而是 continuous/contextualized representation.

好处1：target 不是 predefined (比如有预定义的词表等)，target set 也是无限的 (因为是连续的)，因此可以让模型更好的适配特定的输入

好处2：考虑了更多上下文信息

与其他 CV 自监督算法的对比：

1. 与 BYOL[6]/DINO[7] 等：本文新增了掩码预测任务，而且是对多层进行回归(即参数 K)
2. 与 BEiT[2]/MAE[8] 等带掩码预测任务的算法：本文对 latent representation 进行预测

与其他 Speech 自监督算法的对比:

1. 与 Wav2vec2.0[3]/HuBERT[9] 等：其他工作一般需要另外预测 speech 中的离散单元（或联合学习或交互学习），而本文不需要 这种 quantification.

与多模态预训练的对比：

本文工作重点不在于多模态任务或者多模态训练，而在于如何把不同模态的自监督学习目标统一起来。

实验结果

计算机视觉

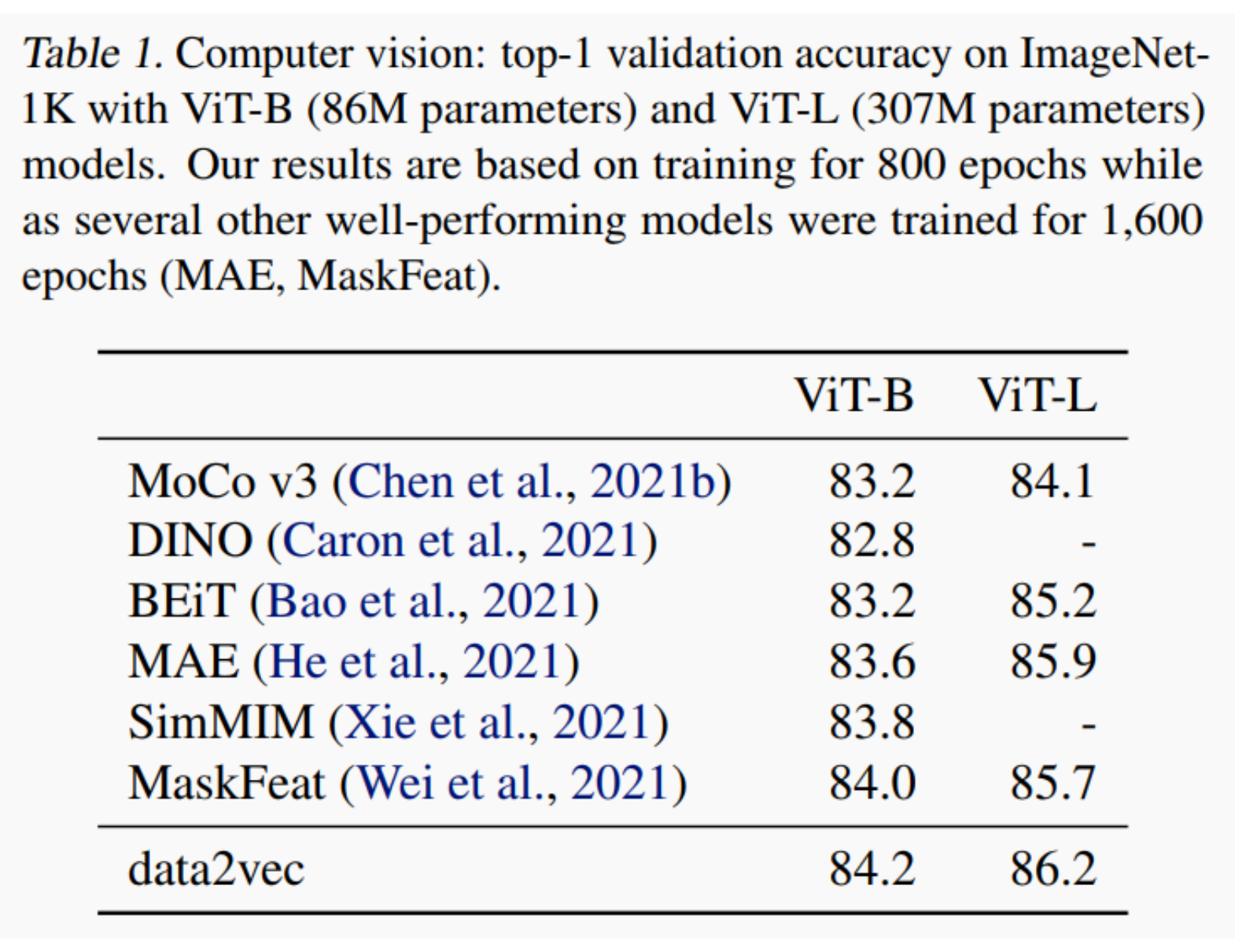


Table 1. Computer vision: top-1 validation accuracy on ImageNet-1K with ViT-B (86M parameters) and ViT-L (307M parameters) models. Our results are based on training for 800 epochs while as several other well-performing models were trained for 1,600 epochs (MAE, MaskFeat).

	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

实验任务：Image Classification

实验结论：可以看到本文工作有较明显的改进

语音

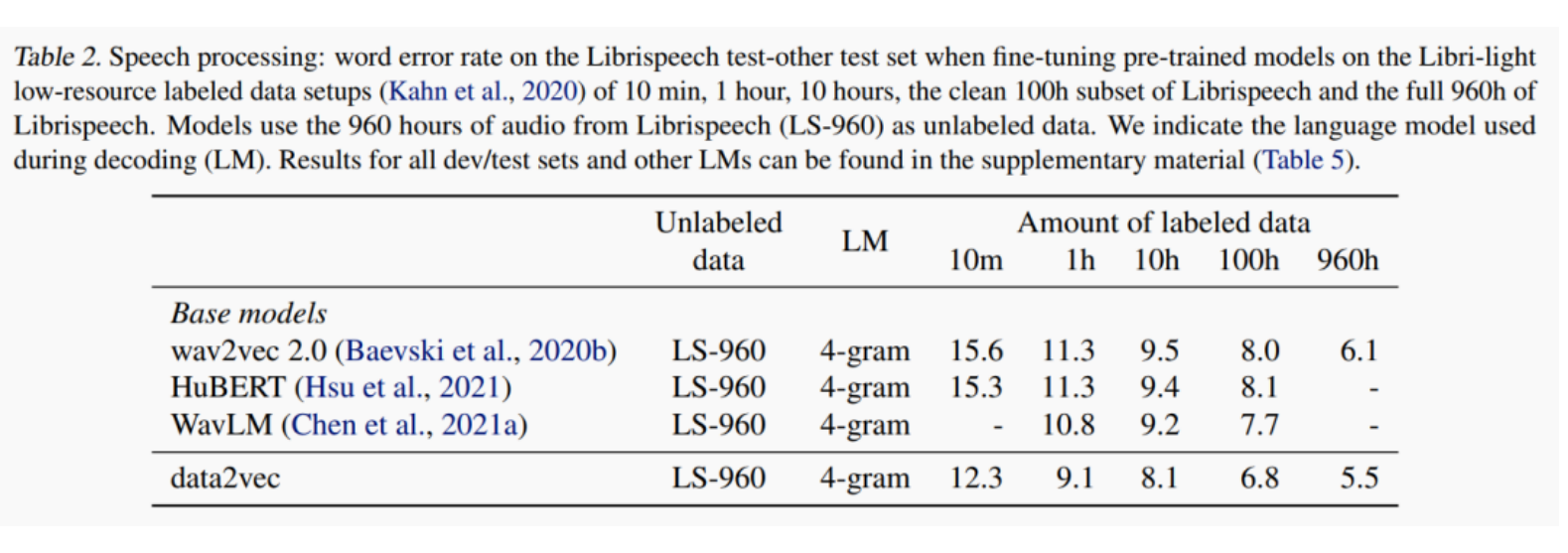


Table 2. Speech processing: word error rate on the Librispeech test-other test set when fine-tuning pre-trained models on the Libri-light low-resource labeled data setups (Kahn et al., 2020) of 10 min, 1 hour, 10 hours, the clean 100h subset of Librispeech and the full 960h of Librispeech. Models use 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

实验任务：Automatic Speech Recognition

实验结论：改进很明显

Natural Language Processing

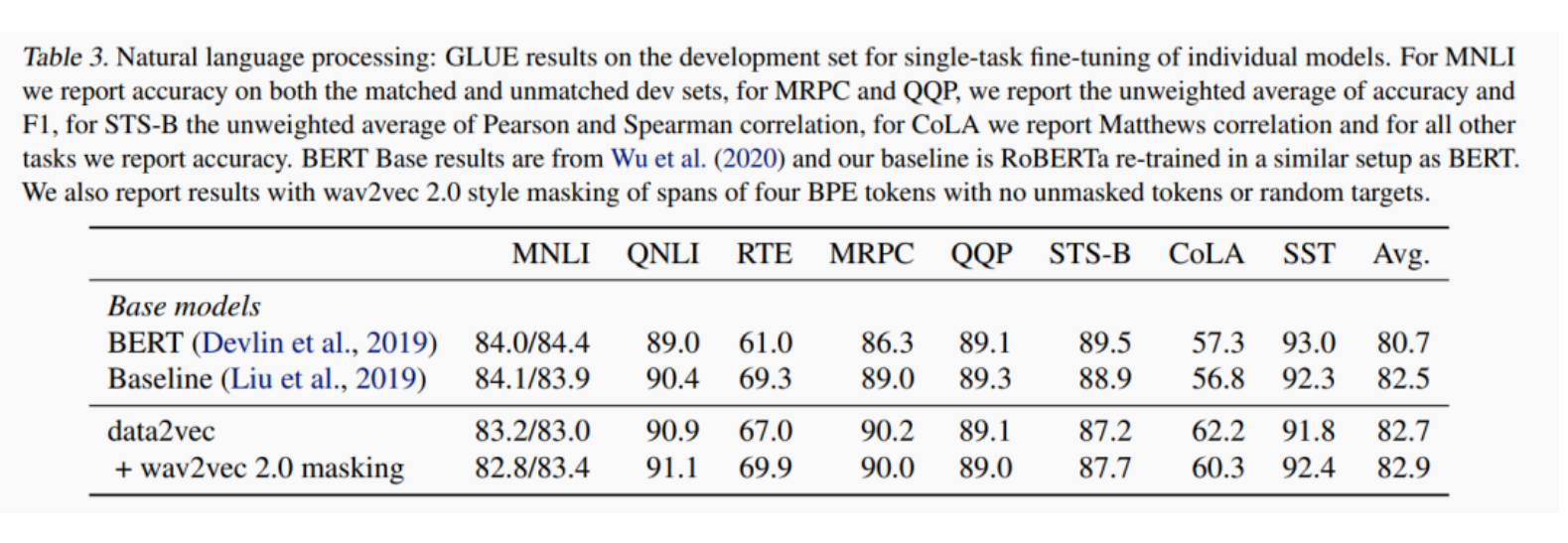


Table 3. Natural language processing: GLUE results on the development set for single-task fine-tuning of individual models. For MNLI we report accuracy on both the matched and unmatched dev sets, for MRPC and QQP, we report the unweighted average of accuracy and F1, for STS-B the unweighted average of 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.

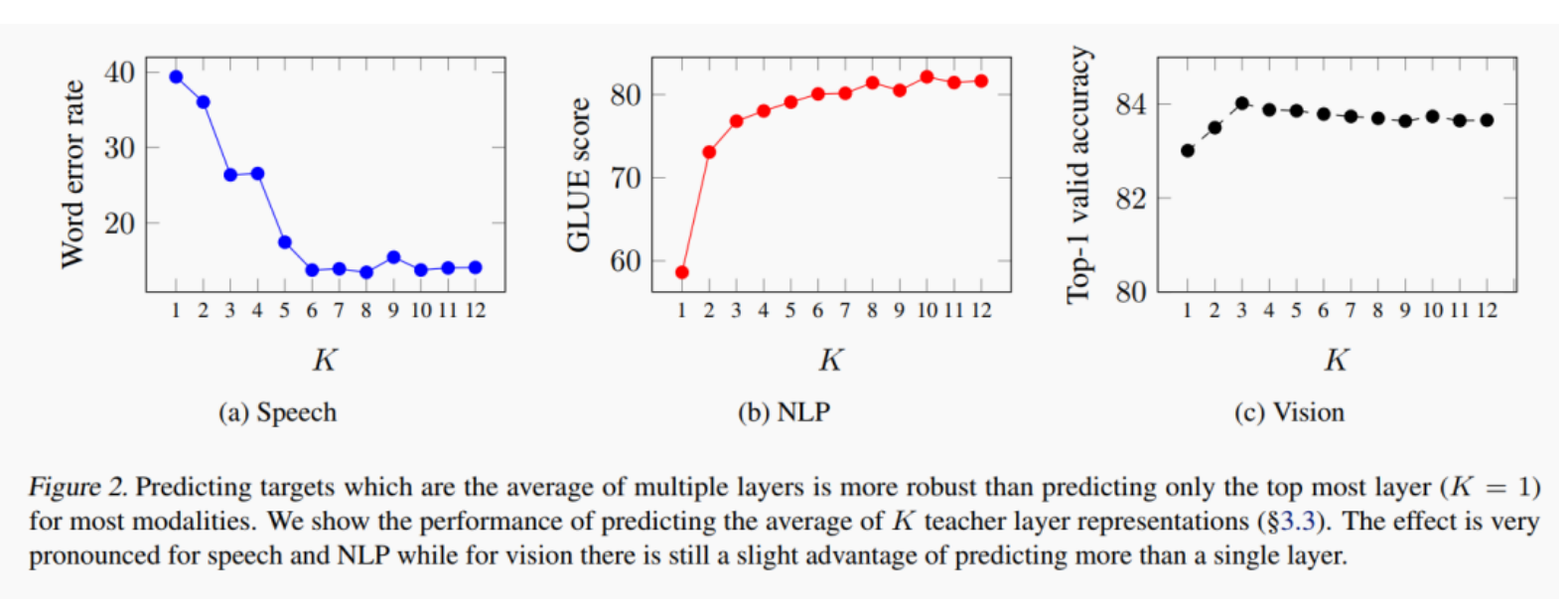
	MNLI	QNLI	RTE	MRPC	QQP	STS-B	CoLA	SST	Avg.
Base models									
BERT (Devlin et al., 2019)	84.0/84.4	89.0	61.0	86.3	89.1	89.5	57.3	93.0	80.7
Baseline (Liu et al., 2019)	84.1/83.9	90.4	69.3	89.0	89.3	88.9	56.8	92.3	82.5
data2vec	83.2/83.0	90.9	67.0	90.2	89.1	87.2	62.2	91.8	82.7
+ wav2vec 2.0 masking	82.8/83.4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

> wav2vec 2.0 masking：masking span of four tokens[3]

实验任务：GLUE

实验结果：作者仅仅对比了 19 年的两个 baseline，说明在文本模态上的改进效果仍然受限，但是这个分数也非常好了

Ablation 1：layer-averaged targets



(a) Speech (b) NLP (c) Vision

Figure 2. Predicting targets which are the average of multiple layers is more robust than predicting only the top most layer ($K = 1$) for most modalities. We show the performance of predicting the average of K teacher layer representations (§3.3). The effect is very pronounced for speech and NLP while for vision there is still a slight advantage of predicting more than a single layer.

这也是和 BYOL[6]/DINO[7] 等模型的一大区分：对多层进行回归

从图表可见，比起只使用 top layer，平均多层输出来构建 target 是很有效的！

Ablation 2：使用 Transformer 的哪一层?

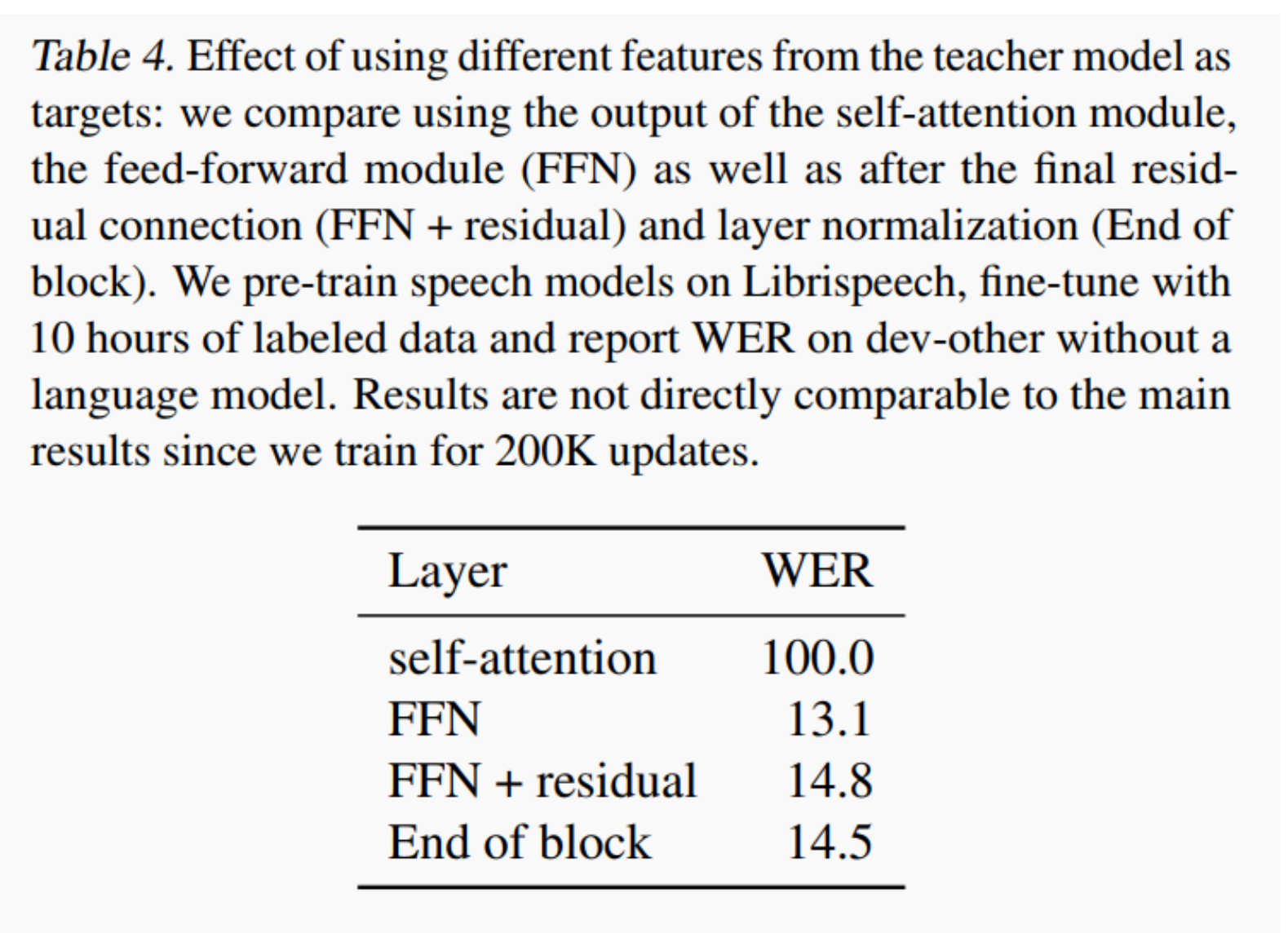


Table 4. Effect of using different features from the teacher model as targets: we compare using the output of the self-attention module, the feed-forward module (FFN) as well as after the final residual connection (FFN + residual) and layer normalization (End of block). We pre-train speech models on 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

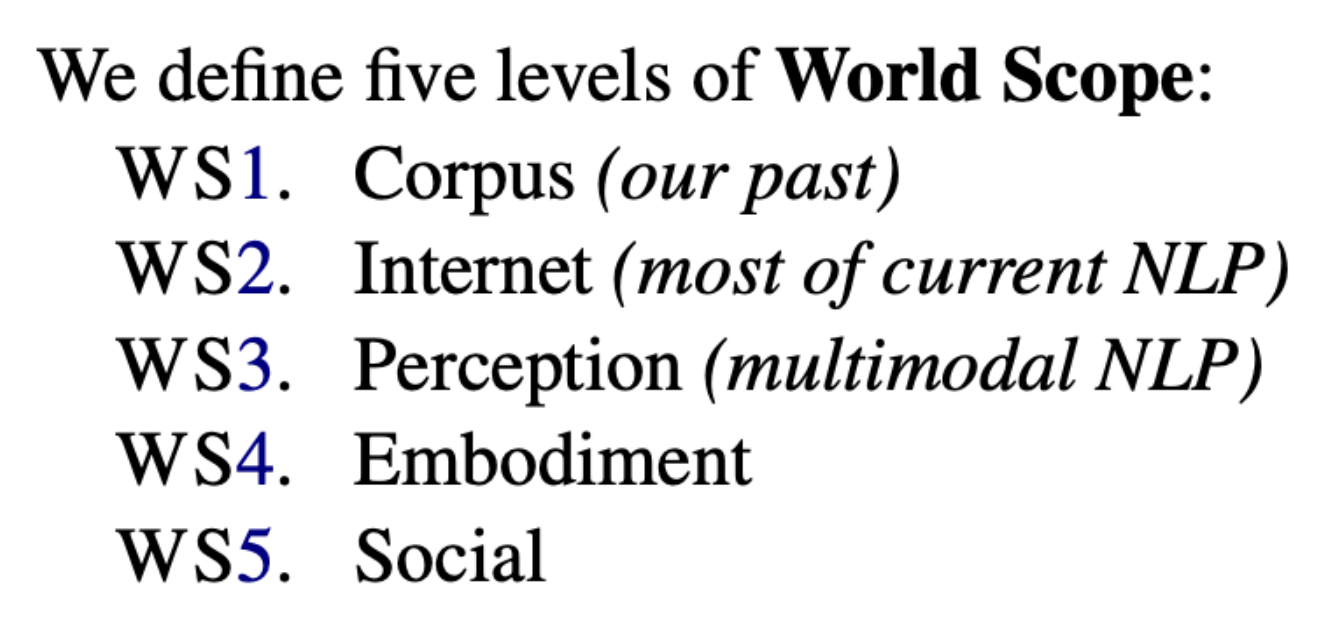
基于语音模态进行实验，发现使用 FFN 层输出最有效，使用自注意力模块的输出基本没用。原因：自注意力模块在残差连接之前，得到的 feature 具有很大的偏差(bias)。

写在最后

也许，**在表示空间中而非输入空间中进行掩码预测的自监督表示学习，是自监督未来的重要方向**！

不过，作者也指出 Data2Vec 的一大局限：**编码方式以及掩码方式仍然是 modality-specific 的**。能否使用类似于 Perceiver[10] 的方式直接在 raw data 上进行操作？或者是否真的有必要统一各个模态的 encoder 呢?

犹记得卖萌屋作者群里有过这么一个分享，是 Yoshua Bengio 等在 EMNLP'20 的文章[11]，里面界定了 NLP 发展的五个阶段：



We define five levels of **World Scope**:
WS1. Corpus *(our past)*
WS2. Internet *(most of current NLP)*
WS3. Perception *(multimodal NLP)*
WS4. Embodiment
WS5. Social

毋庸置疑，多模态的火热标志着我们正在进入第三个阶段：多模态时代。

Data2Vec 巧妙地使用“梦里看花”的方式，让我们看到了自监督的强大威力，也让我们意识到模态统一大业就在眼前！也许，现在的 Data2Vec，只是一颗不能发挥全部威力的宝石，就像 Word2Vec 一样，但相信在不久的将来，从 Data2Vec 出发，能够看到一统多模态的灭霸，就像 BERT 那样！山雨欲来，风满楼！

萌屋作者：ZenMoore
来自北航中法的本科生，数学转码 (AI)，想从 NLP 出发探索人工认知人工情感的奥秘... 个人主页是 zenmoore.github.io，知乎 ID 是 ZenMoore，微信号是 zen1057398161，嘤其鸣矣，求其友声！
作品推荐

1. 一文跟进Prompt进展！综述+15篇最新论文逐一梳理
2. 图灵奖大佬+谷歌团队，为通用人工智能背书！CV 任务也能用 LM 建模！
3. 以4%参数量比肩GPT-3！Deepmind 发布检索型 LM，或将成为 LM 发展新趋势！？

参考文献

[1] An image is worth 16x16 words: Transformers for image recognition at scale. https://arxiv.org/abs/2010.11929

[2] Beit: BERT pre-training of image transformers. https://arxiv.org/abs/2106.08254

[3] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Proc. of NeurIPS, 2020b

[4] The wake-sleep algorithm for unsupervised neural networks https://www.cs.toronto.edu/~hinton/csc2535/readings/ws.pdf

[5] Vicreg: Varianceinvariance-covariance regularization for self-supervised learning. https://arxiv.org/abs/2105.04906

[6] Bootstrap your own latent: A new approach to self-supervised learning https://arxiv.org/abs/2006.07733

[7] Emerging Properties in Self-Supervised Vision Transformers https://arxiv.org/abs/2104.14294

[8] Masked Autoencoders Are Scalable Vision Learners https://arxiv.org/abs/2111.06377

[9] HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units https://arxiv.org/abs/2106.07447

[10] Perceiver: General Perception with Iterative Attention https://arxiv.org/abs/2103.03206

[11] Experience Grounds Language https://arxiv.org/abs/2004.10151