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

Alexei Baevski Wei-Ning Hsu Qiantong Xu Arun Babu Jiatao Gu1
Michael Auli @Meta AI

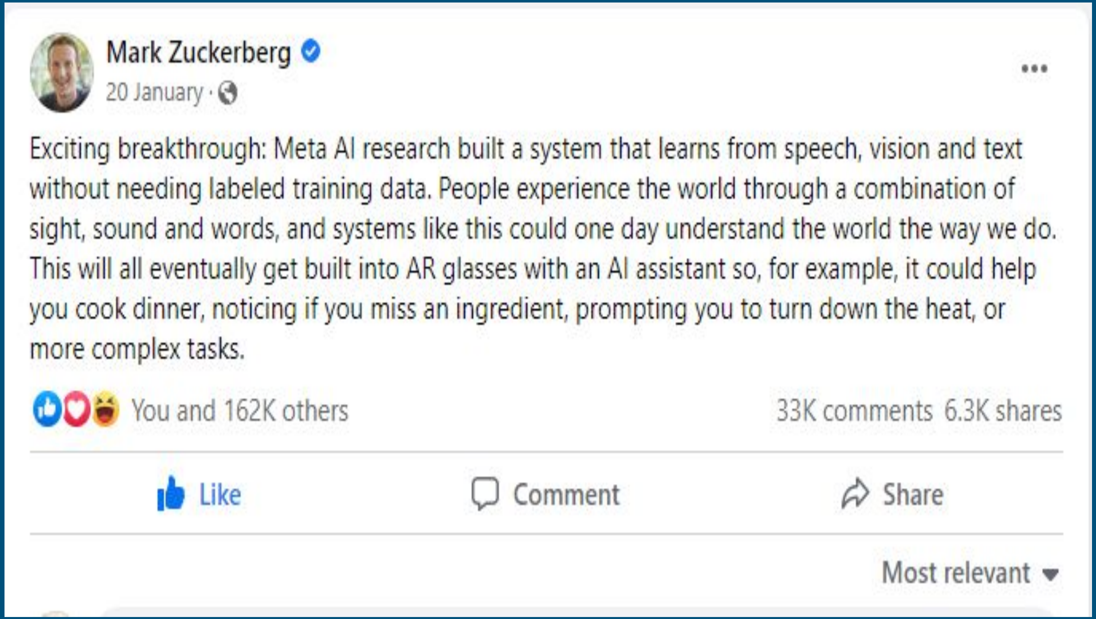
Presented by: Gagan Bhatia

# Outline

1. Motivation
2. Self-supervision Background
3. Training methods
4. Experiments
5. Results
6. Discussions
7. My Thoughts and questions

# Motivation

- One model for speech, text and image
- Make AI understand the world
- Using AI to perform complex tasks

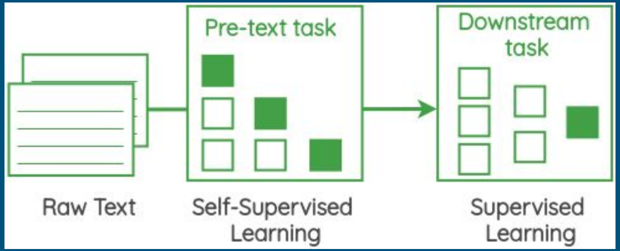


Mark Zuckerberg

20 January ·

Exciting breakthrough: Meta AI research built a system that learns from speech, vision and text without needing labeled training data. People experience the world through a combination of sight, sound and words, and systems like this could one day understand the world the way we do. This will all eventually get built into AR glasses with an AI assistant so, for example, it could help you cook dinner, noticing if you miss an ingredient, prompting you to turn down the heat, or more complex tasks.

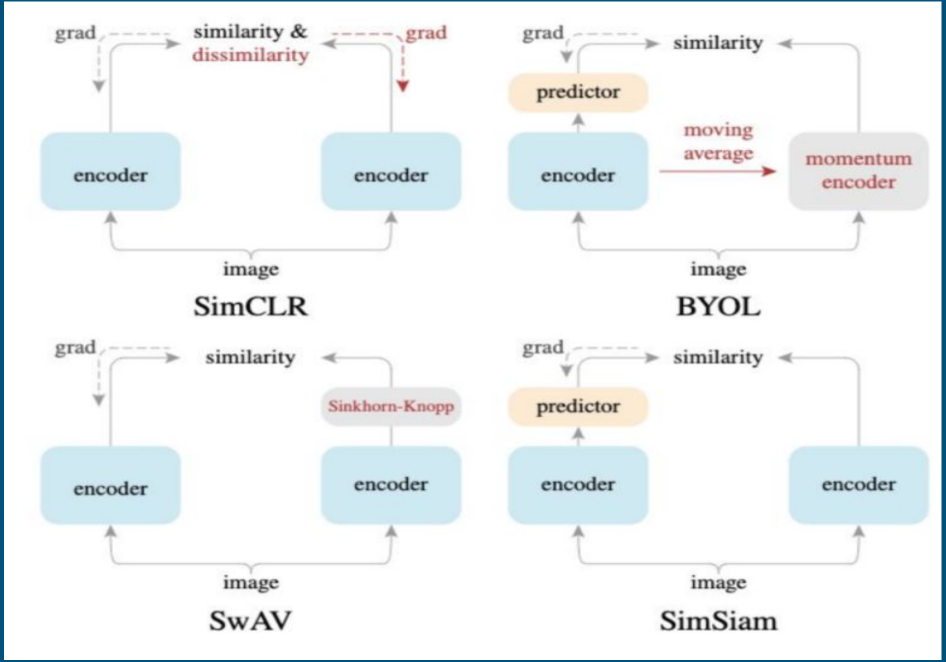
# Self-Supervised Learning



- Instead of using supervised signals from labeled data, SSL exploits the relationship between data.
- Recent self-supervised learning models include frameworks such as Pre-trained Language Models (PTM), GANs, Autoencoder etc.
- The self-supervised learning approach can be described as “the machine predicts any parts of its input for any observed part”

# SSL in Computer Vision - Contrastive

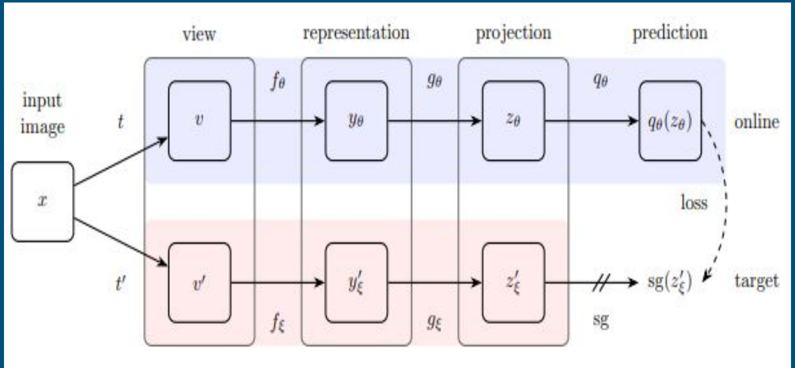
- Unsupervised pre-training for computer vision has been a very active area of research with methods contrasting representations of augmentations of the same image
- Similar to data2vec, BYOL and DINO regress neural network representations of a momentum encoder.



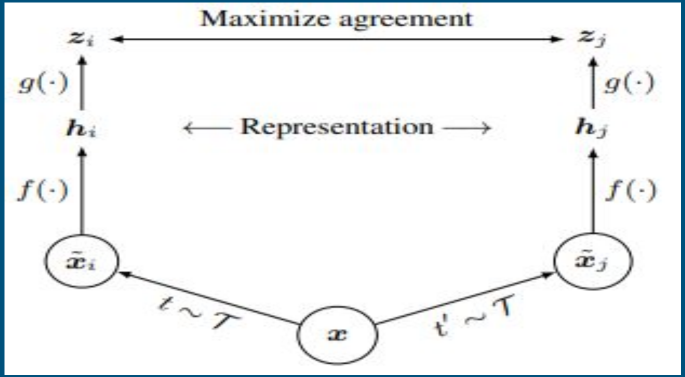
Chen et al., Exploring Simple Siamese Representation Learning
https://arxiv.org/abs/2011.10566

# BYOL and SimCLR

- Contrastive learning (CL) currently achieves state-of-the-art performance in self-supervised learning
- In SimCLR, the input image is transformed by t and t’ to generate two augmented views, and then pass through an encoder f( · ) and a projector g( · ) to get a projected representation.
- The projected representations $z_i$ and z□ are then contrasted to maximize their agreement, which is found to lead to better performance than maximizing agreements between $h_i$ and h□ directly. Negative pairs are constructed by using views from a different input image.
- The difference, however, lies in the factor that in BYOL, two views are generated through different encoders f_θ and f_ξ. These two are of the same architecture but different parameters. Also, in BYOL, there is a predictor and a target network.



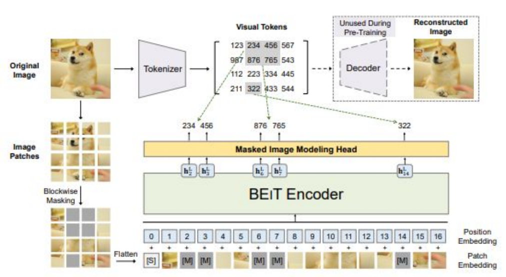
Bootstrap your own latent: A new approach to self-supervised Learning https://arxiv.org/abs/2006.07733

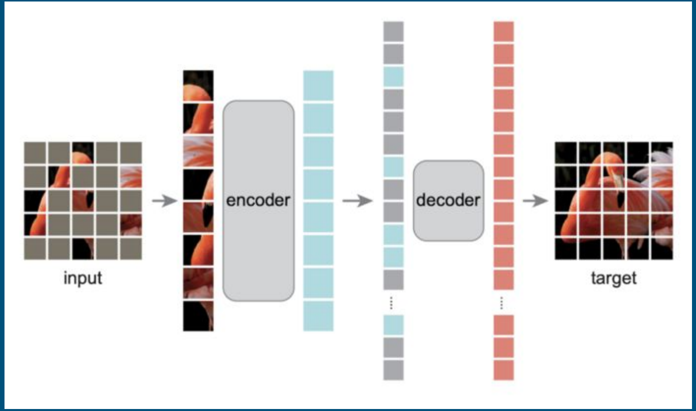


A Simple Framework for Contrastive Learning of Visual Representations https://arxiv.org/abs/2002.05709

# SSL in Computer Vision - ViT, Mask

- The most recent work focuses on training vision Transformers with masked prediction objectives
- MAE, a simple autoencoder that uses partial observation (the input image is not complete) and then turns out the image is entire
- The encoder in this is ViT (Vision Transformer) and is just applied on visible, unmasked patches.
- The input of the decoder is a full set of tokens, including (i) encoded visible patches.

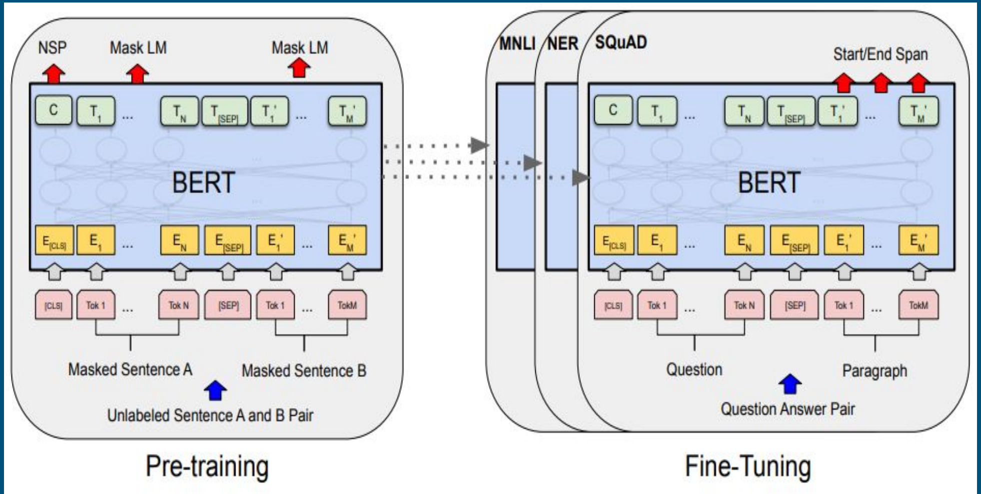




Masked Autoencoders Are Scalable Vision Learners
https://arxiv.org/abs/2111.06377

# SSL in NLP

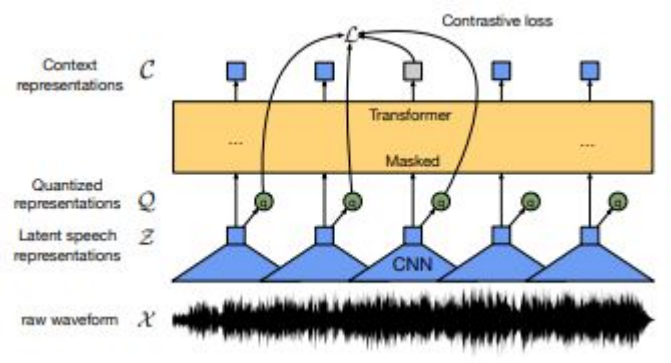
- Pre-training has been very successful in advancing natural language understanding.
- The most prominent model is BERT which solves a masked prediction task where some of the input tokens are blanked out in order to be predicted given the remaining input.



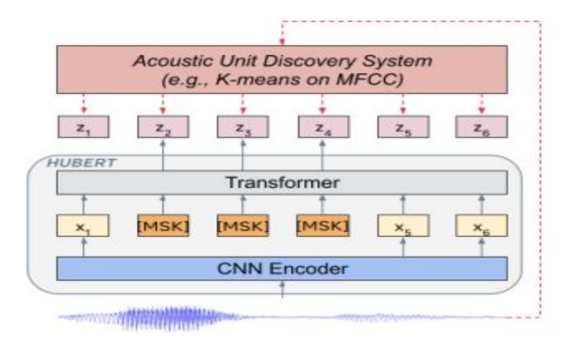
BERT

# SSL in Speech

- Self-supervised learning for speech includes autoregressive models as well as bi-directional models.
- wav2vec 2.0 and HuBERT are based on predicting discrete units of speech.



Wav2vec2



HuBERT

# MultiModal Pretraining

- Video-Audio-Text Transformer, or VATT, is a framework for learning multimodal representations from unlabeled data using convolution-free Transformer architectures.
- VATT borrows the exact architecture from BERT and ViT except the layer of tokenization and linear projection reserved for each modality separately.
- VATT linearly projects each modality into a feature vector and feeds it into a Transformer encoder. A semantically hierarchical common space is defined to account for the granularity of different modalities and noise contrastive estimation is employed to train the model.

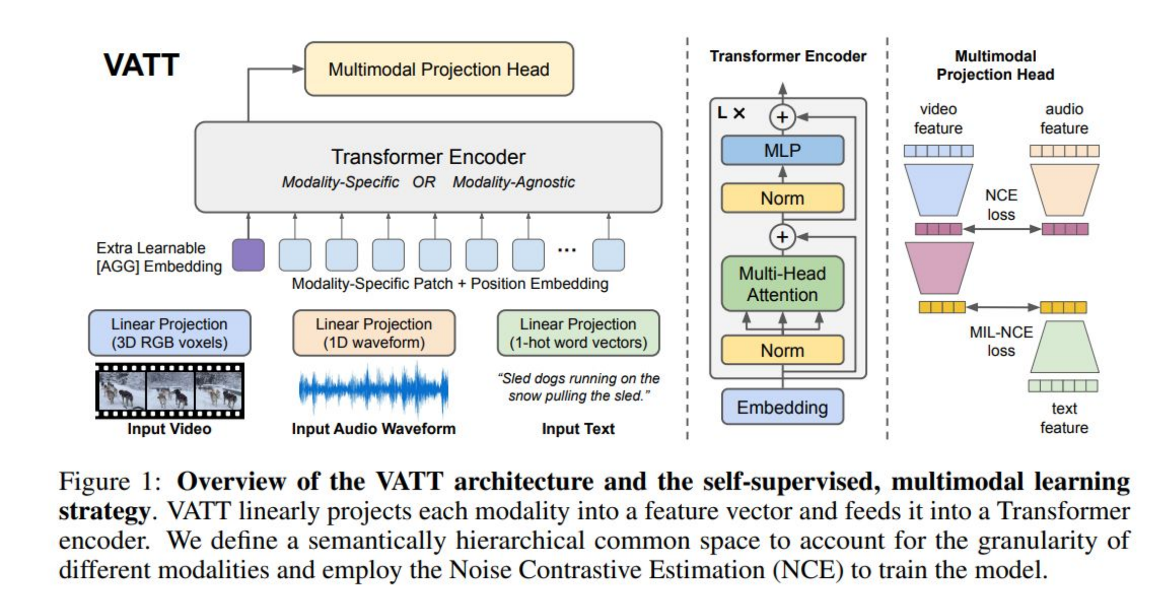

VATT
Multimodal Projection Head
Transformer Encoder
Modality-Specific OR Modality-Agnostic
Extra Learnable [AGG] Embedding
...
Modality-Specific Patch + Position Embedding
Linear Projection (3D RGB voxels)
Linear Projection (1D waveform)
Linear Projection (1-hot word vectors)
Input Video
Input Audio Waveform
"Sled dogs running on the snow pulling the sled."
Input Text
Transformer Encoder
L ×
+
MLP
Norm
+
Multi-Head Attention
Norm
Embedding
Multimodal Projection Head
video feature
audio feature
NCE loss
MIL-NCE loss
text feature

Figure 1: **Overview of the VATT architecture and the self-supervised, multimodal learning strategy**. VATT linearly projects each modality into a feature vector and feeds it into a Transformer encoder. We define a semantically hierarchical common space to account for the granularity of different modalities and employ the Noise Contrastive Estimation (NCE) to train the model.

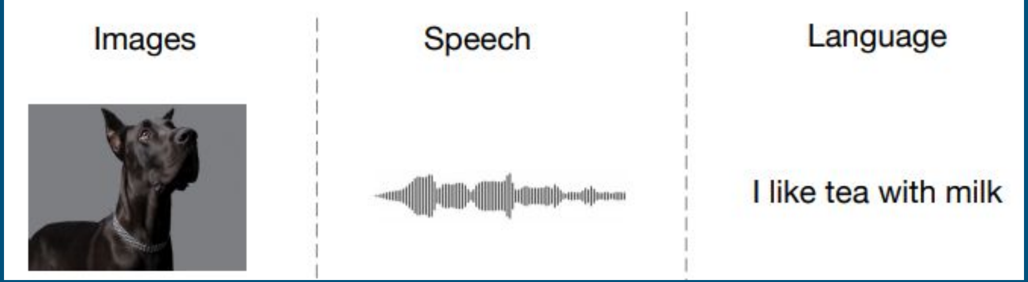
VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text
https://arxiv.org/abs/2104.11178

# Introducing Data2vec

Our method for images

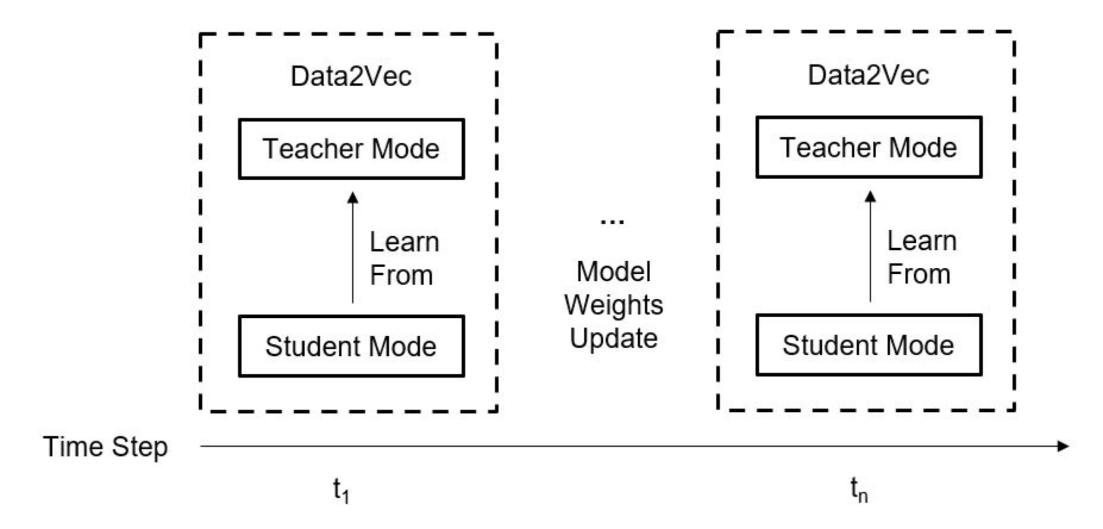
# Data2vec

- This paper hence proposes a unified framework called data2vec for SSL in three modalities:
- Images are 2D structured data
- Texts are discrete 1D data
- Speech is continuous 1D data
- data2vec does not perform multimodal training but aims to unify the learning objective for self-supervised learning in different modalities.

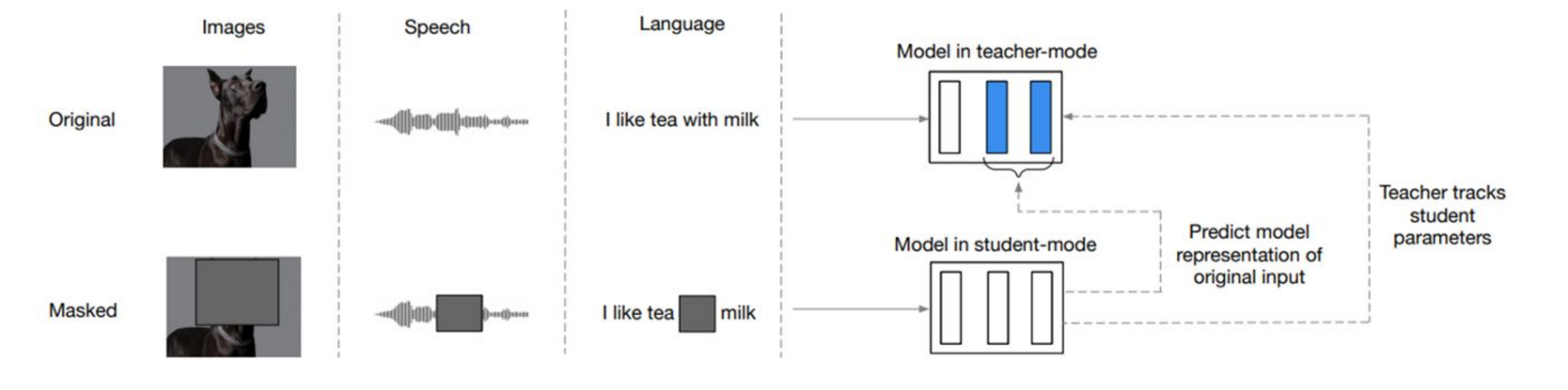


# Method Overview

- data2vec uses one model but has two modes: the teacher mode and the student mode
- In each time step, the student mode of data2vect will try to learn from the teacher mode and update the model parameters



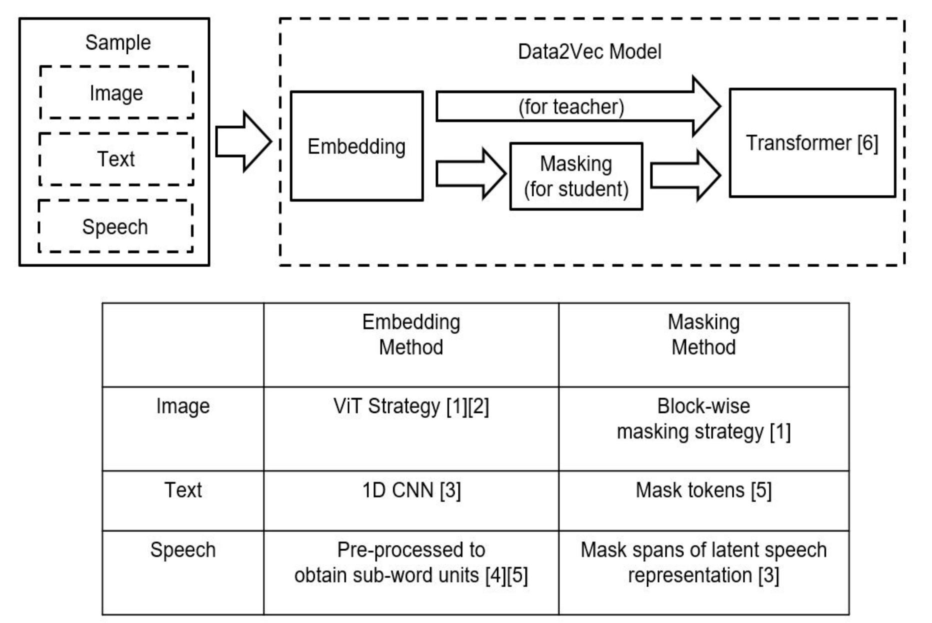
# Method



- 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.

# Architecture

- The model is trained to predict the model representations of the original unmasked training sample based on an encoding of the masked sample.
- They predict model representations only for time-steps which are masked. The representations they predict are contextualized representations, encoding the particular time-step but also other information from the sample due to the use of self-attention in the Transformer network.



Sample

Image

Text

Speech

Data2Vec Model

Embedding

(for teacher)

Masking<br>(for student)

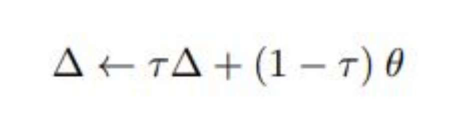
Transformer [6]

| | 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] |

# More details on model setup

- CV: embed 224x224 images into 16x16 patches then linearly transformed into 196 representations
  - 60% mask on randomly sampled adjacent sequences + augmentations (BEiT)
- Speech: fairseq implementation, 16 kHz waveform input into feature encoder (CON + NORM + GELU activation) outputs 50 Hz waveform
  - uniform sample (p = 0.065) time steps and mask subsequent 10 timesteps(49% of all time-steps MASK)
- NLP: RoBERTa-like implementation, byte-pair encoded input
  - BERT-like uniform sample 15% of tokens and replace vo/ a MASK token (80% MASK. 10% unchanged. 10% random vocab token)

# Teacher Parameterization, Targets



$$\Delta \leftarrow \tau \Delta + (1-\tau)\,\theta$$

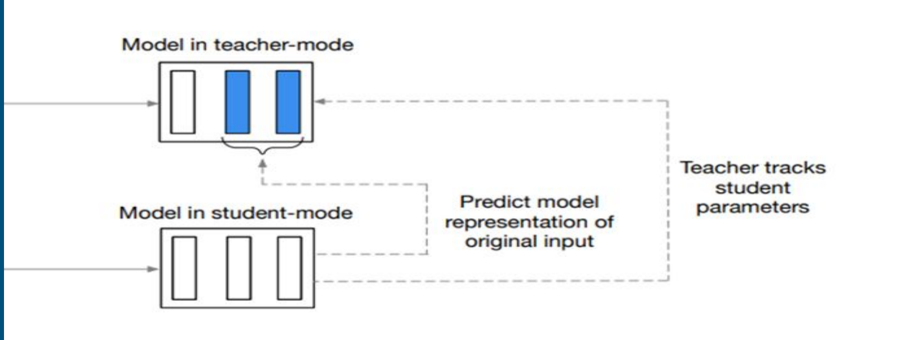
Δ :- model parameters
τ :- tau Learning rate

- Teacher parameterization
    - They use a schedule for τ that linearly increases this parameter from $\tau_0$ to the target value $\tau_e$ over the first $\tau_n$ updates after which the value is kept constant for the remainder of training.

# Teacher Parameterization, Targets

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

t :- current time steps
K :- top k output blocks
L :- Lth block
$a_t^l$ :- output of block l at time t



Model in teacher-mode

Model in student-mode

Predict model representation of original input

Teacher tracks student parameters

- Targets
  - Training targets are constructed based on the output of the top-K blocks of the teacher network for time-steps which are masked in student-mode.
  - Normalizing the targets helps prevent the model from collapsing into a constant representation for all time-steps and it also prevents layers with high norm to dominate the target features.

# Objective Function

- Smooth L1 loss to regress these targets.
- β controls the transition from a squared loss to an L1 loss.
- The advantage of this loss is that it is less sensitive to outliers, however, we need to tune the setting of β.

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

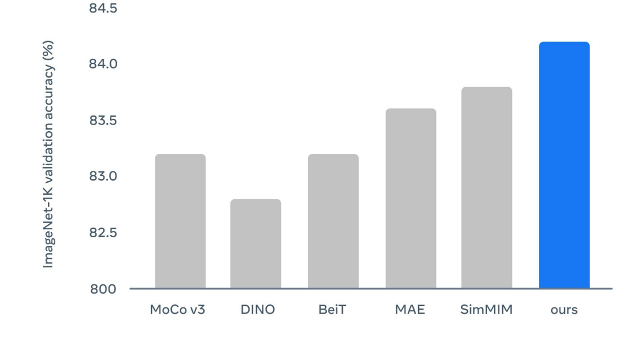
t :- time step
f(x) :- student prediction
β :- transition threshold (beta param)

# How is Data2vec different?

- Vs BERT
  - Predicts continuous and contextualized representations instead of linguistic tokens
- Vs BYOL
  - Used Masked prediction tasks
  - Regress multiple NN layers
- Vs Wav2vec2/HuBERT
  - Learns latent representation without quantization

# Results Computer Vision

- Downstream task of predicting single labels
- Metric: accuracy. Higher value, better performance



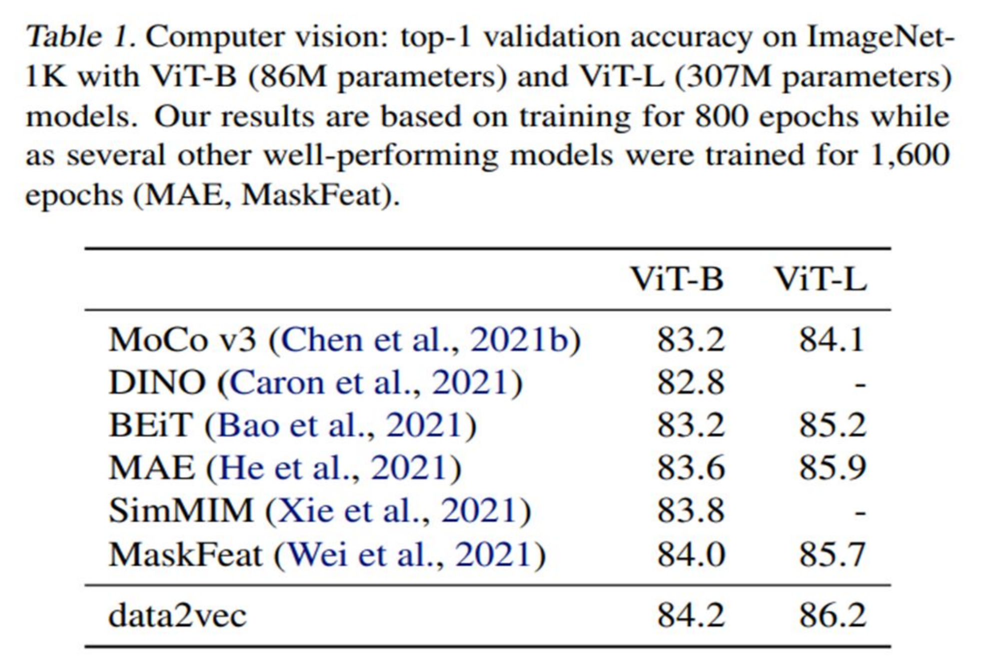
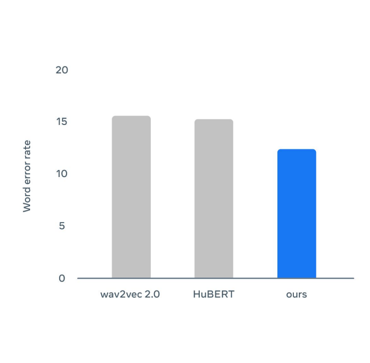


*Table 1.* Computer vision: top-1 validation accuracy on ImageNet-1K with ViT-B (86M parameters) and ViT-L (307M parameters) models. Our results are based on training for 800 epochs while as several other well-performing models were trained for 1,600 epochs (MAE, MaskFeat).

| | 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 |

# Results Speech Processing

- Lowest error rate for all amount of labeled data
- Metric: word error rate. Lower value, better performance



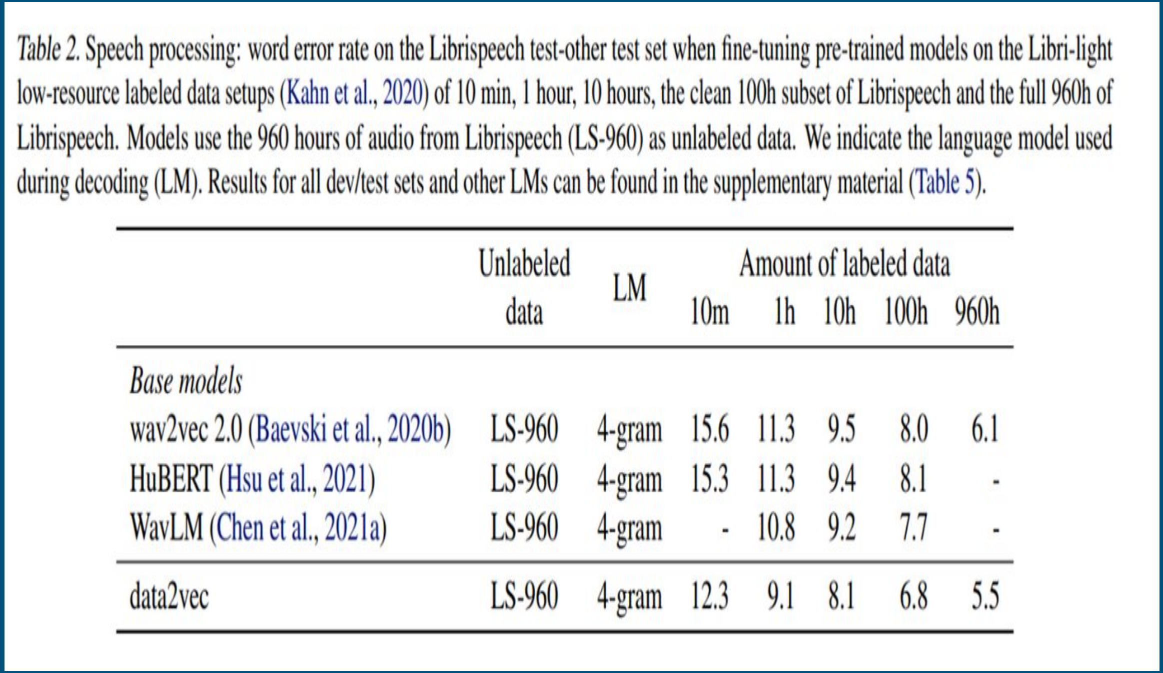


*Table 2.* Speech processing: word error rate on the Librispeech test-other test set when fine-tuning pre-trained models on the Libri-light low-resource labeled data setups (Kahn et al., 2020) of 10 min, 1 hour, 10 hours, the clean 100h subset of Librispeech and the full 960h of Librispeech. Models use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

# Results NLP

- NLP model outperforms RoBERTa Baseline
- Wav2vec2 masking improves accuracy
- metric: GLUE score. Higher value, better performance

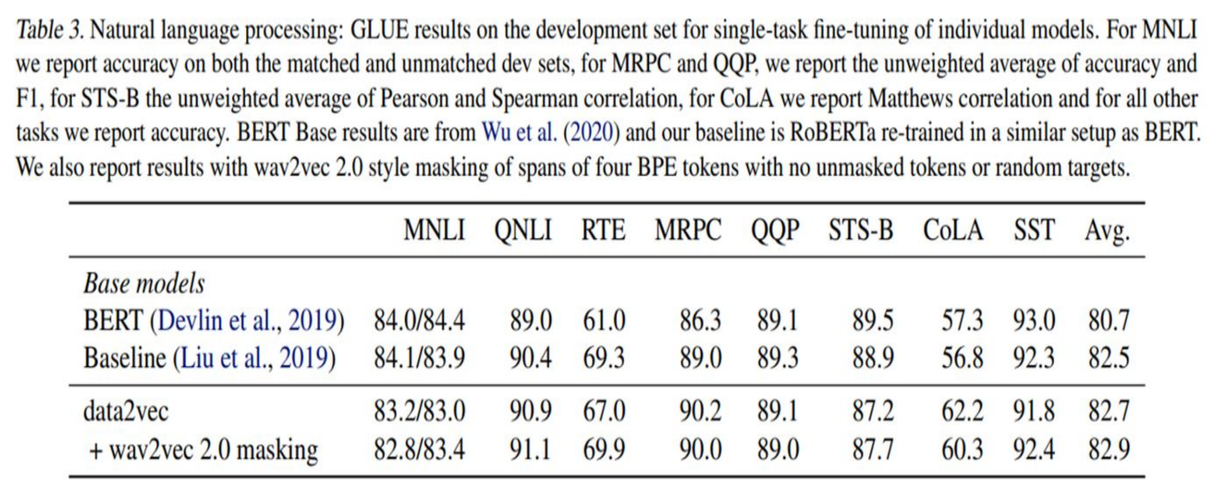
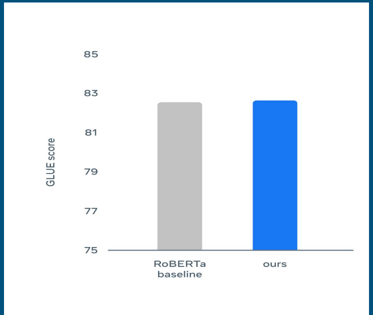


*Table 3.* Natural language processing: GLUE results on the development set for single-task fine-tuning of individual models. For MNLI we report accuracy on both the matched and unmatched dev sets, for MRPC and QQP, we report the unweighted average of accuracy and F1, for STS-B the unweighted average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

- 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
    - The results shown have better performance when the value is lower(speech), higher(NLP, i.e., texts) and higher(Vision, i.e., images ) respectively. The effect is more pronounced in speech and texts than in images.

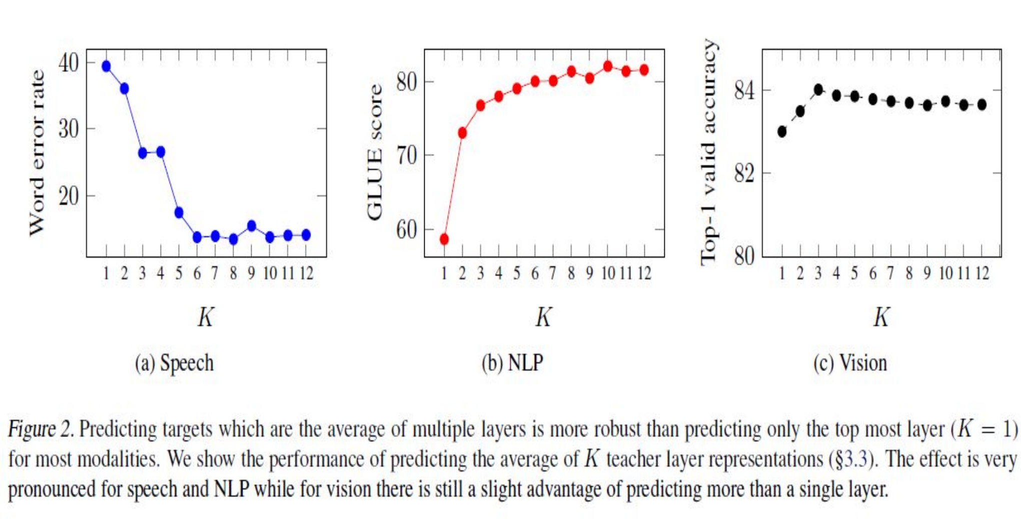


*Figure 2.* Predicting targets which are the average of multiple layers is more robust than predicting only the top most layer ($K = 1$) for most modalities. We show the performance of predicting the average of $K$ teacher layer representations (§3.3). The effect is very pronounced for speech and NLP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

# Ablation Study

- 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.

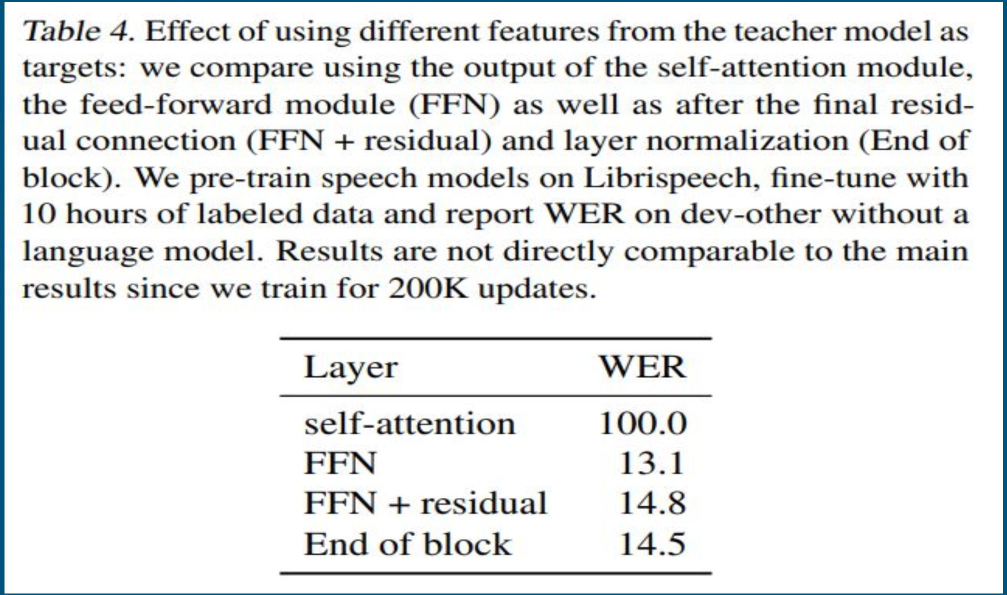


*Table 4.* Effect of using different features from the teacher model as targets: we compare using the output of the self-attention module, the feed-forward module (FFN) as well as after the final residual connection (FFN + residual) and layer normalization (End of block). We pre-train speech models on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

# Discussion

- Modality-specific feature extractors and masking.
    - Despite the unified learning regime, they still use modality-specific features extractors and masking strategies.
- Structured and contextualized targets.
    - For NLP, data2vec is the first work that does not rely on predefined target units.
- Representation collapse.
    - They found that collapse is most likely to happen in the following scenarios:
        - First, the learning rate is too large or the learning rate warmup is too short which can often be solved by tuning the respective hyperparameters.
        - Second, $\tau$ is too low which leads to student model collapse and is then propagated to the teacher.

# My thoughts

- Will this method work for unstructured modality, e.g. graphs
- And Will it be SOTA there too
- Can this be expanded to reach Meta AI’s vision