Data2Vec

Othman Mohamed Othman	- 2002395
Mahmoud Aboud Mohammad	- 2002387

Supervised By:
Prof Hazem Abbas

Introduction

Introduction

- General Framework for Self Supervised Learning.
- Same Architecture and Learning Algorithm for Speech, Vision, NLP.
- Contextualized Latent Representation of the Input.
- Based on Masked Version of the Input.

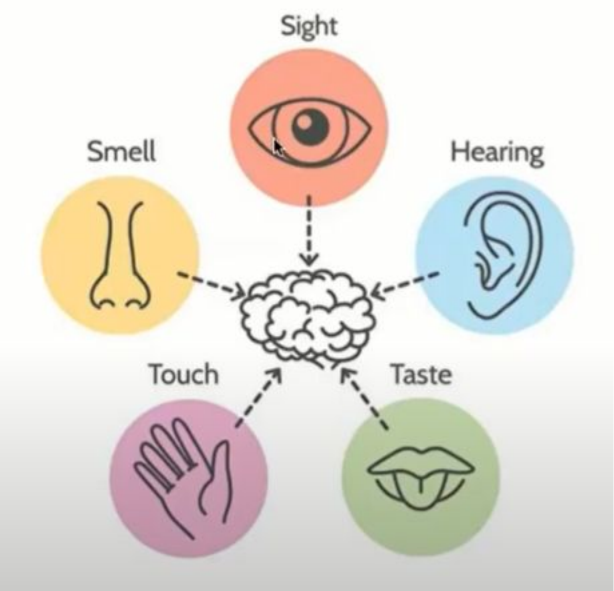
Motivation

- Self Supervised

Supervised	Unsupervised
Data + Labels	Data Only
Classification, Regression	Clustering, Dimensionality Reduction

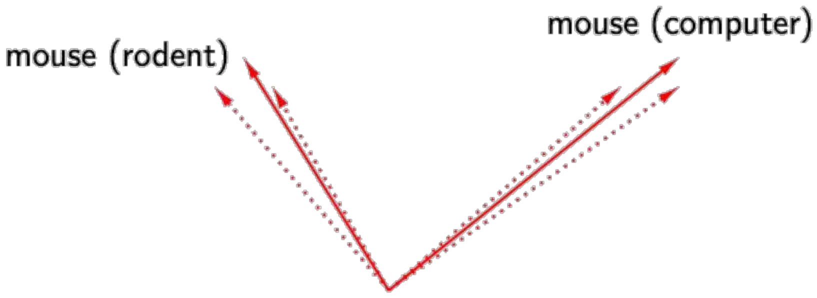
Motivation

- UniModal
  - Similar to the way human brain learns
  - Step towards general AI
  - Unified Algorithm and Architecture for different data types (Images, Audio, Text)



Motivation

- Contextualized
  - Latent Representation of the whole context.
  - Embedding Changes when context change

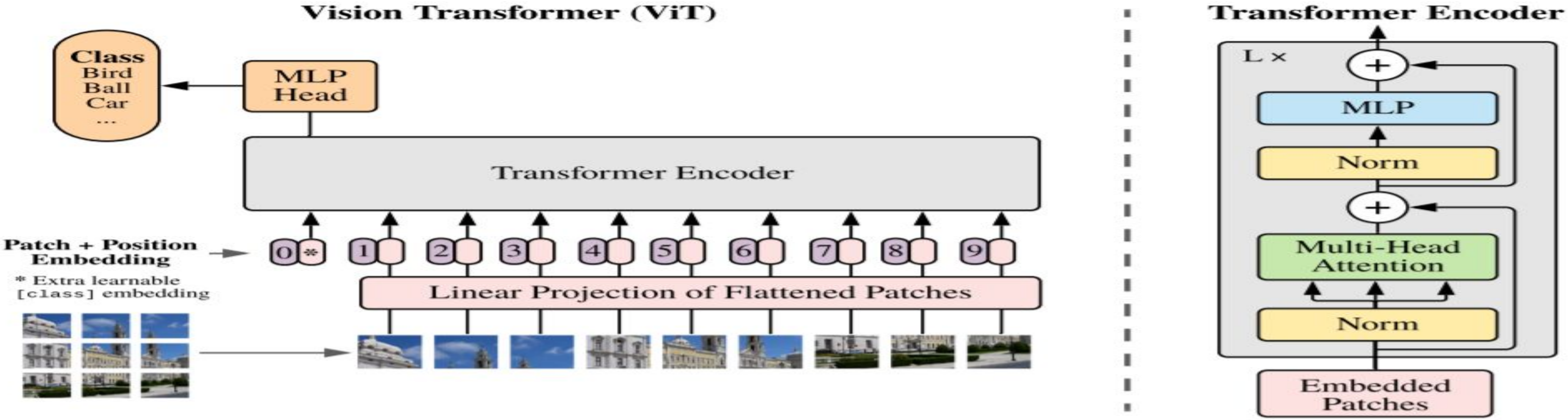


Related Work

Computer Vision

Vision Transformer (ViT)

The 2020 paper: “An Image is Worth 16x16 Pixels: Transformers for Image Recognition at Scale” by Dosovitskiy, A., et al (Google) introduced the Vision Transformer, which at first just seemed like a cool extension of NLP Transformers but which has now proved to be very effective for computer vision tasks.



Computer Vision

DINO

DINO, a new self-supervised system by Facebook AI, can learn incredible representations from unlabeled data. Below is a video visualizing its attention maps and we see the model was able to automatically learn class-specific features leading to accurate unsupervised object segmentation. It was introduced in their paper “Emerging Properties in Self-Supervised Vision Transformers”

Bootstrap Your Own Latent (BYOL)

BYOL uses two same encoder networks referred to as online and target network for obtaining representations and reducing the contrastive loss between the two representations.

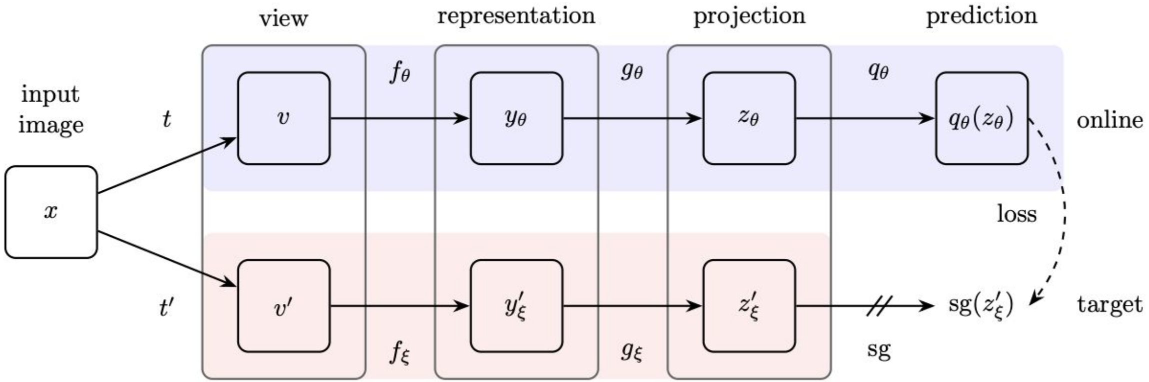
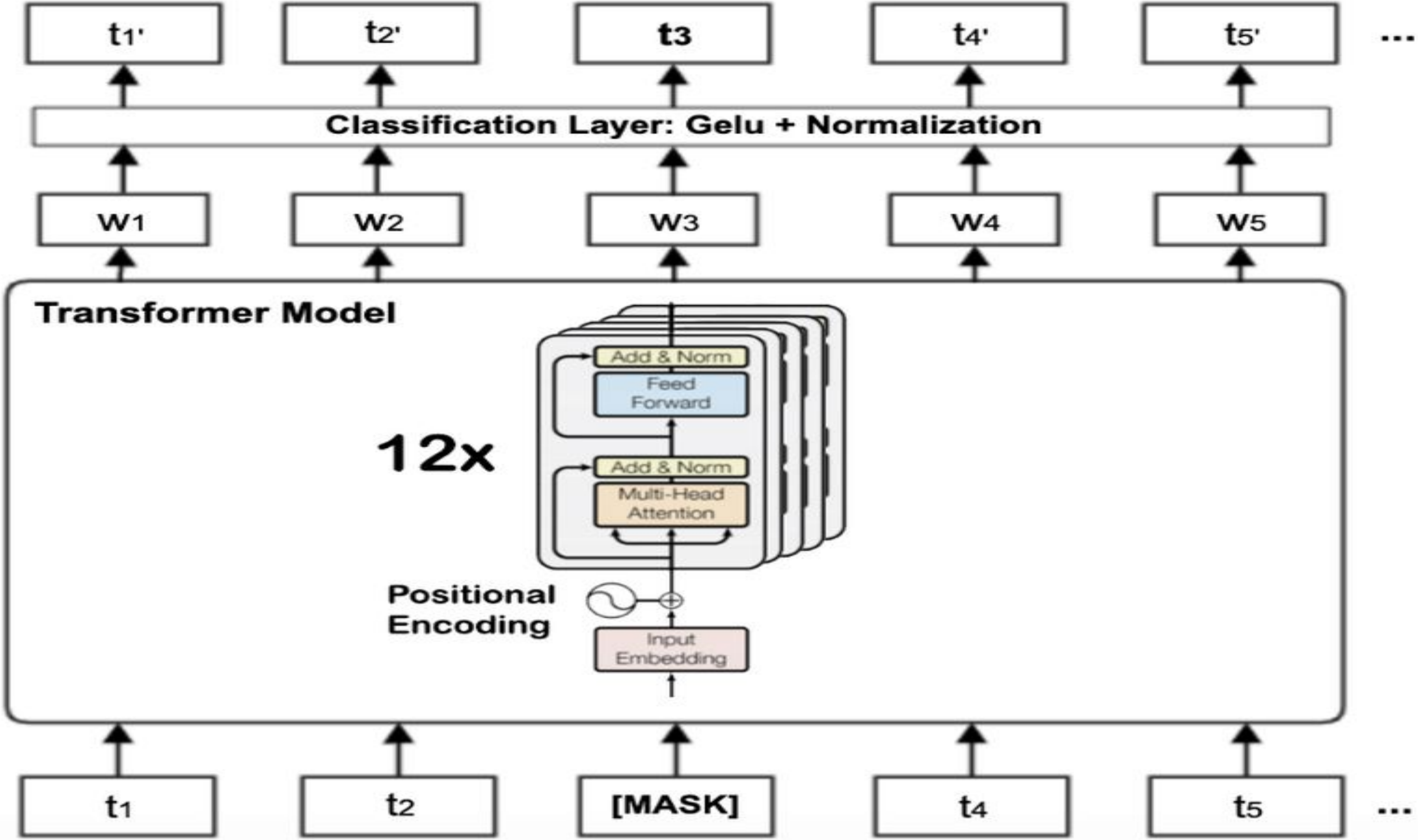


Figure 2: BYOL's architecture. BYOL minimizes a similarity loss between $q_\theta(z_\theta)$ and $\mathrm{sg}(z'_\xi)$, where θ are the trained weights, ξ are an exponential moving average of θ and sg means stop-gradient. At the end of training, everything but f_θ is discarded, and y_θ is used as the image representation.

Natural Language Processing

BERT

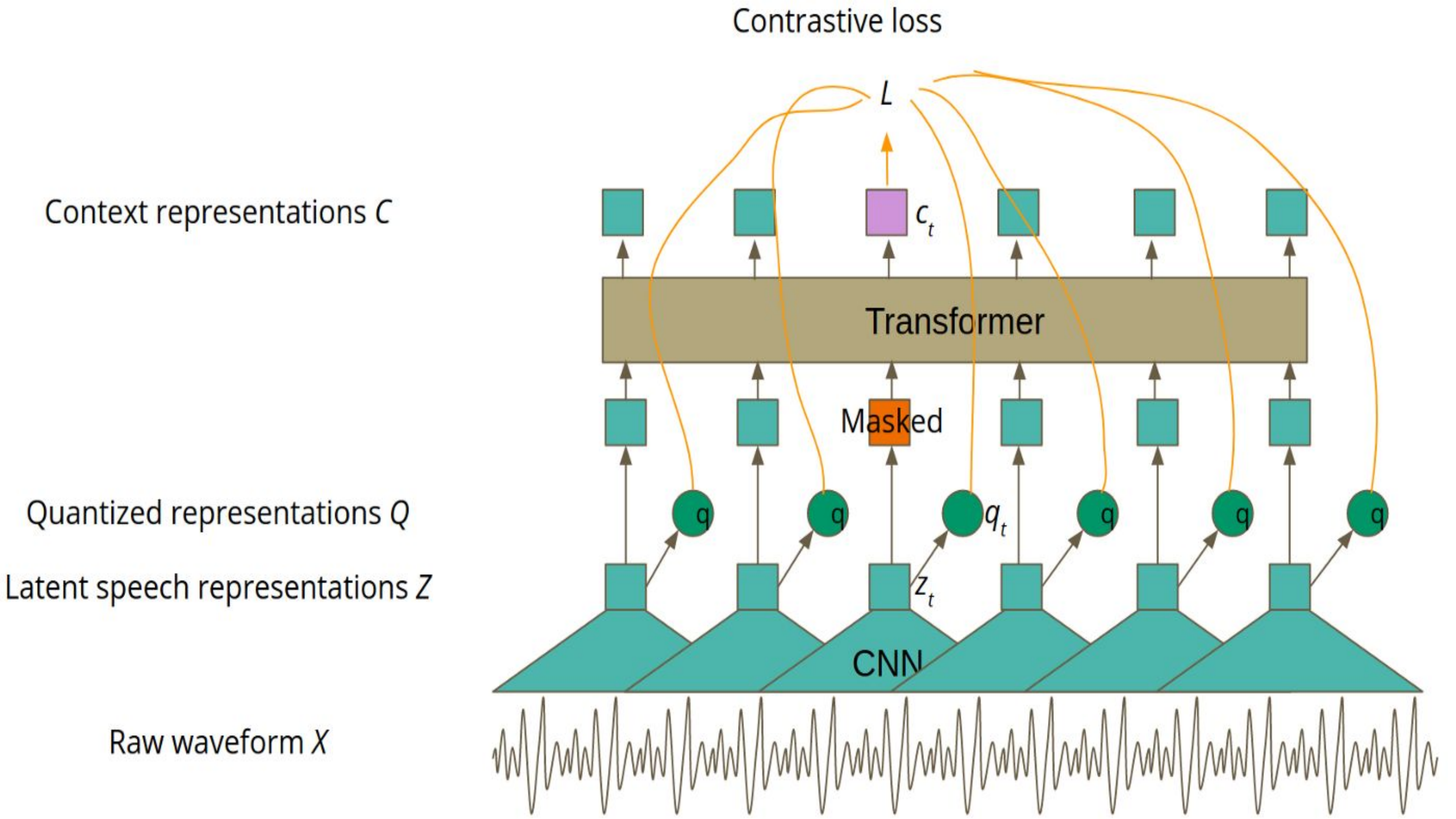
BERT, which stands for Bidirectional Encoder Representations from Transformers is an open-source machine learning framework for natural language processing (NLP). BERT is designed to help computers understand the meaning of ambiguous language in a text by using surrounding text to establish context. The BERT framework was pre-trained using text from Wikipedia and also can be fine-tuned with a question and answer datasets


t1'
t2'
t3
t4'
t5'
...
Classification Layer: Gelu + Normalization
W1
W2
W3
W4
W5
Transformer Model
12x
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Positional Encoding
Input Embedding
t1
t2
[MASK]
t4
t5
...

Speech

Wav2vec

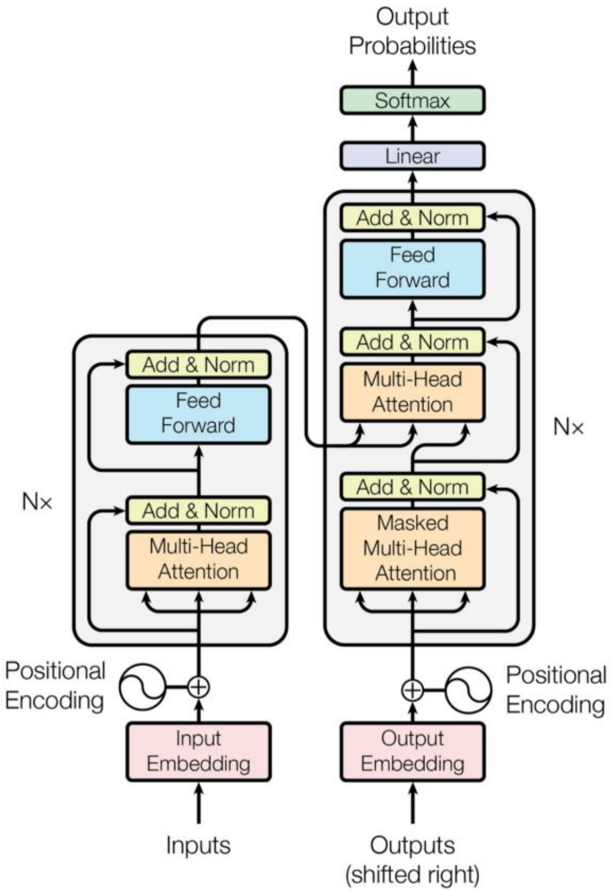
Wav2Vec 2.0 is one of the current state-of-the-art models for Automatic Speech Recognition due to self-supervised training which is quite a new concept in this field. This way of training allows us to pre-train a model on unlabeled data which is always more accessible. Then, the model can be fine-tuned on a particular dataset for a specific purpose. As the previous works show this way of training is very powerful


Contrastive loss
L
Context representations C
c_t
Transformer
Masked
Quantized representations Q
q
q_t
Latent speech representations Z
z_t
CNN
Raw waveform X

Method

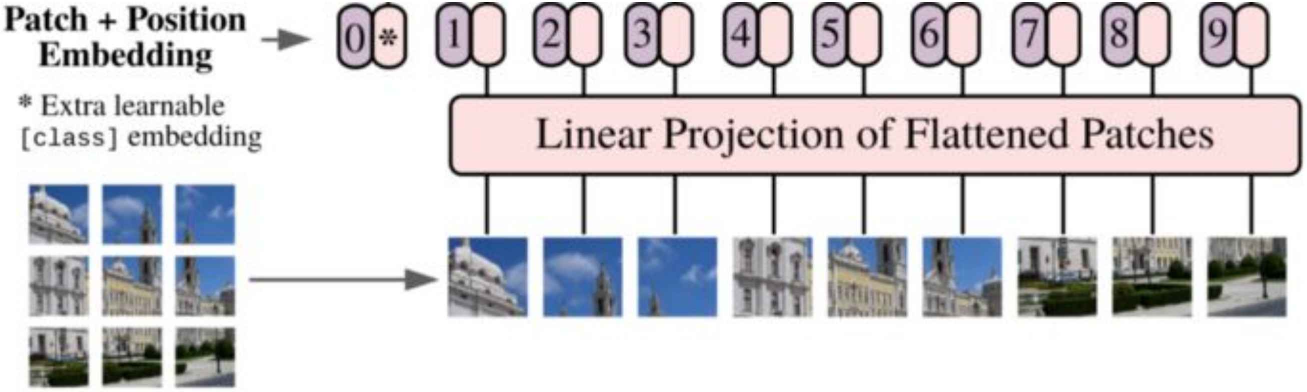
Architecture

- Standard Transformer Architecture.
- Transformers have been used in previous work with Audio, Images and Text.
- Modality Specific Encoding (Input Embedding)



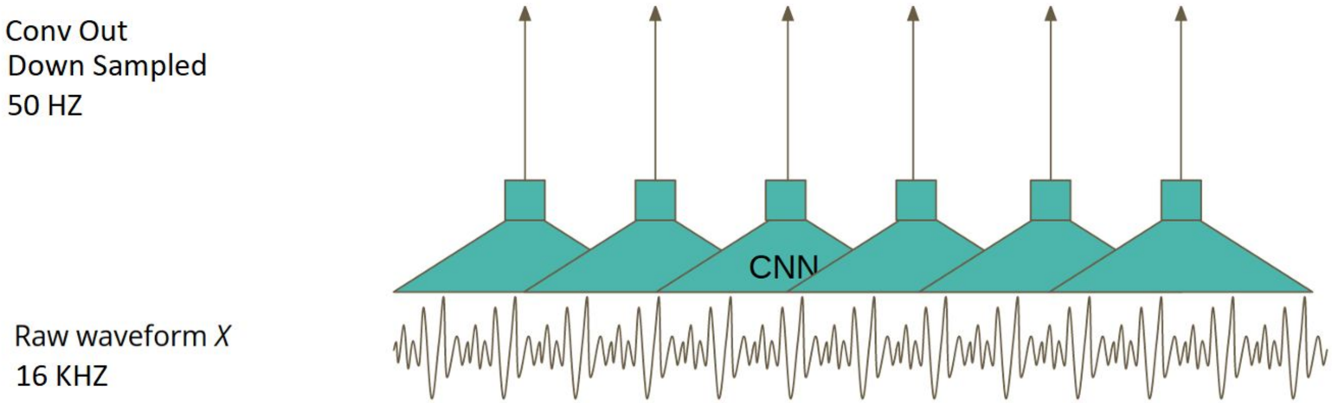
Input Preparation (Computer Vision)

- Vision Transformer (ViT) Encoding Strategy.
- Image converted to sequence of batches, each of size 16x16 pixels.
- batches are then input to a linear transformation to produce the final encoding.


Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

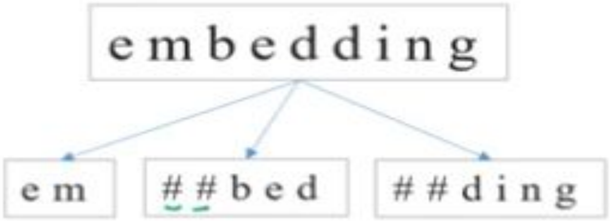
Input Preparation (Speech)

- Wav2Vec 2.0 Encoding Technique
- Multi-Layer 1-D Convolutions
- Down sampling 16kHz waveform to 50 Hz samples



Input Preparation (Text)

- RoBERTa Embedding Technique (Robust Implementation of BERT)
- Sub-Words are obtained from words
- Sub Words are embedded to distributional space via learned embedding vectors



Training Targets

Data2vec is training models to predict their own representations of the input data, regardless of the modality. By focusing on these representations — the layers of a neural network — instead of predicting visual tokens, words, or sounds, a single algorithm can work with completely different types of input. This removes the dependence on modality-specific targets in the learning task. Directly predicting representations is not straightforward, and it required defining a robust normalization of the features for the task that would be reliable in different modalities.

Our method for images

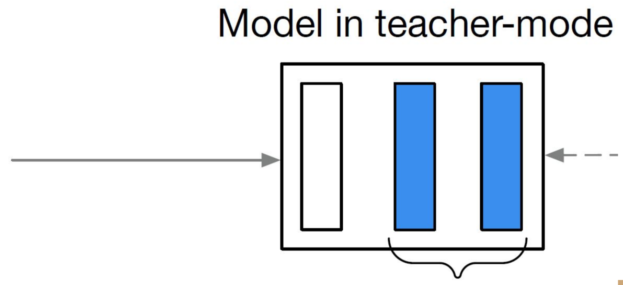
Training Targets

Teacher parameterization

They use a schedule for T that linearly increases this parameter from T0 to the target value Te

over the first Tn updates after which the value is kept constant for the remainder of the training.

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



Training Targets

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. (speech: instance norm, CV & NLP: parameter-less layer norm)

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

Objective (Loss)

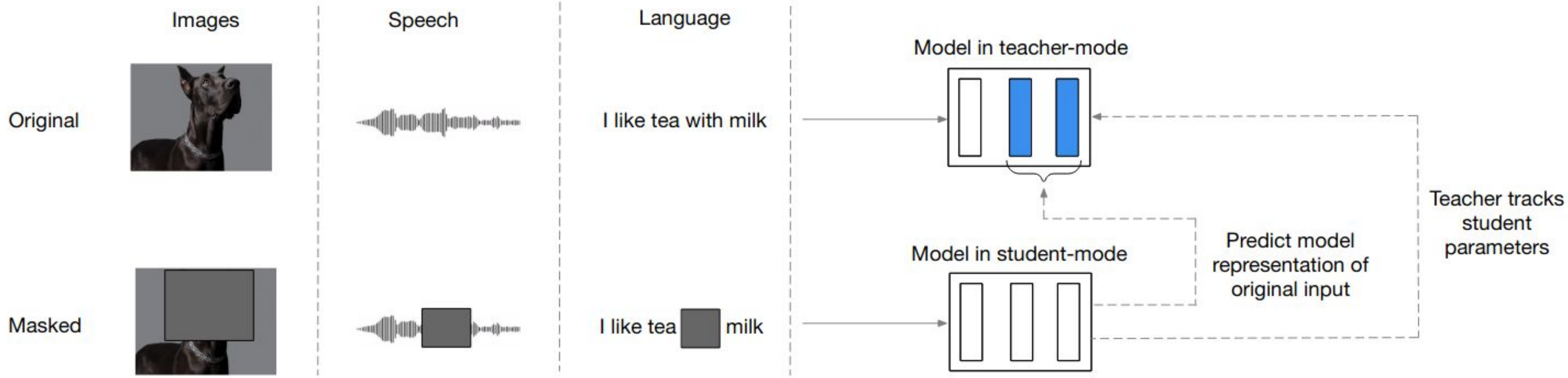
- Smooth L1 loss to regress these targets.

- 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.

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

Experimental Setup


Images
Speech
Language
Original
I like tea with milk
Masked
I like tea
milk
Model in teacher-mode
Model in student-mode
Predict model representation of original input
Teacher tracks student parameters

Computer Vision

- Images of 224x224 pixels are embedded as patches of 16x16 pixels like the normal ViT.
- Augmentation by: randomly applied resized image crops, horizontal flipping, and color jittering
- 60% of the image is masked instead of 40% as in ViT.
- The same modified image is used both in teacher mode and student mode.

Computer Vision

Hyper-parameter	ViT-B	ViT-L
Epochs	800	1600
Batch size	2048	8192
Warm-up Epochs	40	80
Learning Rate Value	0.002	0.001
Learning Rate Scheduling	Learning rate is annealed following the cosine schedule	
optimizer	Adam	Adam
β	2	2
K	6	6

Speech

- Feature Encoder: input 16 KHz and output of 50 Hz with the help of 7 convolutional layers.
- Sample with p = 0.065 of all time-steps to be starting indices and mask the subsequent ten time-steps. This results in approximately 49% of all time-steps to be masked.
- Adam optimizer is used, with a peak learning rate of 5x10-4 for data2vec Base.
- Tri-stage scheduler is used: linearly warms up the learning rate for first 3% of updates, holds it for 90% of updates, then linearly decays over the remaining 7%.

Natural Language Processing

- They build on the BERT re-implementation RoBERTa.
- The input data is tokenized using a byte-pair encoding of 50K types and the model learns an embedding for each type.
- Once the data is embedded, BERT masking strategy is applied to 15% of uniformly selected tokens: 80% are replaced by a learned mask token, 10% are left unchanged and 10% are replaced by randomly selected vocabulary tokens.
- They also consider the wav2vec 2.0 strategy of masking spans of four tokens.

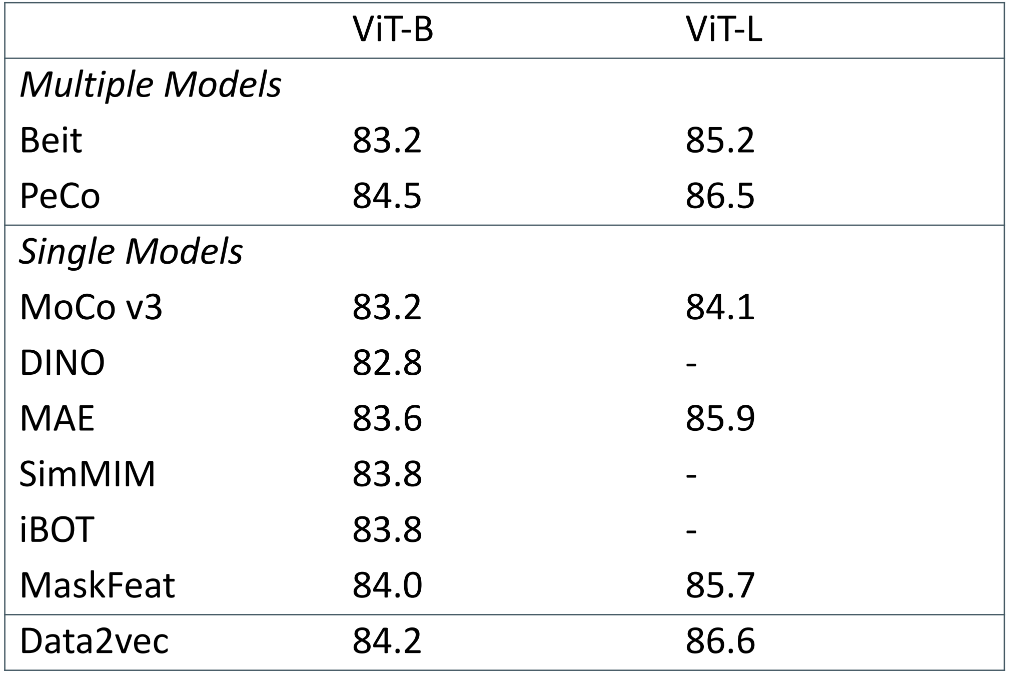
Results

Results

- Experiments on the major benchmarks of speech recognition, image classification, and natural language understanding show that the proposed framework outperforms or at least is comparable to the previous state-of-the-art in all the domains.

Results (Computer Vision)

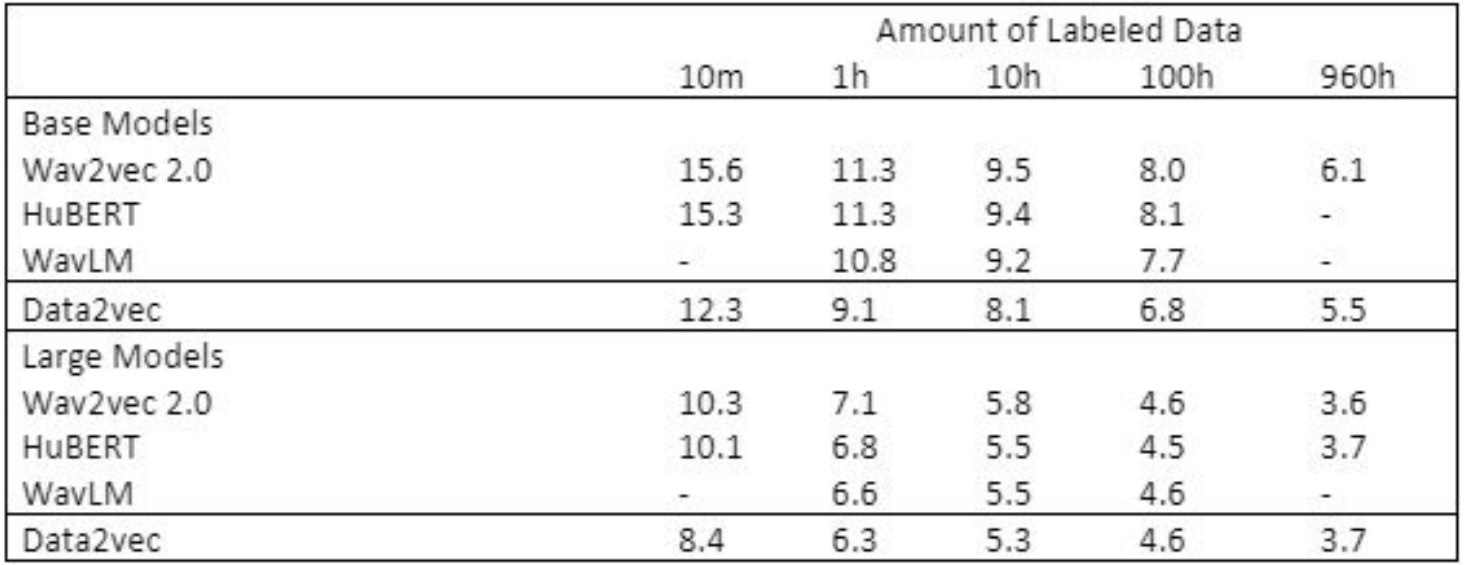
- Trained on ImageNet1K Training Set
- Fine Tuned for image classification using labeled data from same dataset
- top-1 accuracy of the classification task is calculated
- It is distinguished between the results based on a single self-supervised model, and results that train a separate visual tokenizer or distill other self-supervised models.



	ViT-B	ViT-L
Multiple Models		
Beit	83.2	85.2
PeCo	84.5	86.5
Single Models		
MoCo v3	83.2	84.1
DINO	82.8	-
MAE	83.6	85.9
SimMIM	83.8	-
iBOT	83.8	-
MaskFeat	84.0	85.7
Data2vec	84.2	86.6

Results (Speech Recognition)

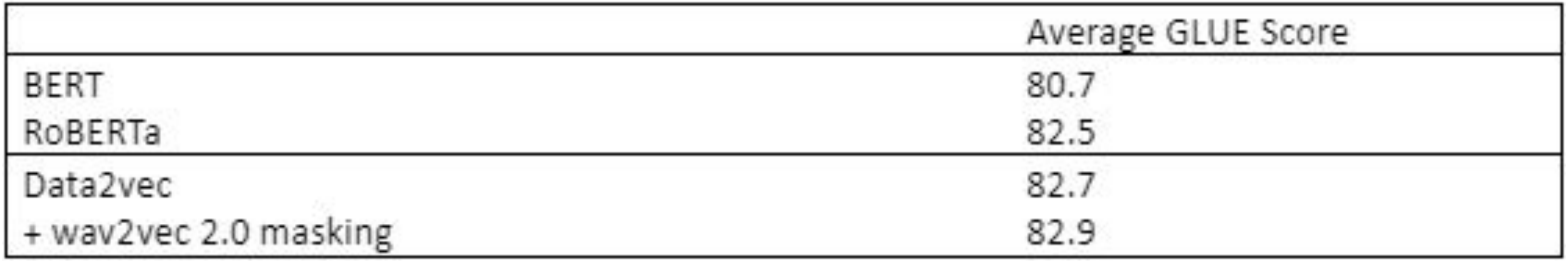
- Pretrained on 960 hours unlabeled data (LibriSpeech Dataset)
- Fine Tuned on different amount of labeled data
- The table shows Word Error Rate (WER) for different amount of labeled data
- Biggest improvement when amount of labeled data is low



	Amount of Labeled Data				
	10m	1h	10h	100h	960h
Base Models					
Wav2vec 2.0	15.6	11.3	9.5	8.0	6.1
HuBERT	15.3	11.3	9.4	8.1	-
WavLM	-	10.8	9.2	7.7	-
Data2vec	12.3	9.1	8.1	6.8	5.5
Large Models					
Wav2vec 2.0	10.3	7.1	5.8	4.6	3.6
HuBERT	10.1	6.8	5.5	4.5	3.7
WavLM	-	6.6	5.5	4.6	-
Data2vec	8.4	6.3	5.3	4.6	3.7

Results (NLP)

- PreTrained on the Book Corpus and English Wikipedia (Same as BERT)
- General Language Understanding Evaluation (GLUE) benchmark.
- GLUE includes tasks for natural language inference, sentence similarity, grammatical analysis, and sentiment analysis.
- Fine Tuning for each task on labeled data.
- data2vec outperforms BERT and RoBERTa for the average GLUE score when span of 4 tokens is masked with a masking probability of 0.35 (wav2vec style masking).

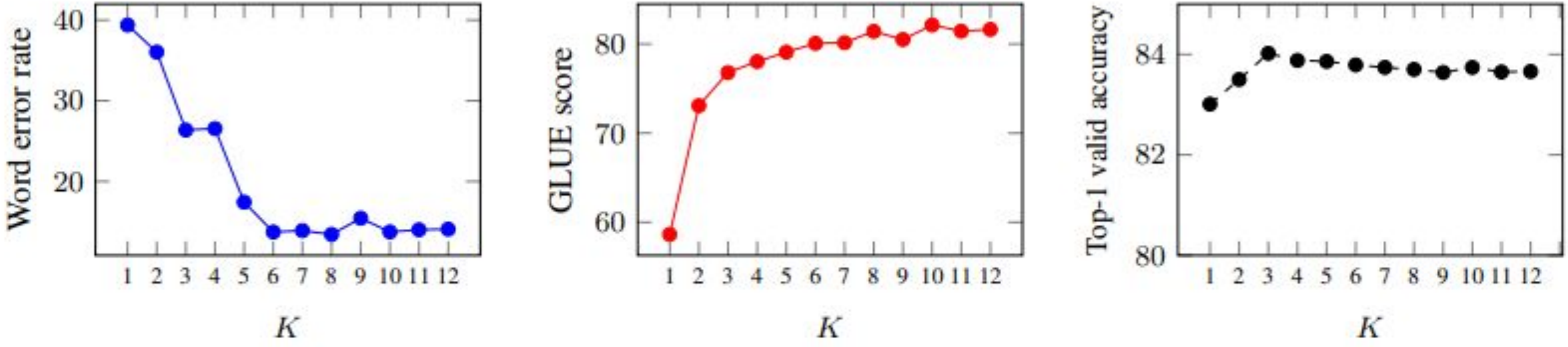


	Average GLUE Score
BERT	80.7
RoBERTa	82.5
Data2vec	82.7
+ wav2vec 2.0 masking	82.9

Ablation

K Averaging Effect

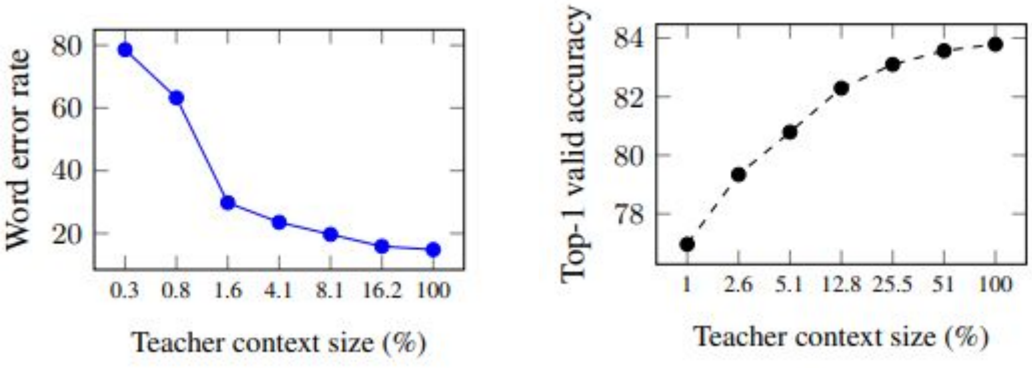
- Experiments on wav2vec have shown the top layer of network doesn’t perform as well as the middle layers.
- Targets based on multiple layers improve over using only the top layer (K = 1)
- Using all layers is generally a good choice and only slightly worse than a carefully tuned value of K.



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

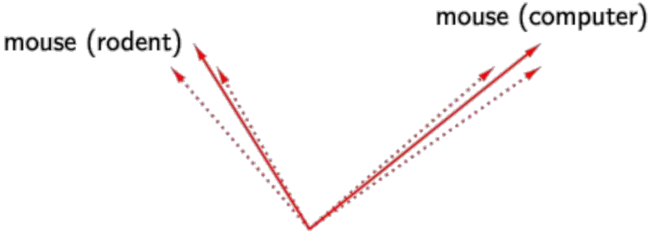
Target Contextualization

- Fully contextualized output which means that for every sample the output carries information about the whole sequence.
- This is different from most of the previous work where the Teacher has access to only sub part of the context.



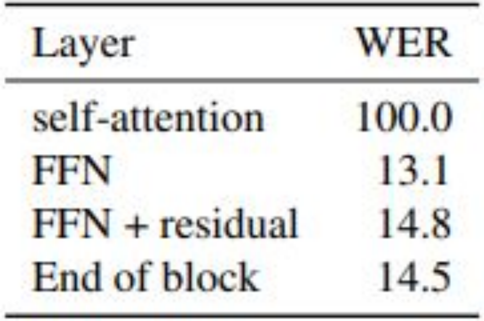
(a) Speech

(b) Vision

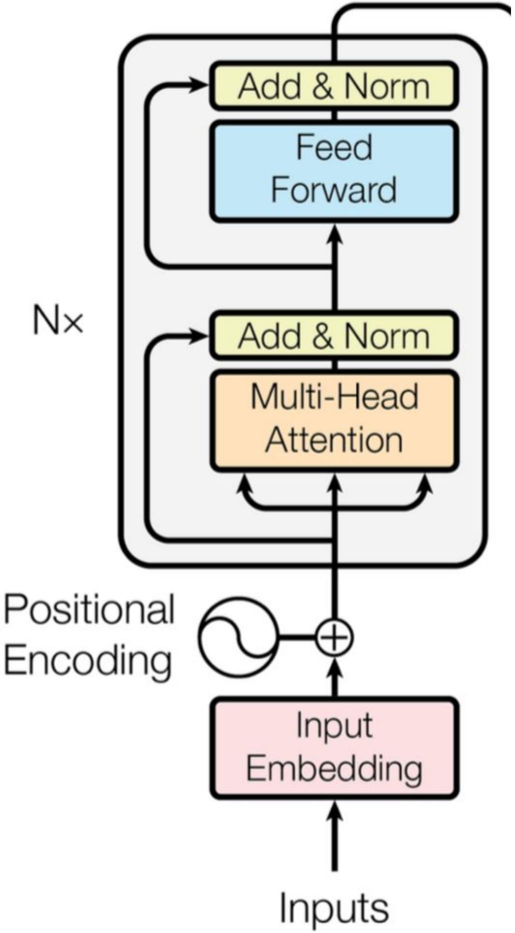


Target Feature Type

- The authors try different features from different layers of the transformer block.
- It is shown that the output of the feedforward block works the best for the latent representation.
- This is shown by fine-tuning on speech recognition task and calculating the Word Error Rate.



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



Discussion

Discussion

Modality-specific feature extractors and masking

- Despite the unified learning regime, they still use modality-specific feature extractors and masking strategies.
- This makes sense given the vastly different nature of the input data.

Representation collapse.

- This occurs when the model produces very similar representations for all masked segments.

Discussion

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, is too low which leads to student model collapse and is then propagated to the teacher.
- Third, we found collapse to be more likely for modalities where adjacent targets are very correlated and where longer spans need to be masked, e.g., speech. We address this by promoting variance through normalizing target representations over the sequence or batch.

Conclusion

Strengths and Limitations

Strengths

- Unified Architecture and Training Algorithm for different modalities.
- Strong Contextualized Latent Representation, Achieving State of the Art in various downstream tasks.
- Self Supervised Training, which means they need for labeled data is less than supervised approaches which decrease resources cost.

Limitations

- Modality Specific encoding is needed for each model.
- The models are based on transformer architecture which means model sizes are large and require huge amount of data to pretrain.

Training vs Inference

- data2vec follows the teacher-student procedure for training. That means that there are 2 networks involved in the training process.
- This leads to the need for larger memory specs while training to be able to load the 2 networks.
- However, at inference time or test time, the teacher network is not needed. The student network will be used alone to produce the latent representations.

- During Training, the Percentage of the input samples input to the student network is masked.
- During Inference, This masking procedure is dropped. The full input sequence is fed to the student network.

Conclusion

- Self Supervised Approaches have greatly improved the ability of Artificially Intelligent models to learn about the real world without the need for much-labeled data.
- Self Supervised Learning has been widely used to produce powerful contextualized latent representations of data that can be used in downstream tasks.
- Data2vec Shows that a single self-supervised learning regime can be effective for vision, speech, and language.
- A single learning method for multiple modalities will make it easier to learn across modalities and future work may investigate tasks such as audio-visual speech recognition or cross-modal retrieval.