Data2Vec

...

https://arxiv.org/pdf/2202.03555.pdf

Compiled by - Vaibhav

Abstract

data2vec, a framework that uses the same learning method for either speech, NLP or computer vision.

The core idea is to predict latent representations of the full input data based on a masked view of the input in a self distillation setup using a standard Transformer architecture.

Abstract

data2vec, a framework that uses the same learning method for either speech, NLP or computer vision.

The core idea is to predict latent representations of the full input data based on a masked view of the input in a self distillation setup using a standard Transformer architecture.

Instead of predicting modality-specific targets such as words, visual tokens or units of human speech which are local in nature, data2vec predicts **contextualized latent representations** that contain information from the entire input.

Abstract

data2vec unifies the learning algorithm but still learns representations ***individually*** for each modality.

The method combines masked prediction with the learning of latent target representations but generalizes the latter by using multiple network layers as targets.

Abstract

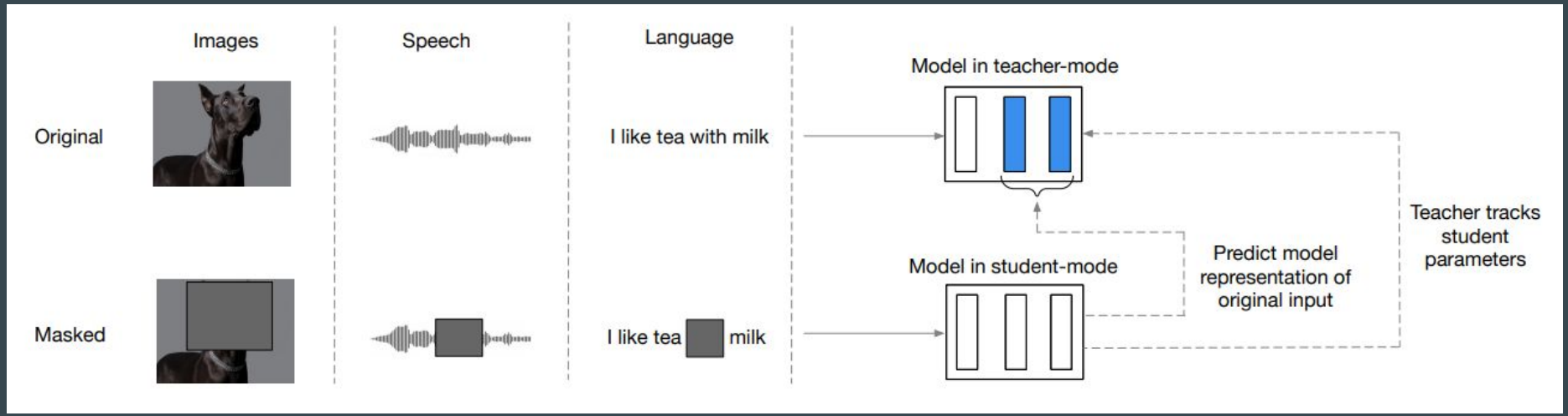
data2vec unifies the learning algorithm but still learns representations ***individually*** for each modality.

The method combines masked prediction with the learning of latent target representations but generalizes the latter by using multiple network layers as targets.

1. build representations of the full input data whose purpose is to serve as targets in the learning task (teacher mode)
2. encode a masked version of the input sample with which we predict the full data representations (student mode)

The weights of the teacher are an exponentially decaying average of the student

Say.. what?



Since different modalities have vastly different inputs, e.g., pixels vs. words, we use modality-specific feature encoders and masking strategies from the literature.

But, what does that mean for speech?

In comparison to wav2vec 2.0, data2vec directly predicts contextualized latent representations without quantization.

HuBERT discretizes representations from different layers across iterations and predicts these discretized units whereas data2vec predicts the average over multiple layers.

Alright.. let's talk architecture

Standard Transformer architecture with a modality-specific encoding of the input data.

CV: ViT-strategy of encoding an image as a sequence of patches, each spanning 16x16 pixels, input to a linear transformation is used.

Speech: data is encoded using a multi-layer 1-D convolutional neural network that maps 16 kHz waveform to 50 Hz representations.

Text: is pre-processed to obtain sub-word units, which are then embedded in distributional space via learned embedding vectors.

Alright.. let's talk architecture

Masking

After the input sample has been embedded as a sequence of tokens, we mask part of these units by replacing them with a learned MASK embedding token and feed the sequence to the Transformer network.

Alright.. let's talk architecture

Training targets

The model is trained to predict the model representations of the original unmasked training sample based on an encoding of the masked sample. We predict model representations only for time-steps which are masked.

The representations we predict are contextualized representations, encoding the particular time-step but also other information from the sample due to the use of self-attention.

Alright.. let's talk architecture

Objective function

Given contextualized training targets yt, they use a Smooth L1 loss to regress these targets:

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

where β controls the transition from a squared loss to an L1 loss, depending on the size of the gap between the target yt and the model prediction ft(x) at time-step t.

Specifically, speech

Feature encoder is same as Wav2Vec2.0 - takes 16 kHz waveform.

Masking strategy is also the same as Wav2Vec2.0 - approximately 49% of all time-steps to be masked for a typical training sequence.

Fine-tuning strategy is also the same as Wav2Vec2.0

Results

	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
Large models							
wav2vec 2.0 (Baevski et al., 2020b)	LS-960	4-gram	10.3	7.1	5.8	4.6	3.6
HuBERT (Hsu et al., 2021)	LS-960	4-gram	10.1	6.8	5.5	4.5	3.7
WavLM (Chen et al., 2021a)	LS-960	4-gram	-	6.6	5.5	4.6	-
data2vec	LS-960	4-gram	8.4	6.3	5.3	4.6	3.7

Conclusion

The authors show that a single self-supervised learning regime can be effective for vision, speech and language. The key idea is to regress contextualized latent representations based on a partial view of the input.

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.

Future work may investigate a single masking strategy that is modality-agnostic as well as jointly training multiple modalities.

Try Data2Vec models on Hugging Face

-> All the models - Hugging Face Data2Vec models

-> Test out the ASR performance on the base model via hosted inference or code - https://huggingface.co/facebook/data2vec-audio-base-960h