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

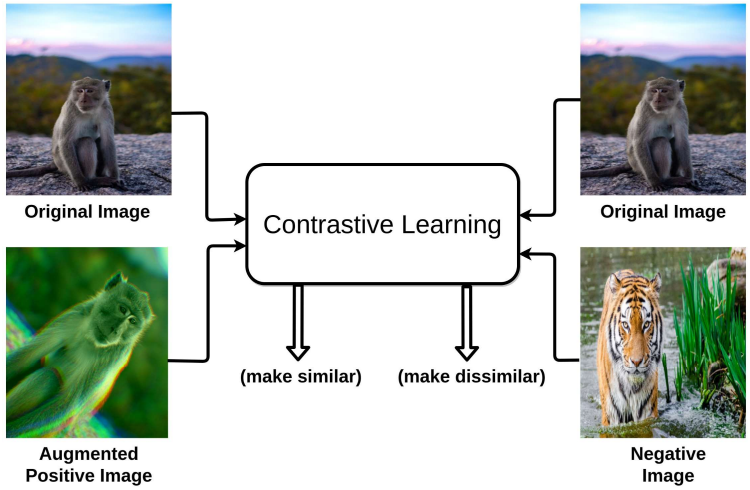
Presented by: Van Thao NGUYEN

Presentation Outline

1. Problem
   - Challenges that data2vec solves
2. Data2vec
   - Training methods
3. Experiments & Results
4. Questions

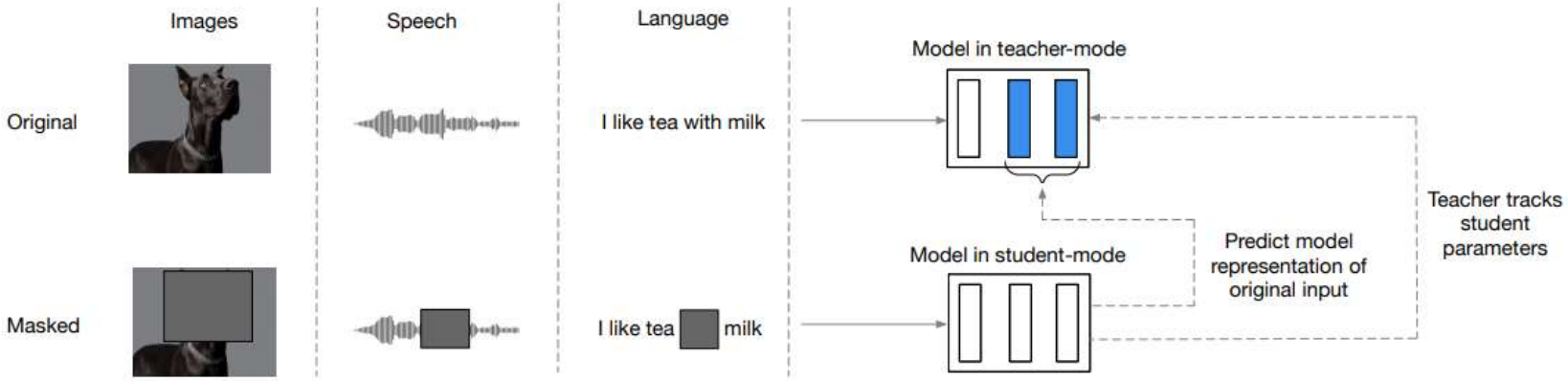
Self-supervised learning(SSL) lets machines learn from data itself, not labels

- Supervised learning is not scalable.
- The task defines a proxy loss, and the network is forced to learn what we really care about.
- Research on SSL today almost always focuses on a particular modality.
- Drawbacks of SSL is that each training process needs a unique algorithm required to that specific domain



An introduction to data2vec

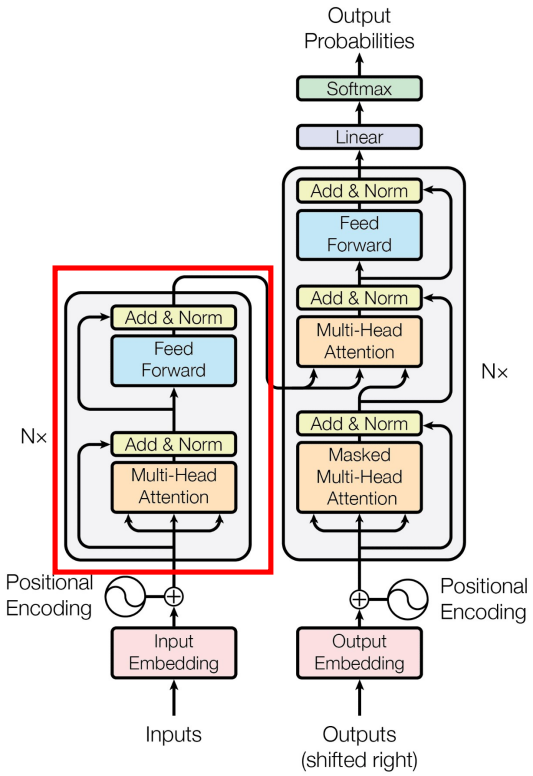
- A general framework that combines:
  - Masked prediction
  - Latent target representation learning
  - Two Transformer networks (**self-distillation**)



- The weights of the teacher are an exponentially moving average of the student: $\Delta \leftarrow \tau\Delta + (1-\tau)\,\theta$

Training methods

Model Architecture is a standard Transformer NN



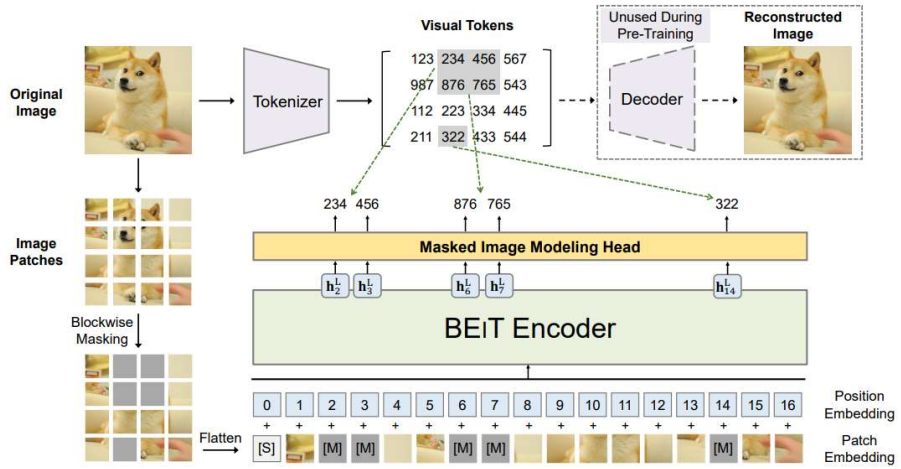
Standard Transformer architecture:

- CV: the ViT-strategy of encoding an image as a sequence of patches, each spanning 16x16 pixels, input to a linear transformation
- Speech: data is encoded using a multi-layer 1D CNN that maps 16kHz 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.

Training strategies on different modalities

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

- CV: embed 224x224 images as patches of 16x16 pixels then linearly transformed into sequence of 196 representations.
  - Follow the masking strategies of BEiT but mask 60% instead of 40% + data augementation.



Training strategies on different modalities

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

- CV: embed 224x224 images as patches of 16x16 pixels then linearly transformed into sequence of 196 representations.
  - Follow the masking strategies of BEiT but mask 60% instead of 40% + data augementation.
- Speech: 16kHz waveform input into feature encoder which results an encoder output frequency of 50 Hz waveform
  - Sample p=0,065 of all time-steps and mask subsequent 10 timesteps (49% of all time-step MASK)
- NLP: input data is tokenized by byte-pair encoding
  - applied to a 15% of the tokens, where 80% are replaced by a mask token, 10% are left unchanged, and 10% are replaced by randomly selected vocabulary tokens.

- **Training Targets:** The representations to predict are *contextualized* representations.
  - This is an important difference to BERT, wav2vec 2.0 or BEiT, MAE, SimMIM, and MaskFeat, which predict targets lacking contextual information.

- **Objective:** Minimizes difference between the teacher output y_t, and student prediction, $f_t(x)$. Beta param controls transition thresold.

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

t: current time-step
f(x): student prediction
B: beta param

Results

Fig1: Computer Vision

Table 1. Computer vision: top-1 validation accuracy on ImageNet-1K with ViT-B and ViT-L models. data2vec ViT-B was trained for 800 epochs and ViT-L for 1,600 epochs. We distinguish between individual models and setups composed of multiple models (BEiT/PeCo train separate visual tokenizers and PeCo also distills two MoCo-v3 models).

	ViT-B	ViT-L
Multiple models		
BEiT (Bao et al., 2021)	83.2	85.2
PeCo (Dong et al., 2022)	84.5	86.5
Single models		
MoCo v3 (Chen et al., 2021b)	83.2	84.1
DINO (Caron et al., 2021)	82.8	-
MAE (He et al., 2021)	83.6	85.9
SimMIM (Xie et al., 2021)	83.8	-
iBOT (Zhou et al., 2021)	83.8	-
MaskFeat (Wei et al., 2021)	84.0	85.7
data2vec	84.2	86.6

Fig2: Speech

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 6).

	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

Fig3: NLP

Table 4. 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.
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