Data2vec for Speech

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, Michael Auli

What is data2vec?

- Generalized self-supervised learning algorithm that works on audio, images and text
- SOTA results in a like-for-like setup on speech recognition (Librispeech) and image classification (Imagenet), and competitive to leading algorithms on text classification tasks (GLUE).
- Self distillation setup: uses a momentum teacher to generate contextualized targets and learns by reconstructing them to solve a masked prediction task (more details later)

Motivation

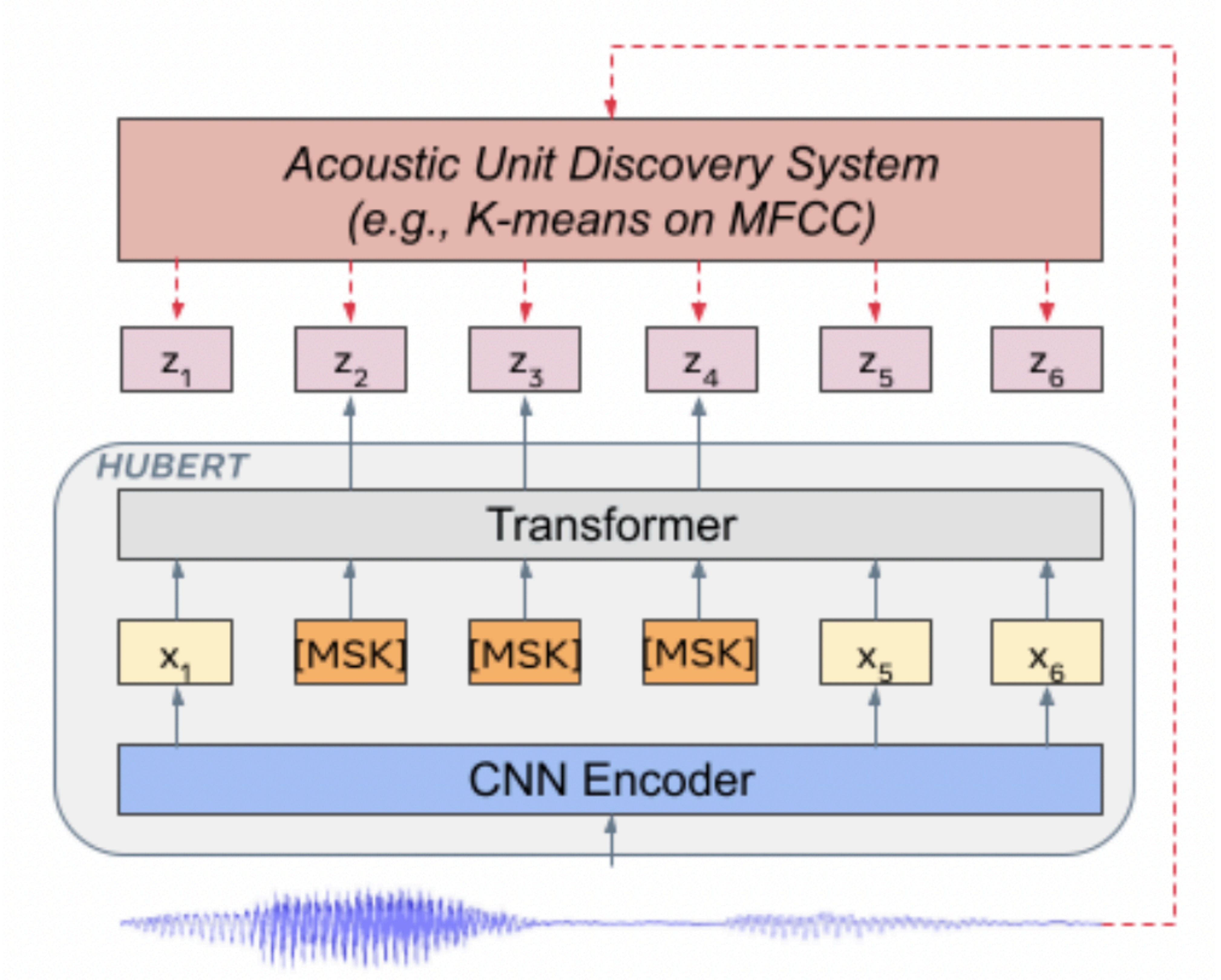
- Hypothesis: a good self-supervised learning algorithm learns representations that are **contextualized** and **predictive**
- The same algorithm should work on any kind of data that is structured (i.e. context can be used to infer unseen data points)
- Most leading SSL techniques are based on predicting or reconstructing local input (e.g. BERT, wav2vec 2.0, MAE), or learning a data augmentation invariant representation (e.g. BYOL, DINO)
- Can we do better?

Inspiration

- HUBERT (Hsu, et al) learns **contextualized** representations by clustering intermediate transformer representations

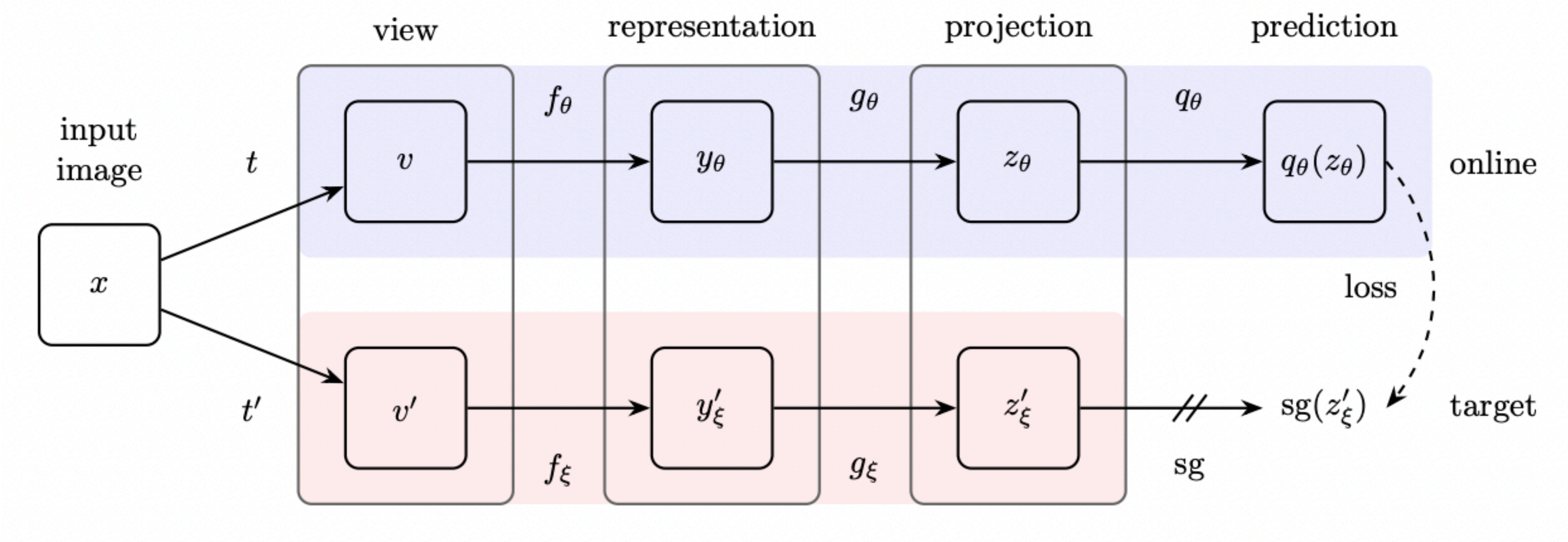
- BYOL (Grill, et al) / DINO (Caron, et al) learn data augmentation invariant representation via self-distillation from a momentum teacher

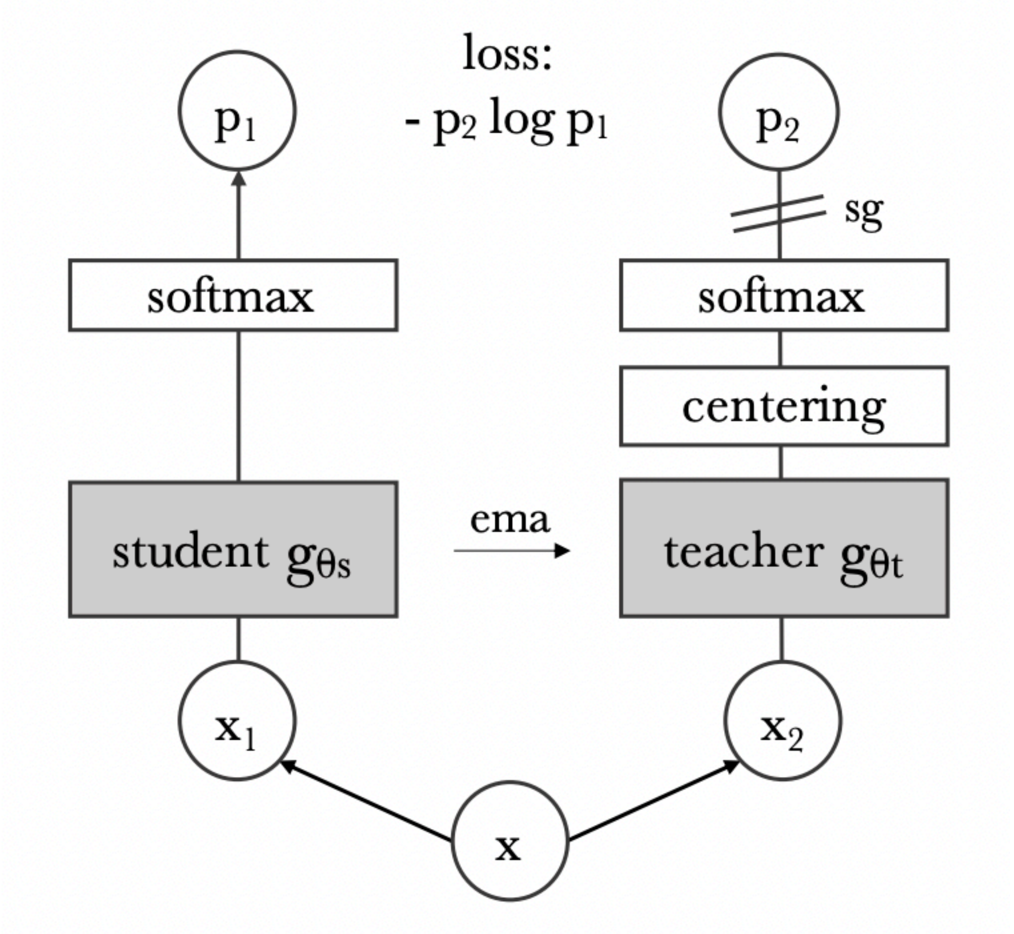
HUBERT



- Works very well for speech recognition and other speech tasks!
- Targets are cluster identities which are discrete
- Requires a pipeline approach with a supervised layer selection step

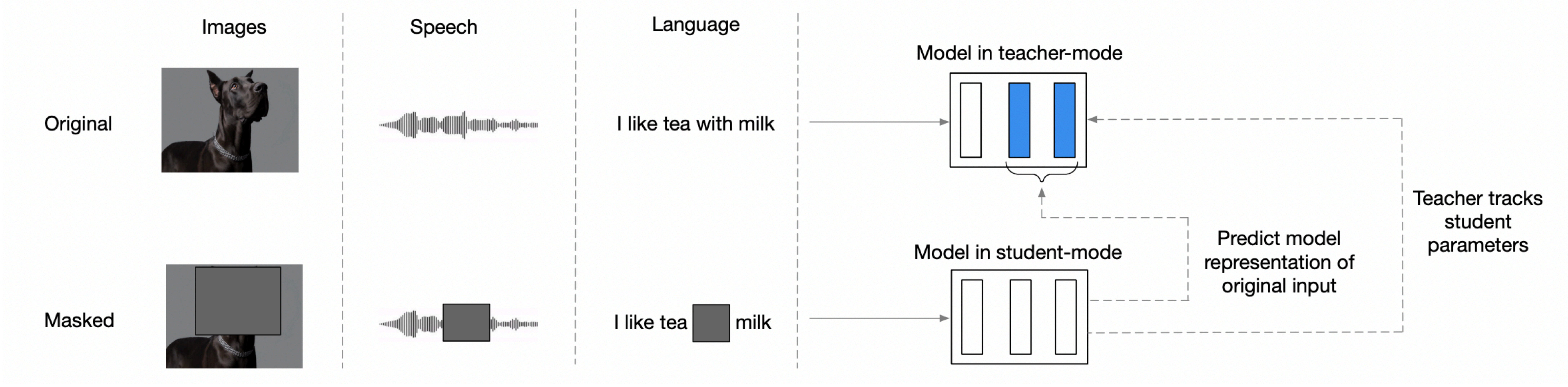
BYOL & DINO



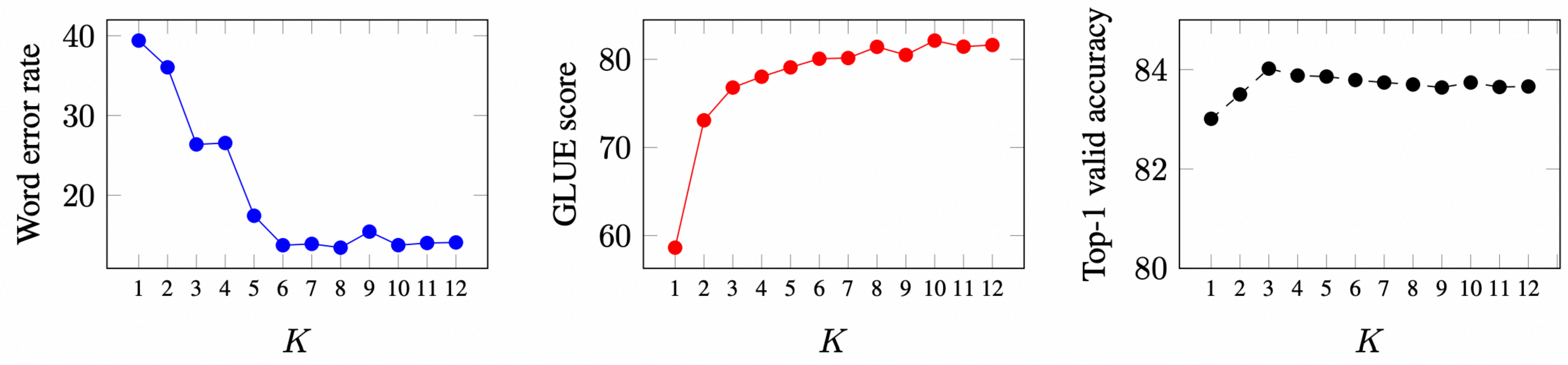


- Learns very good representations
- Improves targets over time through momentum updates of the teacher
- Relies on hand-crafted augmentation policy

Data2vec



- Modality specific feature encoder
- Common masking policy, but modality / dataset specific parameterization
- Identical context encoder (transformer)
- Identical learning task



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

Results

Librispeech test-other word error rate decoded with a 4-gram language model

Base models (95M parameters), 960 hours

	Wav2vec 2.0	HUBERT	WavLM	Data2vec
Pretrain updates	400k	250k + 400k	250k + 400k	400k
10 min	15,6	15,3	-	12,3
1 hour	11,3	11,3	10,8	9,1
10 hours	9,5	9,4	9,2	8,1
100 hours	8,0	8,1	7,7	6,8

Large models (300M parameters), 60k hours

	Wav2vec 2.0	HUBERT	WavLM *	Data2vec
Pretrain updates	1 million	250k + 400k + 400k	250k + 400k + 700k	600k
10 min	10,3	10,1	-	8,4
1 hour	7,1	6,8	6,6	6,3
10 hours	5,8	5,5	5,5	5,3
100 hours	4,6	4,5	4,6	4,6

* Pretrained on additional data

Limitations

- Modality specific feature encoder + masking parameters
- Sensitive to hyper parameter choices
  - Model collapses or plateaus if not well-tuned
- Requires two forward passes during pre-training
  - Can re-use feature encoder output
  - Can encode fewer examples, but generate several masks

Future work

- Recipes for additional modalities (videos, off-line RL, etc)
- Modality agnostic feature encoders (e.g. like in Perceiver (Jangle et al.))
- Multi-modal representation learning
- Causal / generative pre-training

Thank you!

- Questions?