Building Intelligent Machines that Learn from Human Speech

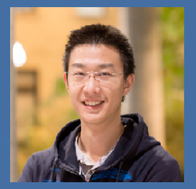
Michael Auli

FAIR
California


Arun Babu


Alexis Conneau


Steffen Schneider


Henry Zhou


Abdelrahman Mohamed


Jiatao Gu

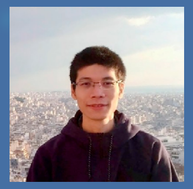

Naman Goyal


Wei-Ning Hsu


Alexei Baevski


Michael Auli


Kushal Lakhotia


Andros Tjandra


Kritika Singh


Yatharth Saraf


Geoffrey Zweig


Qiantong Xu


Tatiana Likhomanenko


Paden Tomasello


Ronan Collobert


Gabriel Synnaeve

Speech is Rich in Information

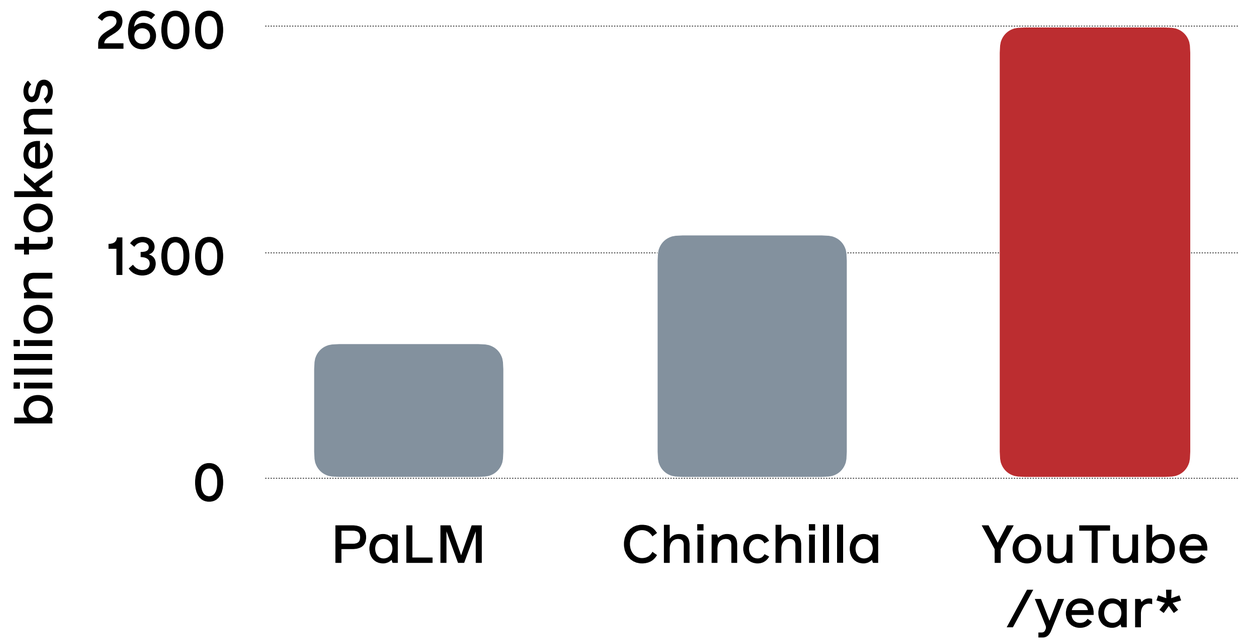
- Voice carries a lot of information: what you say/how you say it.
- Stress in our voice, intonation, and other paralinguistic features.
- Children learn a lot by listening to others.

Speech is Natural & Interactive

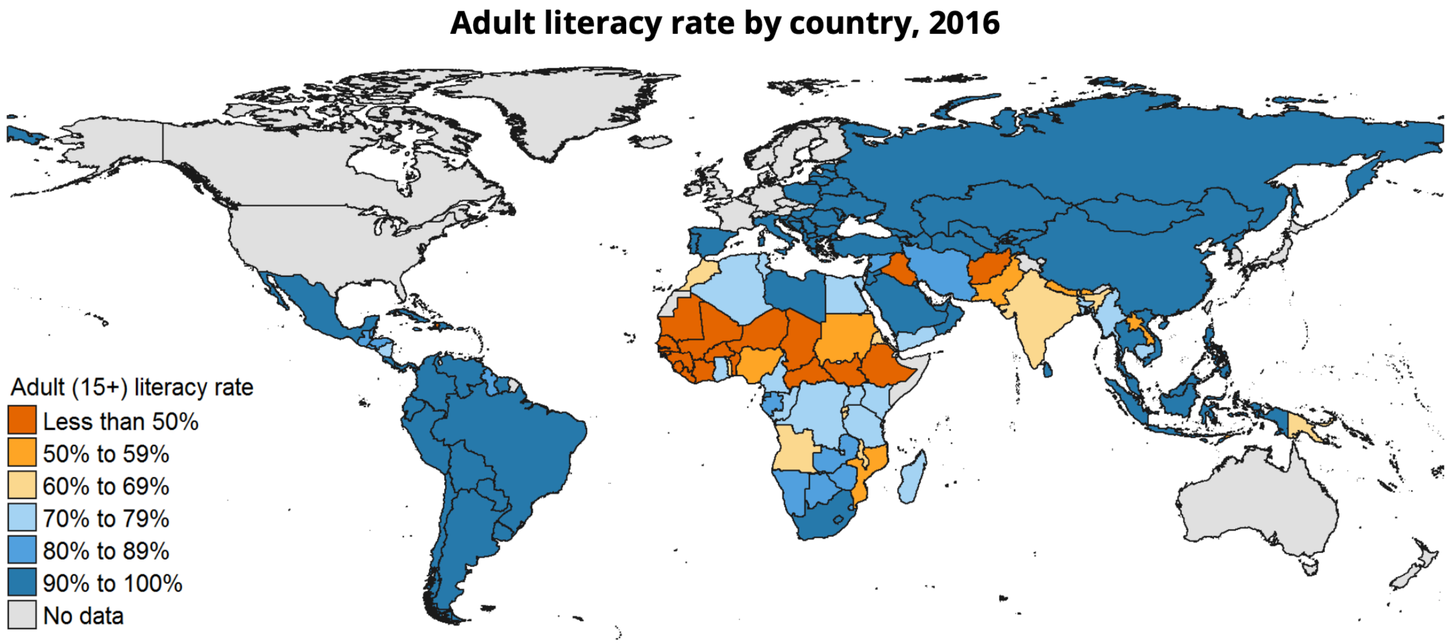
- Intelligent machine communicating using speech.
- High-latency (text) vs. low-latency (speech).
- We speak faster than we can type (2-3 words/sec).

Speech is Ubiquitous

- Large language models are trained on a lot of data
- YouTube adds 500 hours of video data/minute. Up to 2.6T tokens/year.



Access to Technology and Information



Language Diversity

The New York Times

World's Languages Dying Off Rapidly

Give this article

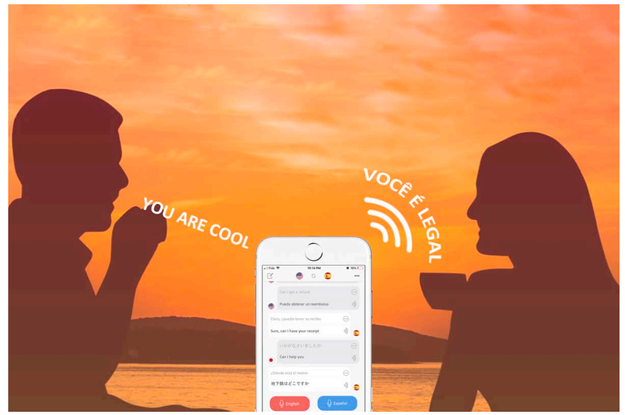
By John Noble Wilford

Sept. 18, 2007

Of the estimated 7,000 languages spoken in the world today, linguists say, nearly half are in danger of extinction and are likely to disappear in this century. In fact, they are now falling out of use at a rate of about one every two weeks.

Speech Applications

- Speech to text/speech recognition - dictation etc.
- Text to speech - reading out aloud
- Keyword spotting - ”Hey Alexa/Portal”
- Speaker identification - is it your voice?
- Language identification
- Speech translation



This Talk

- **wav2vec**: a self-supervised algorithm for speech representations.
- **wav2vec-U**: self-supervised learning enables unsupervised speech recognition.
- **data2vec**: unified objective for self-supervised learning in multiple modalities.

Self-supervised Speech Representation Learning

Supervised Machine Learning

( , cat)

Not how humans learn!

potential train/test mismatch

Need to annotate lots of data!

Supervised Machine Learning

Self-supervised Learning

- Learn good data representations (structure, features etc.) **without labels**
- |Unlabeled data| >> |Labeled data|
- Use representations to solve the task

Neural network

cat

Input

Output

Supervised learning simultaneously performs representation learning of the data and associating these features with labels

Limitation: relies on labeled data to learn feature encoding

Input

Feature encoder

Data representation

Classifier

cat

Output

Self-supervised learning:

1/ representation learning of the data

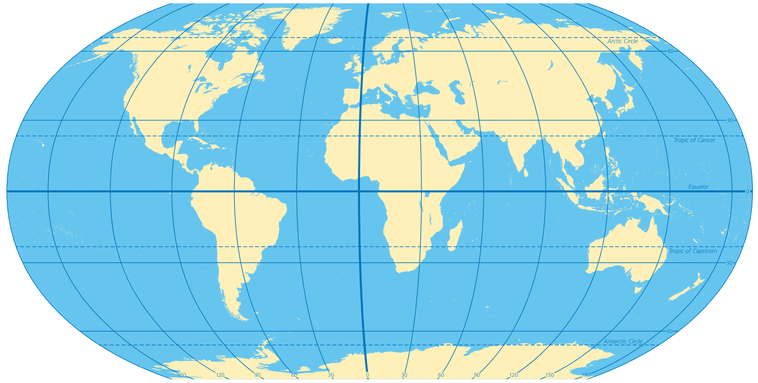
2/ learn to associate labels with the representations

Reduces reliance on labeled data!

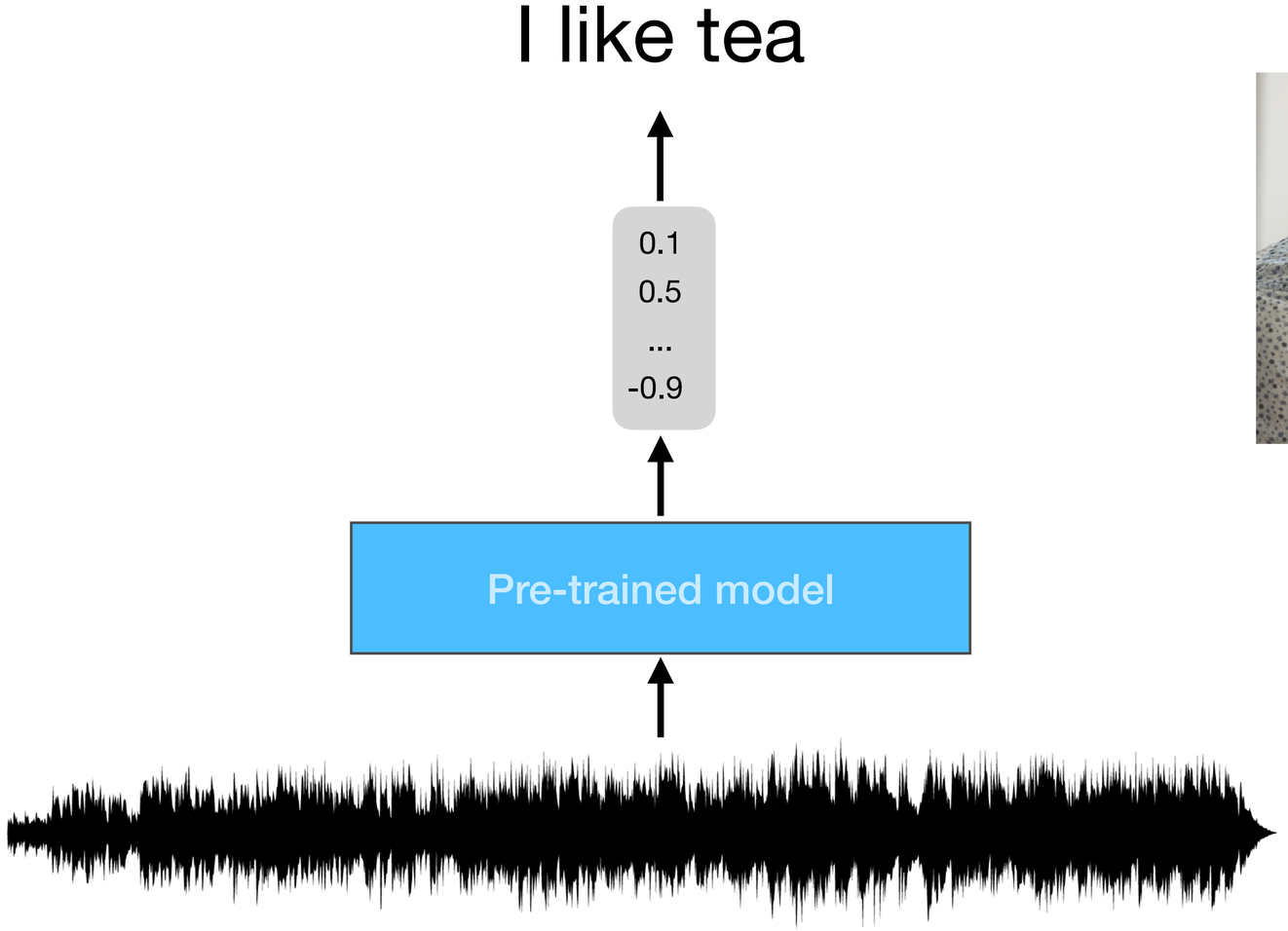
Training Speech Recognition Models

I like black tea with milk

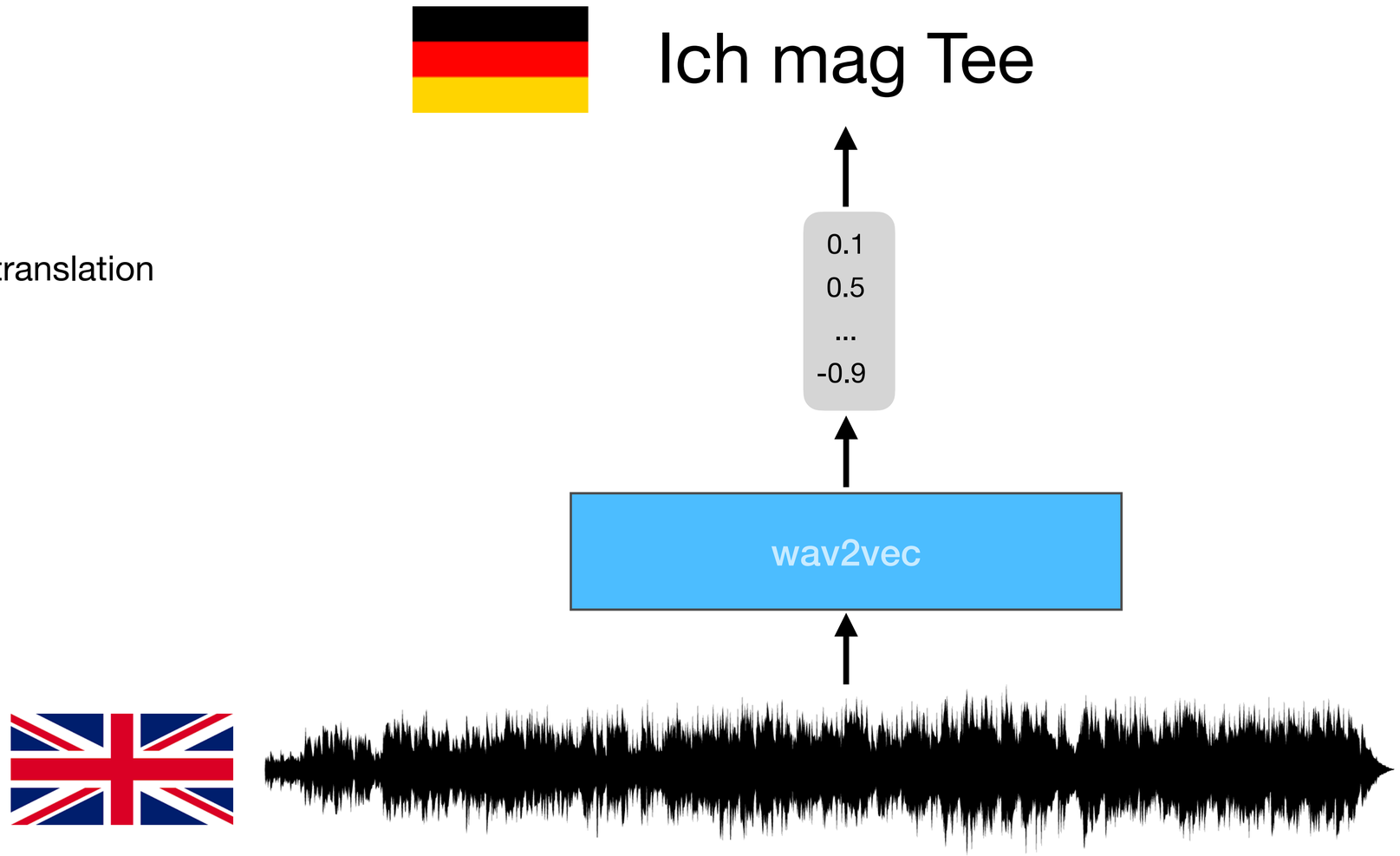
- Train on 1,000s of hours of data for good systems.
- Many languages, dialects, domains etc.



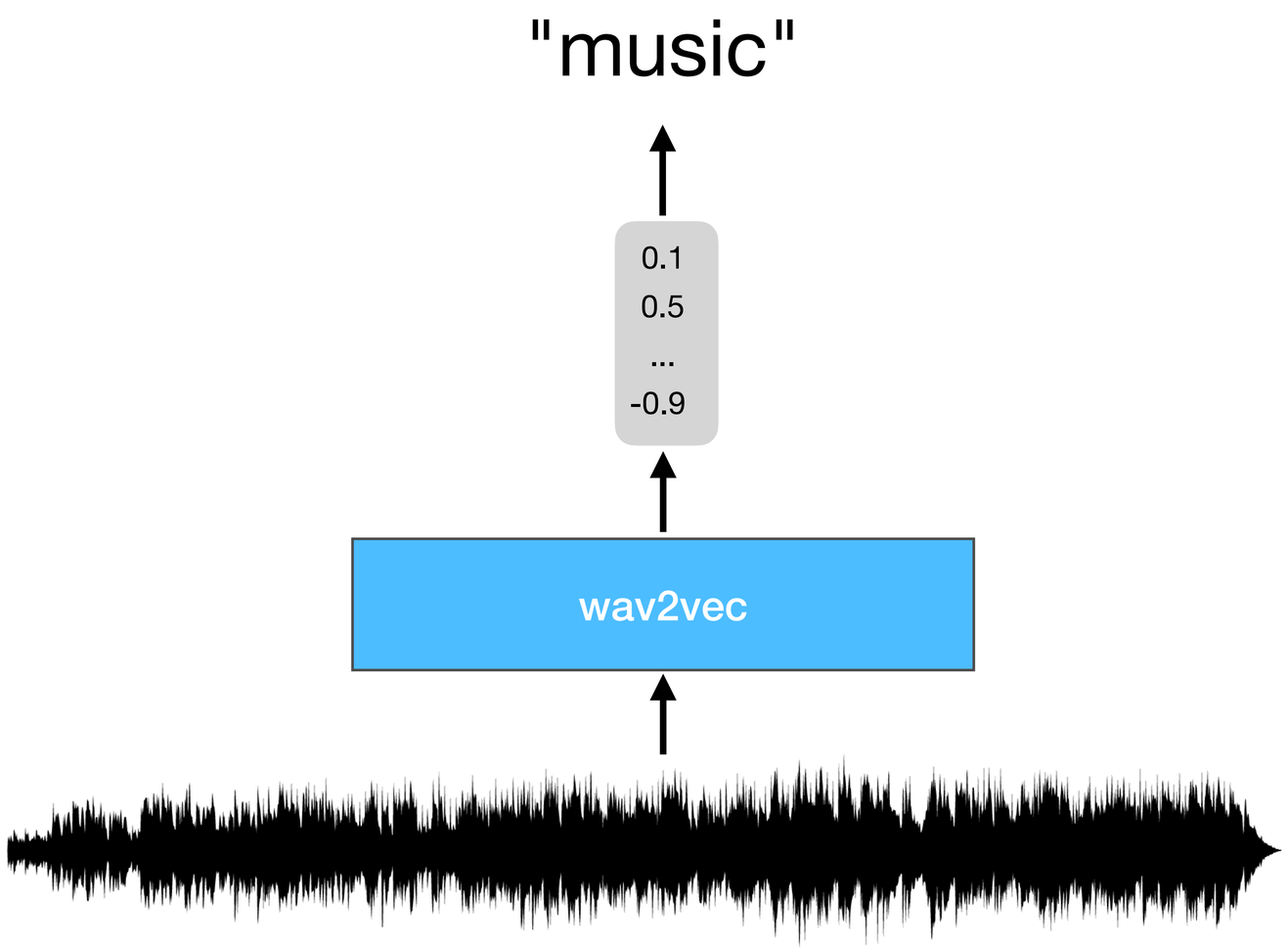
Speech recognition



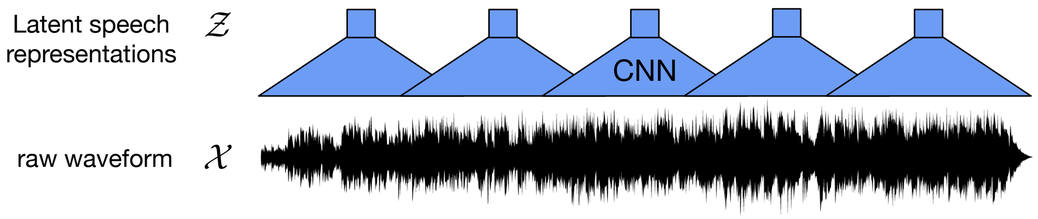
Speech translation



Audio event detection

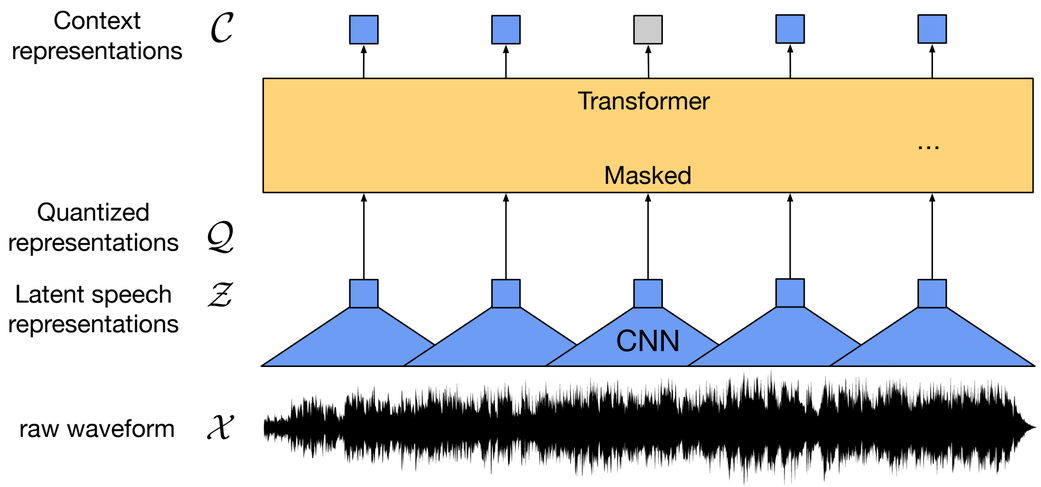


wav2vec 2.0



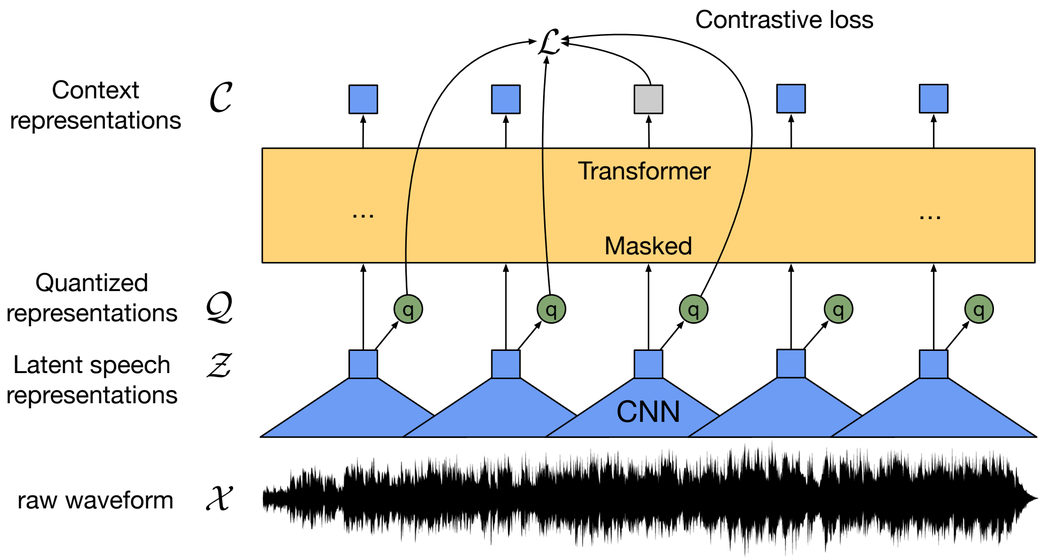
- Masked prediction with transformer, bi-directional contextualized representations (similar to BERT).
- But predict what? Learn an inventory of speech units with vector quantization via Gumbel softmax.
- Learning task: Joint VQ & context representation learning.
- Contrast true quantized latent with distractor latents.

wav2vec 2.0



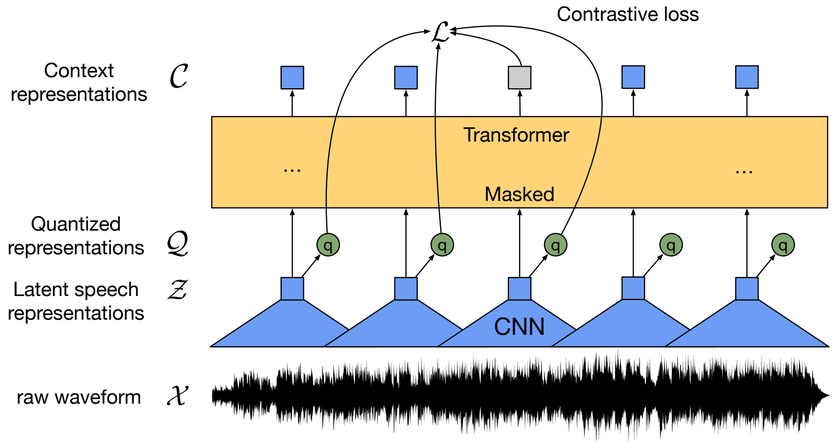
- Masked prediction with transformer, bi-directional contextualized representations (similar to BERT).
- But predict what? Learn an inventory of speech units with vector quantization via Gumbel softmax.
- Learning task: Joint VQ & context representation learning.
- Contrast true quantized latent with distractor latents.

wav2vec 2.0



- Masked prediction with transformer, bi-directional contextualized representations (similar to BERT).
- But predict what? Learn an inventory of speech units with vector quantization via Gumbel softmax.
- Learning task: Joint VQ & context representation learning.
- Contrast true quantized latent with distractor latents.

Objective



Cosine similarity

Context representation

Discrete latent speech representation

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

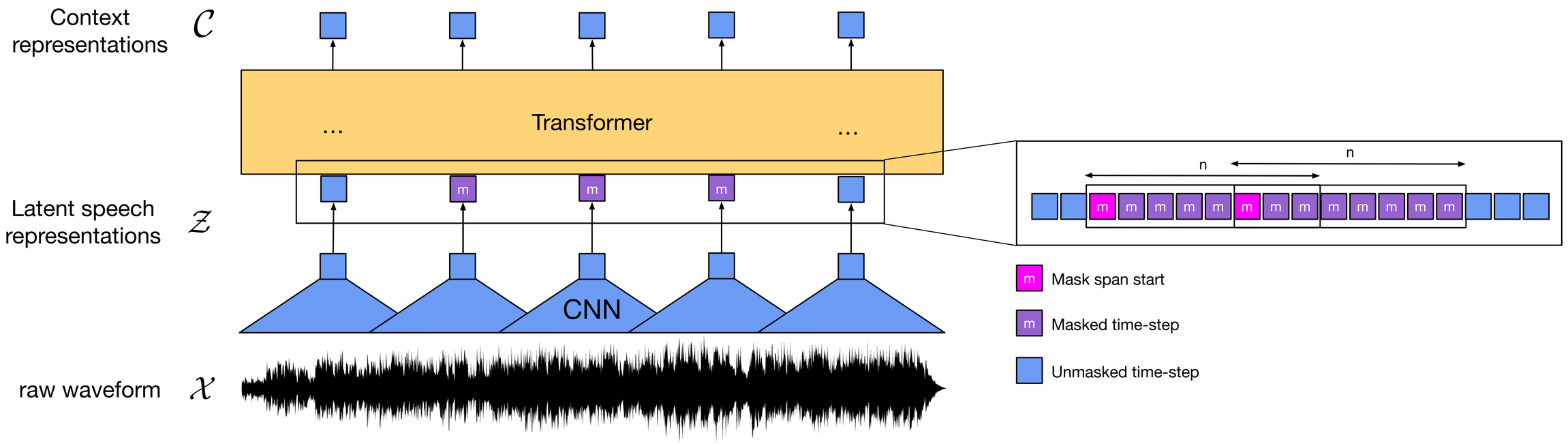
Negative samples

Temperature

Codebook diversity penalty to encourage more codes to be used

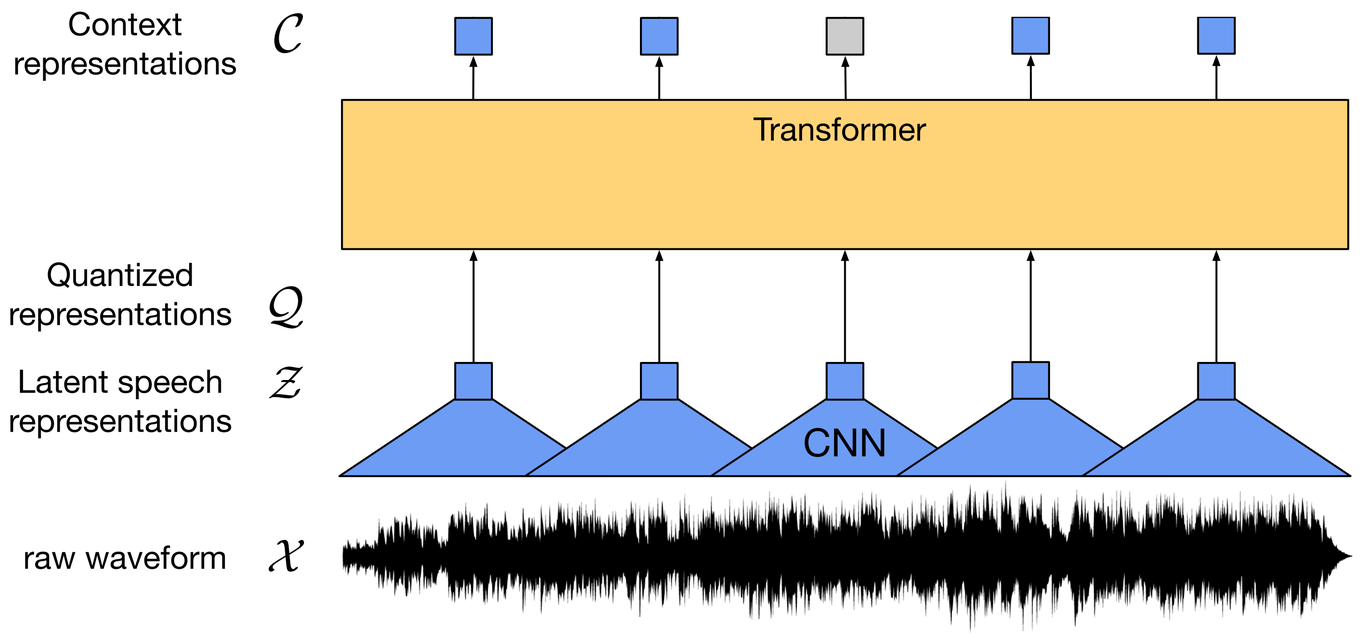
Masking

- Sample starting points for masks without replacement, then expand to 10 time-steps
- Spans can overlap
- For a 15s sample, ~49% of the time-steps masked with an average span length of ~300ms

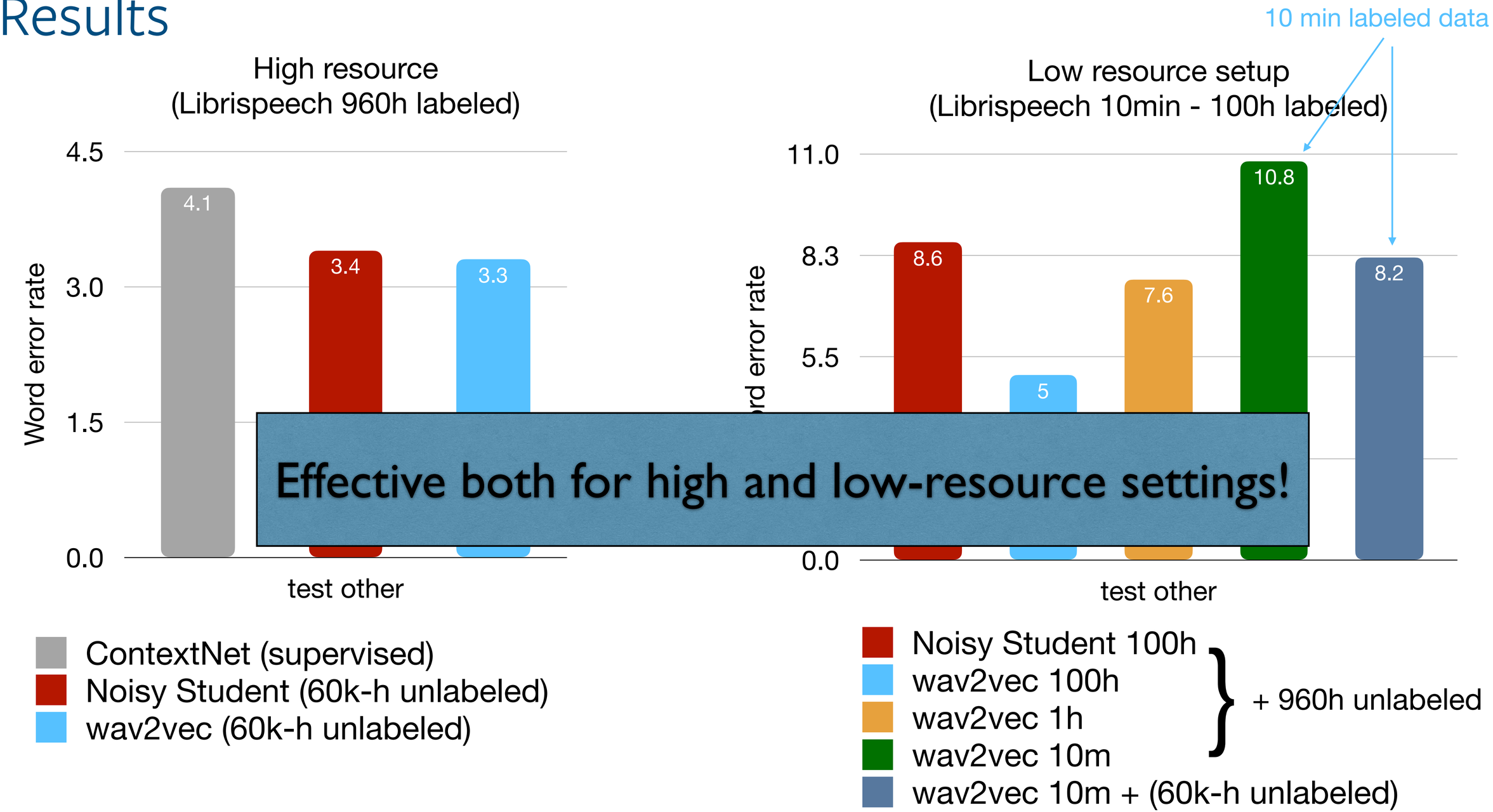


Fine-tuning

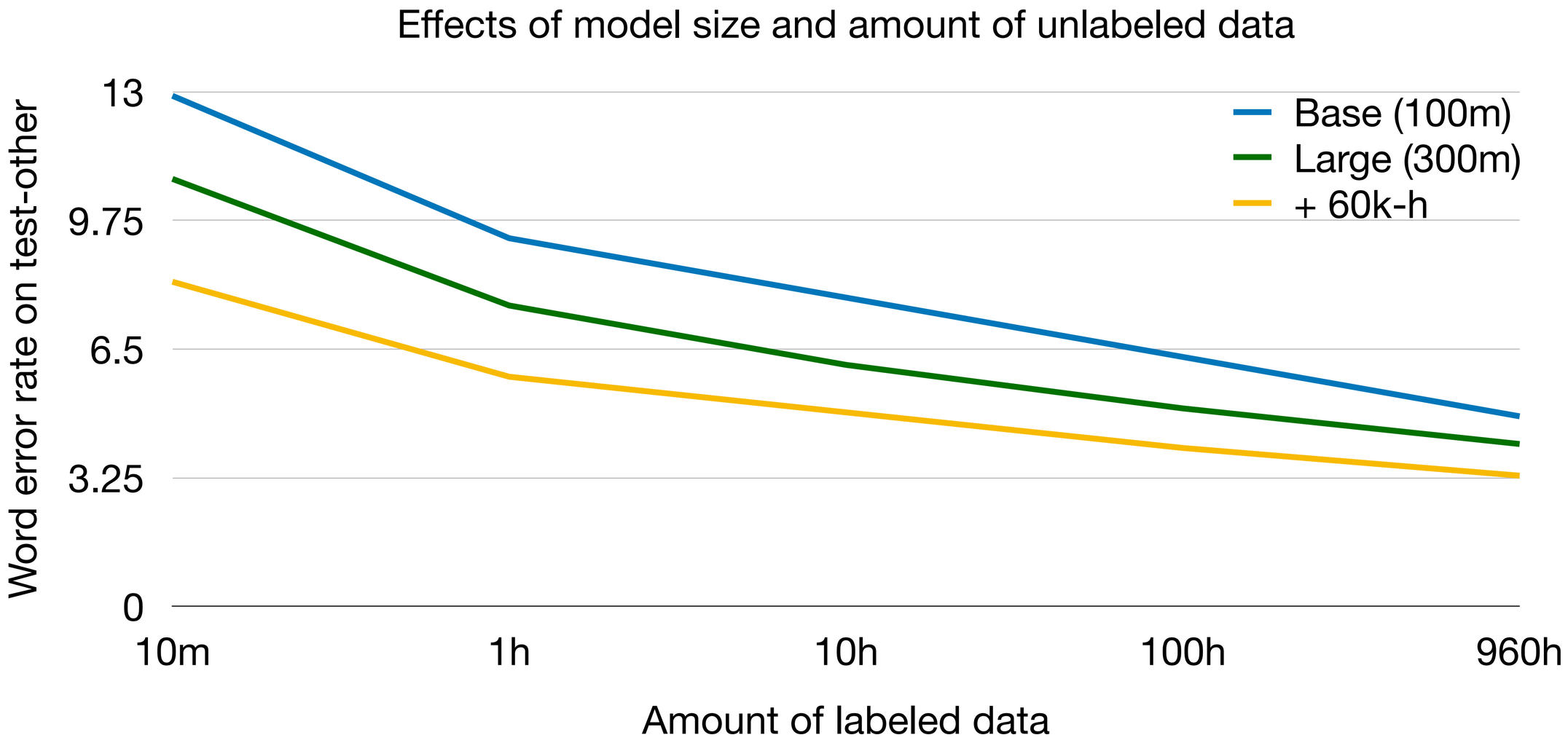
- Fine-tune model on labeled data for ASR with CTC (or other speech tasks)
- SpecAugment-style regularization & remove quantization



Results


High resource
(Librispeech 960h labeled)
Word error rate
4.5
3.0
1.5
0.0
4.1
3.4
3.3
test other
Low resource setup
(Librispeech 10min - 100h labeled)
10 min labeled data
11.0
8.3
5.5
0.0
8.6
5
7.6
10.8
8.2
test other
Effective both for high and low-resource settings!
ContextNet (supervised)
Noisy Student (60k-h unlabeled)
wav2vec (60k-h unlabeled)
Noisy Student 100h
wav2vec 100h
wav2vec 1h
wav2vec 10m
wav2vec 10m + (60k-h unlabeled)
+ 960h unlabeled

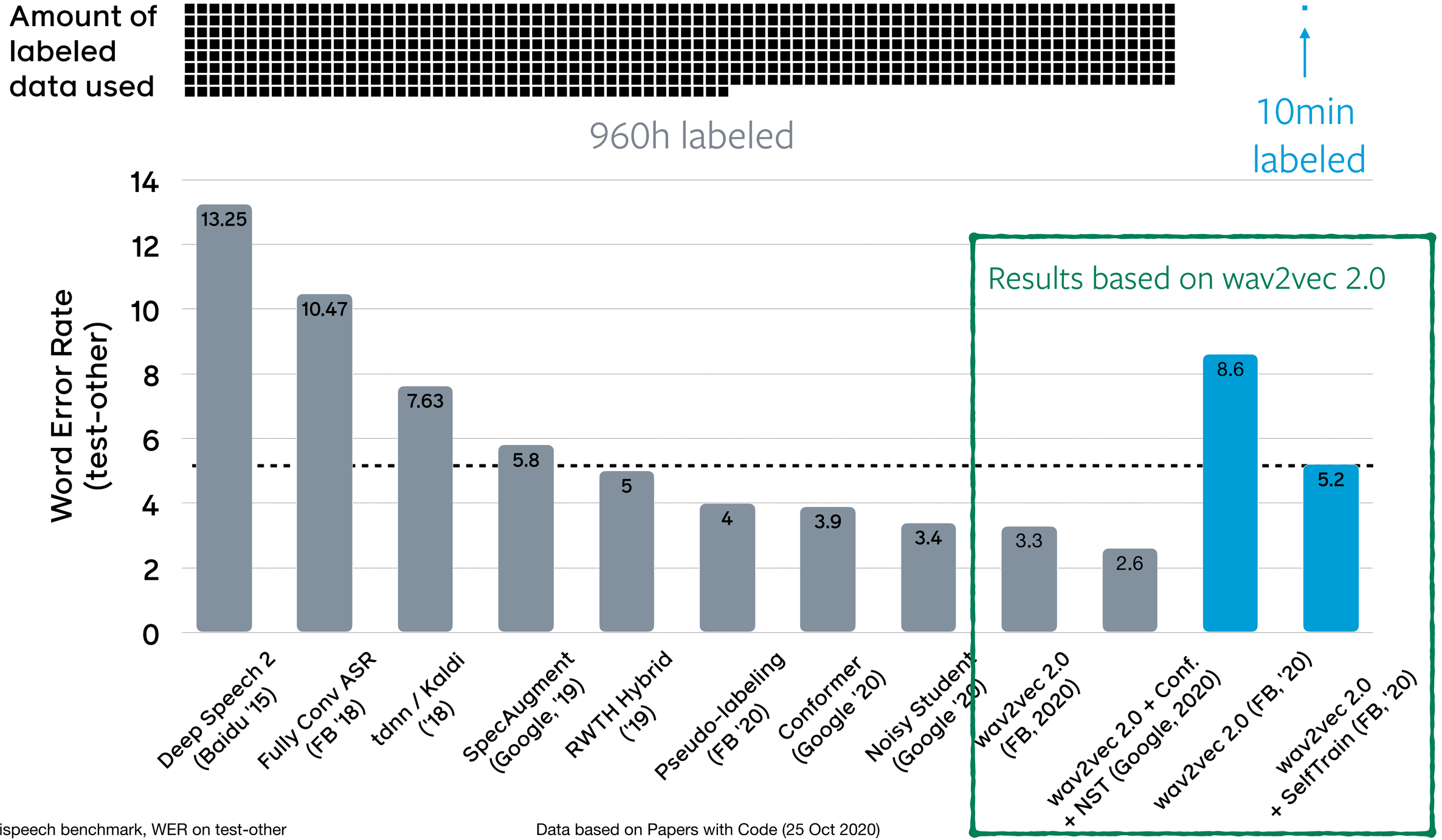
Results


Effects of model size and amount of unlabeled data
Word error rate on test-other
13
9.75
6.5
3.25
0
Base (100m)
Large (300m)
+ 60k-h
10m
1h
10h
100h
960h
Amount of labeled data

Examples (10 min Labeled Data)

HYP (no LM): she SESED and LUCHMAN GAIVE A SENT won by her GENTAL argument
HYP (w/ LM): she ceased and LUCAN gave assent won by her gentle argument
REF: she ceased and lakshman gave assent won by her gentle argument

HYP (no LM): but NOT WITH STANDING this boris EMBRAED him in a QUIAT FRENDLY way and CISED him THRE times
HYP (w/ LM): but NOT WITHSTANDING this boris embraced him in a quiet friendly way and kissed him three times
REF: but notwithstanding this boris embraced him in a quiet friendly way and kissed him three times



Librispeech benchmark, WER on test-other

Data based on Papers with Code (25 Oct 2020)

Summary

- For the first time, pre-training for speech works very well in both low-resource and high-resource setup.

- Using only 10 minutes (48 utterances) of transcribed data rivals best system trained on 960h from 1 year ago.

- Code and models are available in the fairseq GitHub repo + Hugging Face.

Unsupervised Speech Recognition

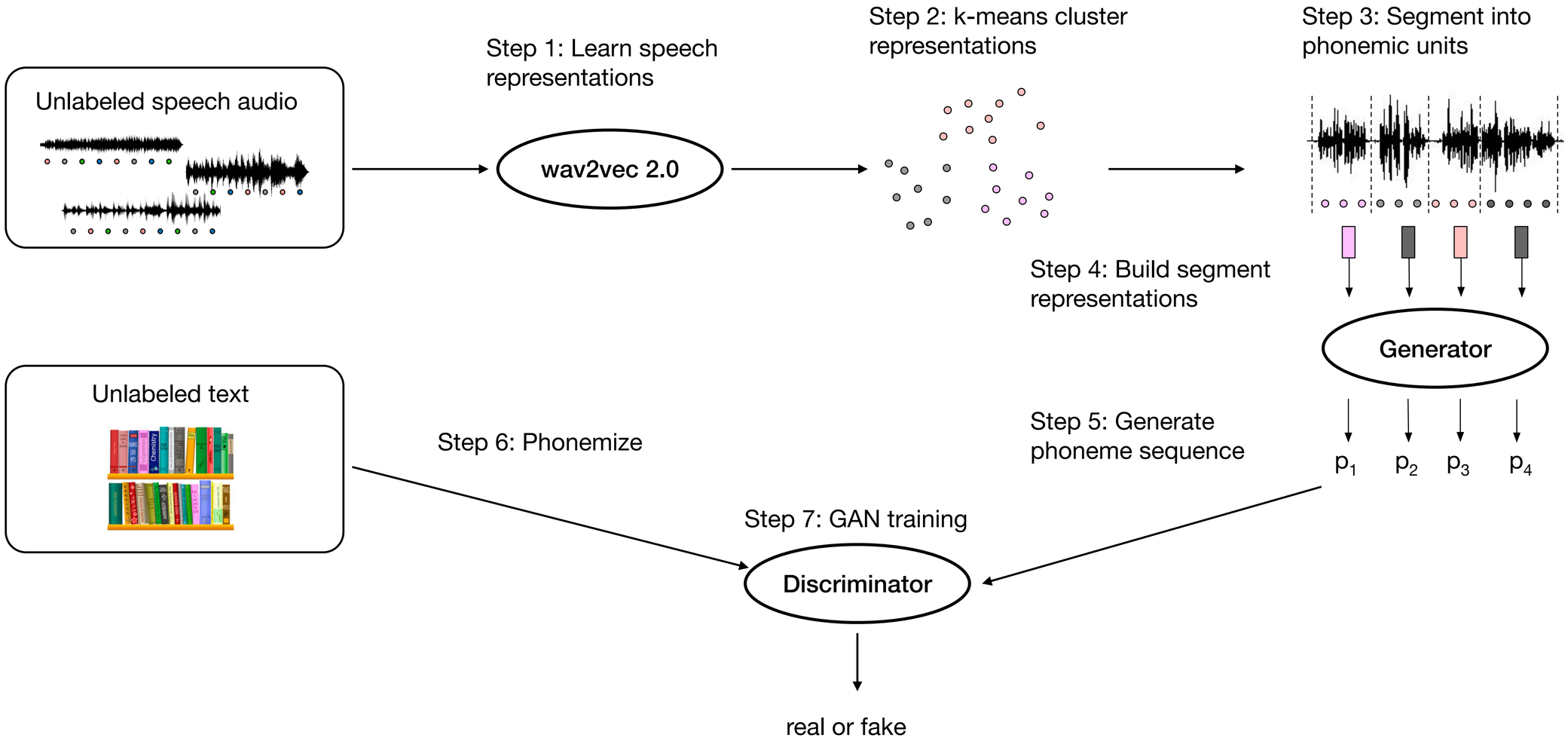
Unsupervised Speech Recognition

- Important step towards agents that can learn without supervision.
- Unsupervised machine translation exists, what about speech?
- Key problem: what are the units in the speech audio?

wav2vec Unsupervised: Key Ideas

- Learn good representations of speech audio
- Unsupervised segmentation of the speech audio into phonemic units
- Learn mapping between speech segments and phonemes using adversarial learning

wav2vec Unsupervised



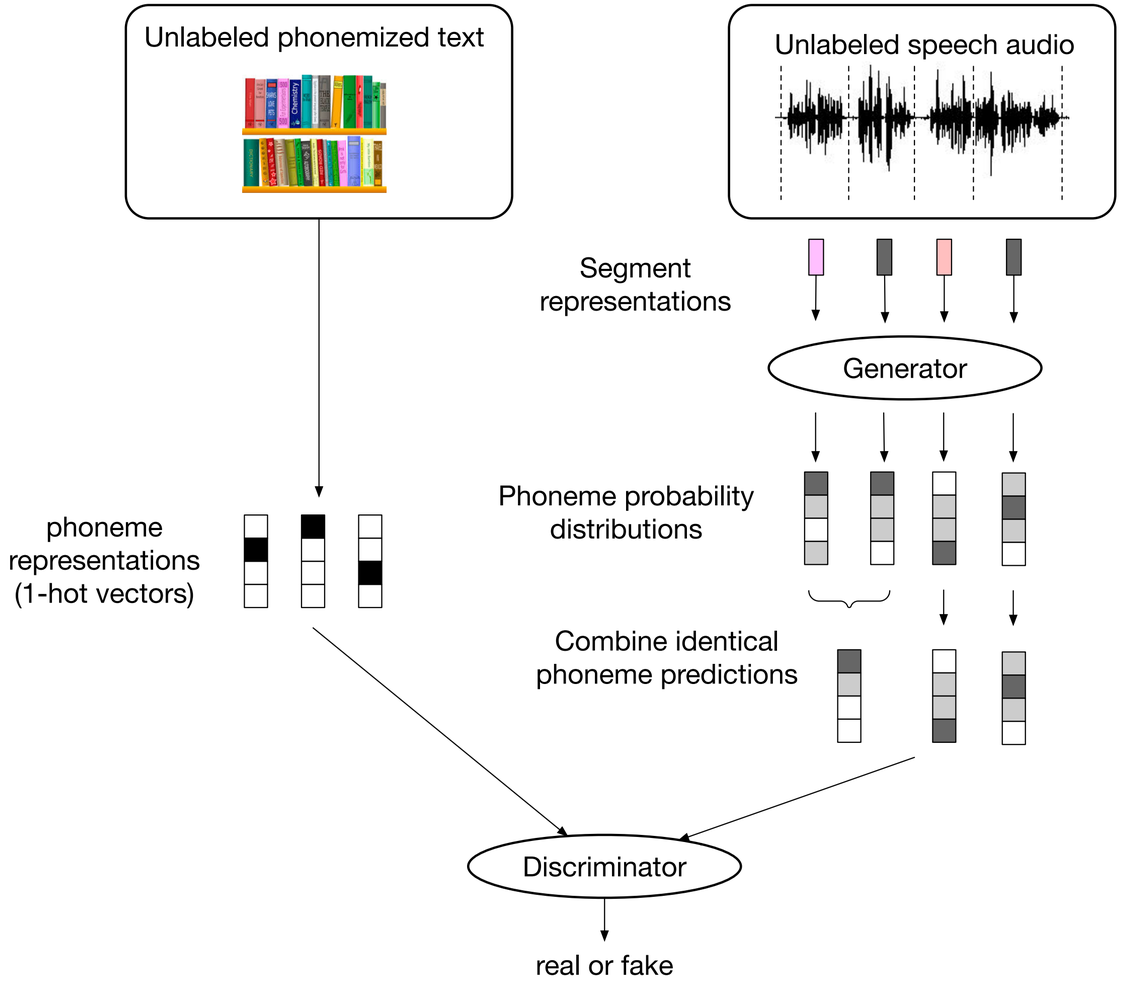
Text Data Pre-processing

he spoke soothingly

Phonemize

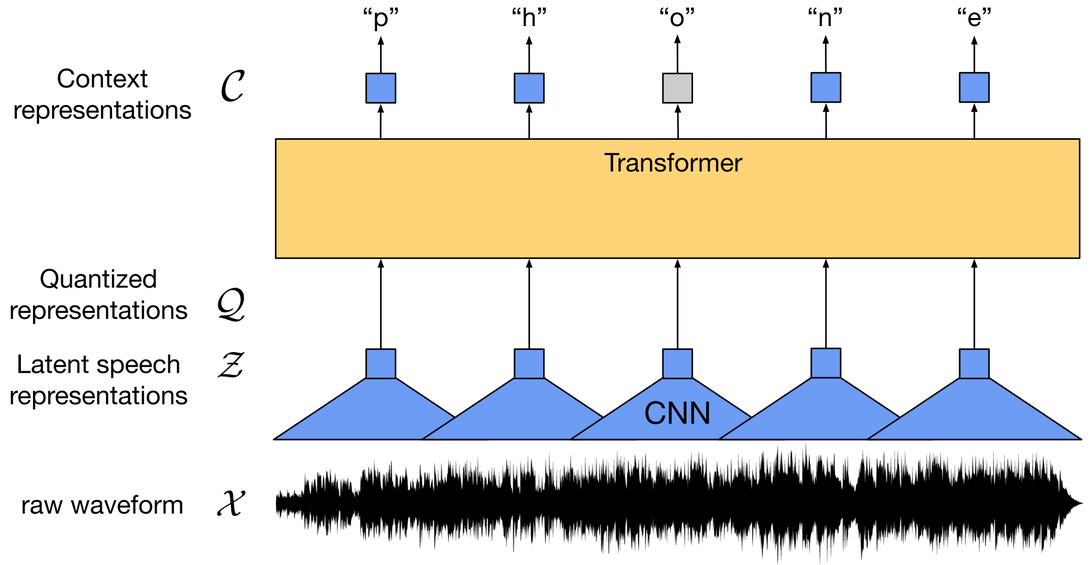
hh iy s ow k s uw dh ih ng l iy

GAN inputs


Unlabeled phonemized text
Unlabeled speech audio
Segment representations
Generator
Phoneme probability distributions
phoneme representations (1-hot vectors)
Combine identical phoneme predictions
Discriminator
real or fake

Generator / Discriminator

- Generator: 1 layer CNN with 90k parameters
  w2v features frozen
- Discriminator: 3 layer CNN
- Train time: 12-15h on a single V100



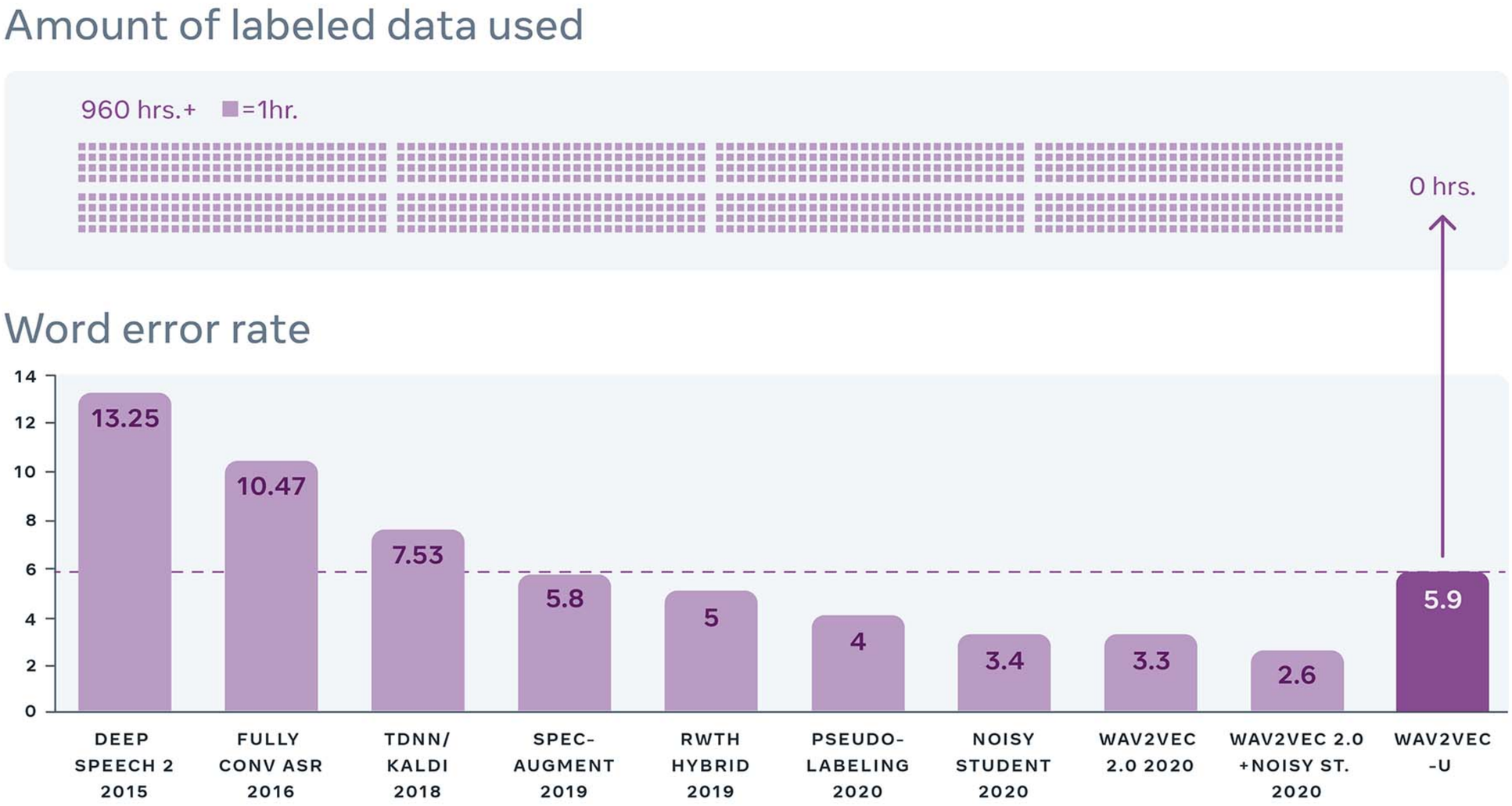
Training Details

- Unsupervised metric for early stopping, hyper-parameter selection
- Self-training after GAN training (HMM and fine-tuning w2v)

Comparison to Prior Unsupervised Work

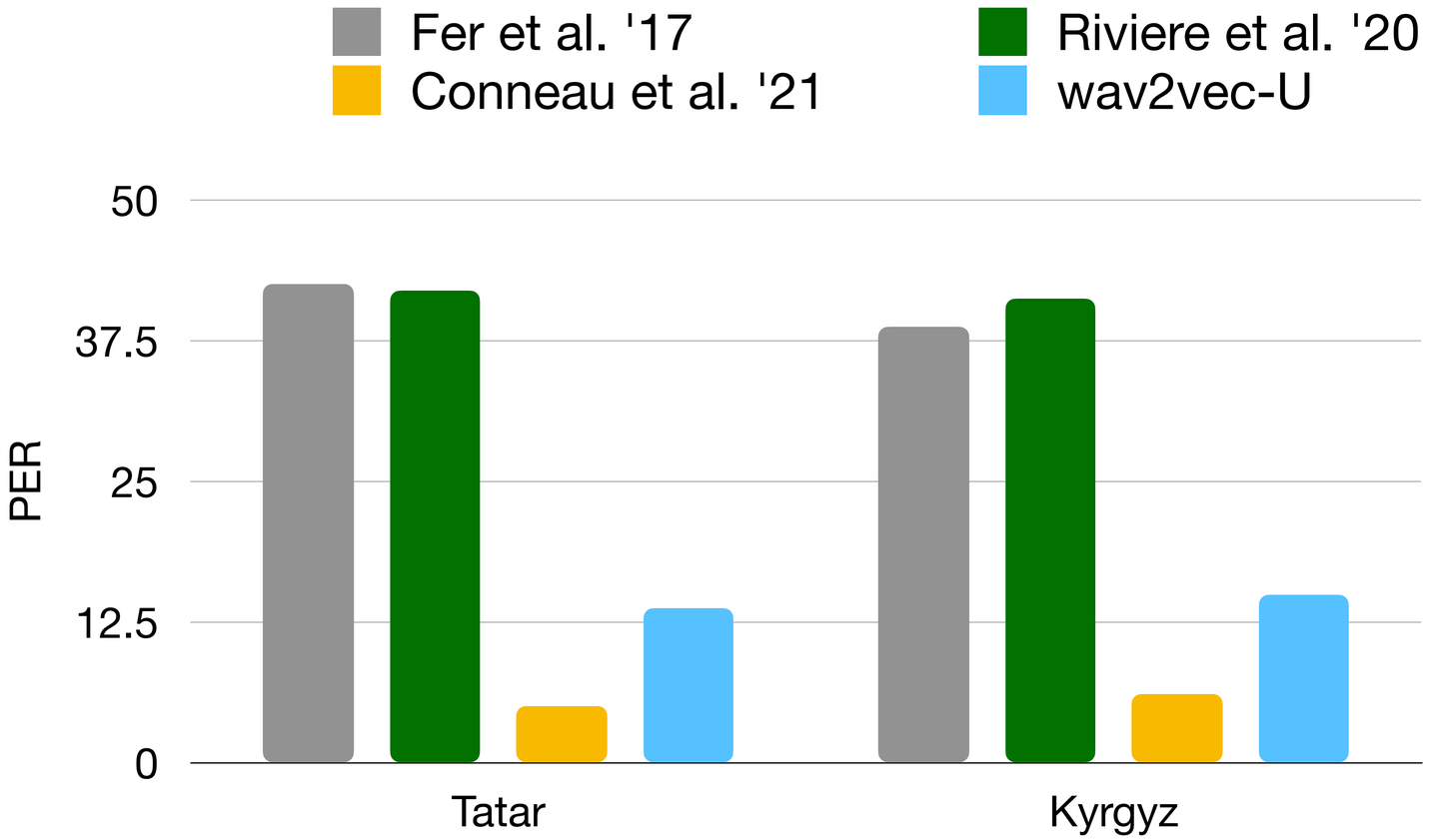
TIMIT Benchmark

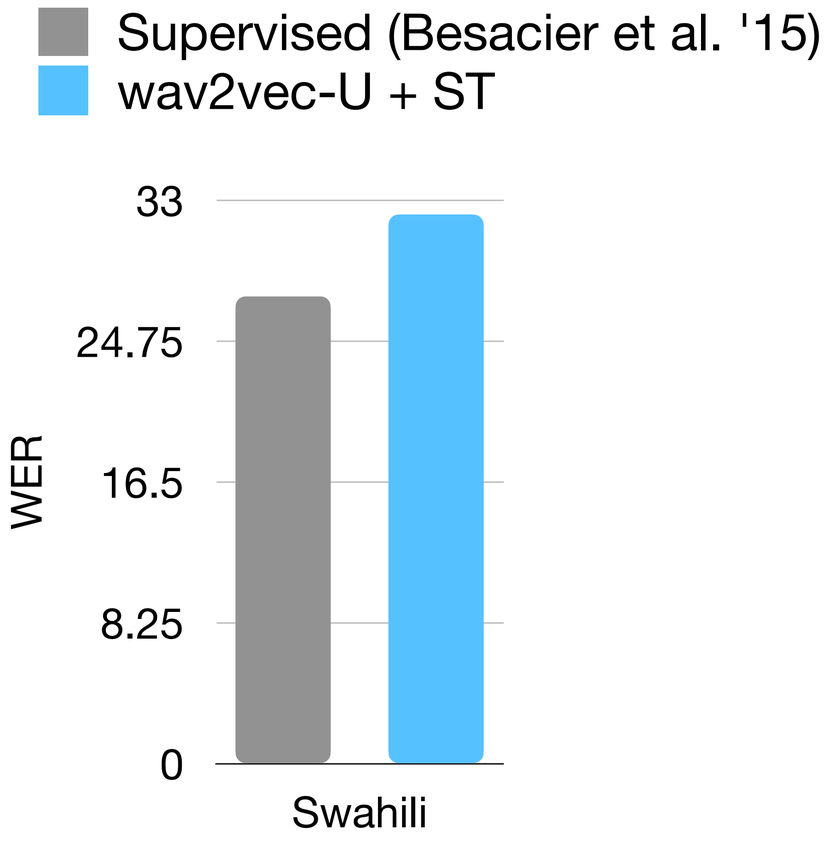
Comparison to Best Supervised Systems



Librispeech benchmark

Low-resource Languages





*wav2vec-U uses much less speech audio than prior work:
1.8h vs. 17h for Kyrgyz, 4.6h vs. 17h for Tatar

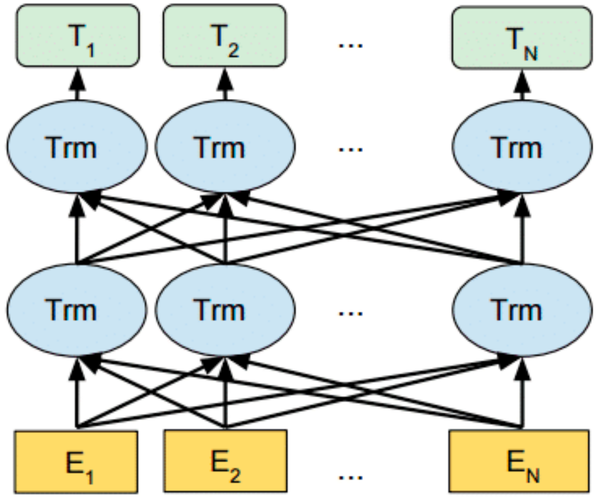
Discussion

- Very lightweight approach (except for wav2vec 2.0)
- Why does it work? Good audio features are main driver of performance
- Phonemizer still required
- Segment construction

data2vec: A Unified Objective for Self-supervised Learning

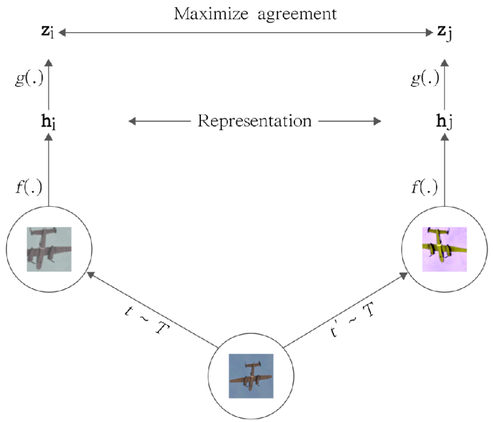
Natural Language Processing

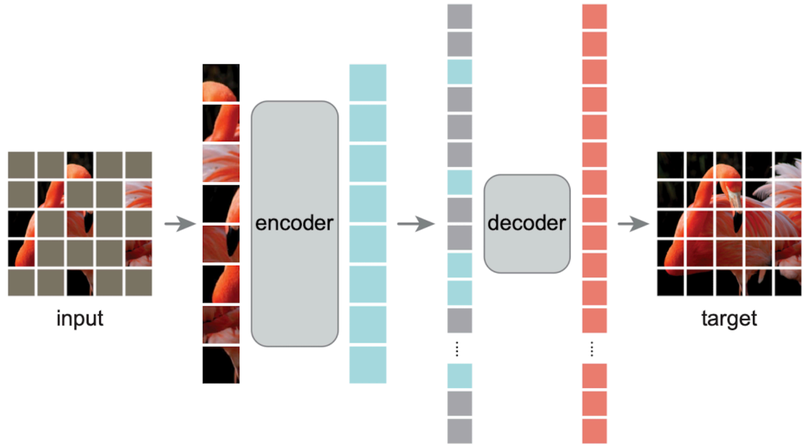
BERT

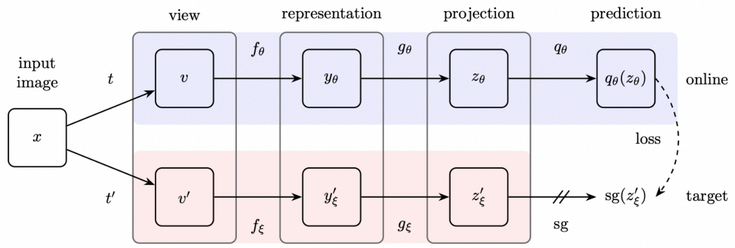


Computer Vision

SimCLR, BYOL, Masked AutoEncoders (MAE), …

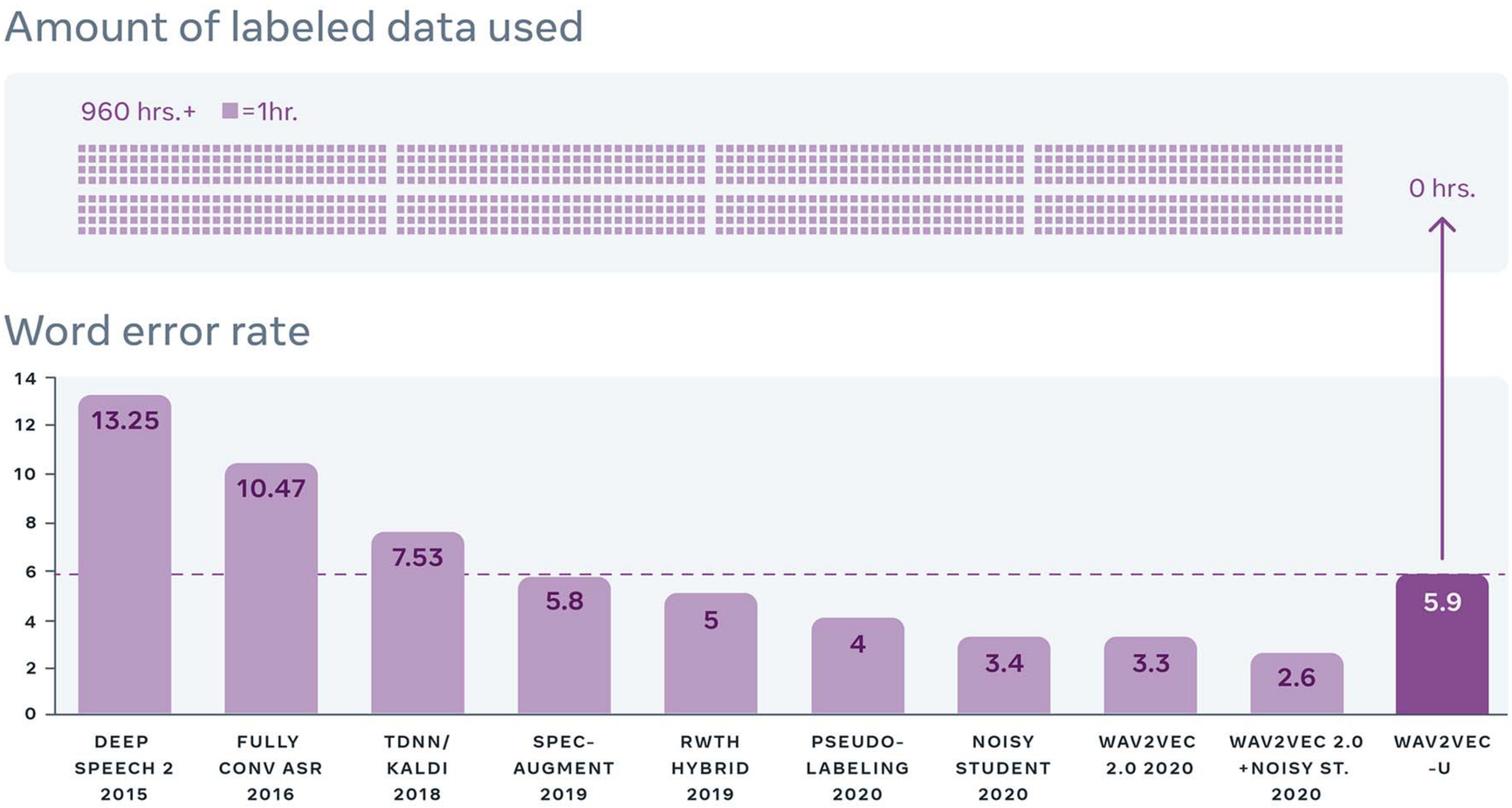






Speech: Unsupervised Speech Recognition

CPC, wav2vec 2.0, wav2vec Unsupervised, WavLM, w2v-BERT, HuBERT, …



Librispeech benchmark (test-other) compared with the best systems over time. Source: paperswithcode.com

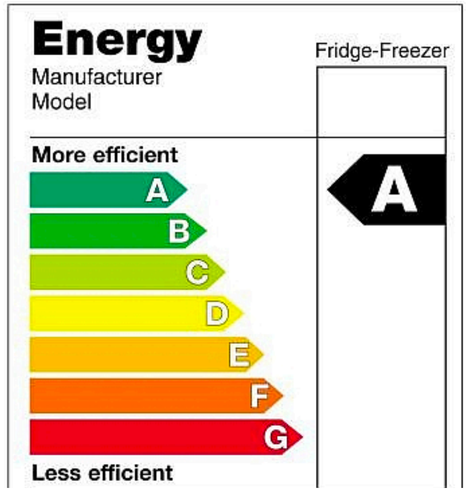
Two Challenges

Modality-specific Learning Algorithms

- Most algorithms developed for one modality - specific designs and learning biases.
- General idea of SSL. Biology of learning (Friston, ’10).
- **This talk:** single objective for vision, speech and text.

Little Focus on Efficiency

- Great progress but model sizes and compute requirements are ever growing.
- Are we using the best algorithms to push the boundaries?
- Scaling an efficient learner may ultimately get you further than an inefficient one.
- **This talk:** compute efficient SSL

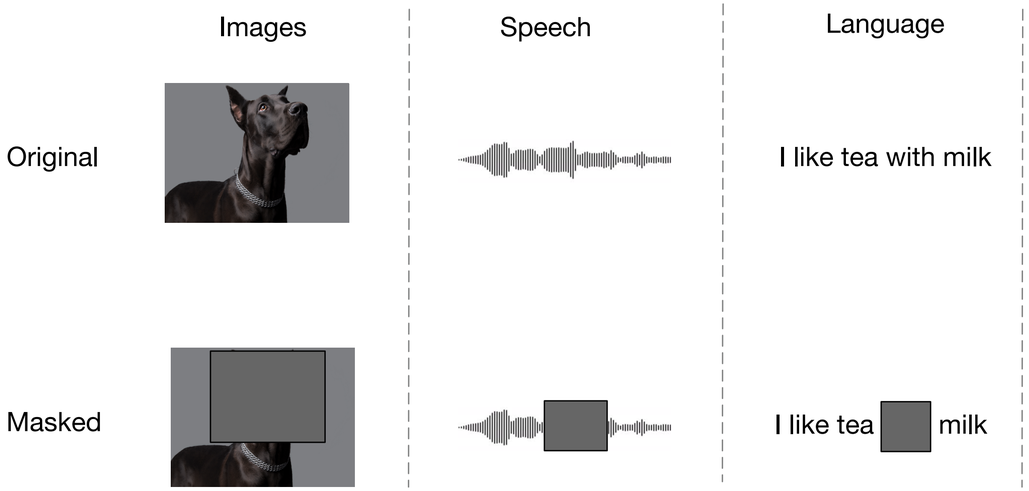


A Single Learning Objective

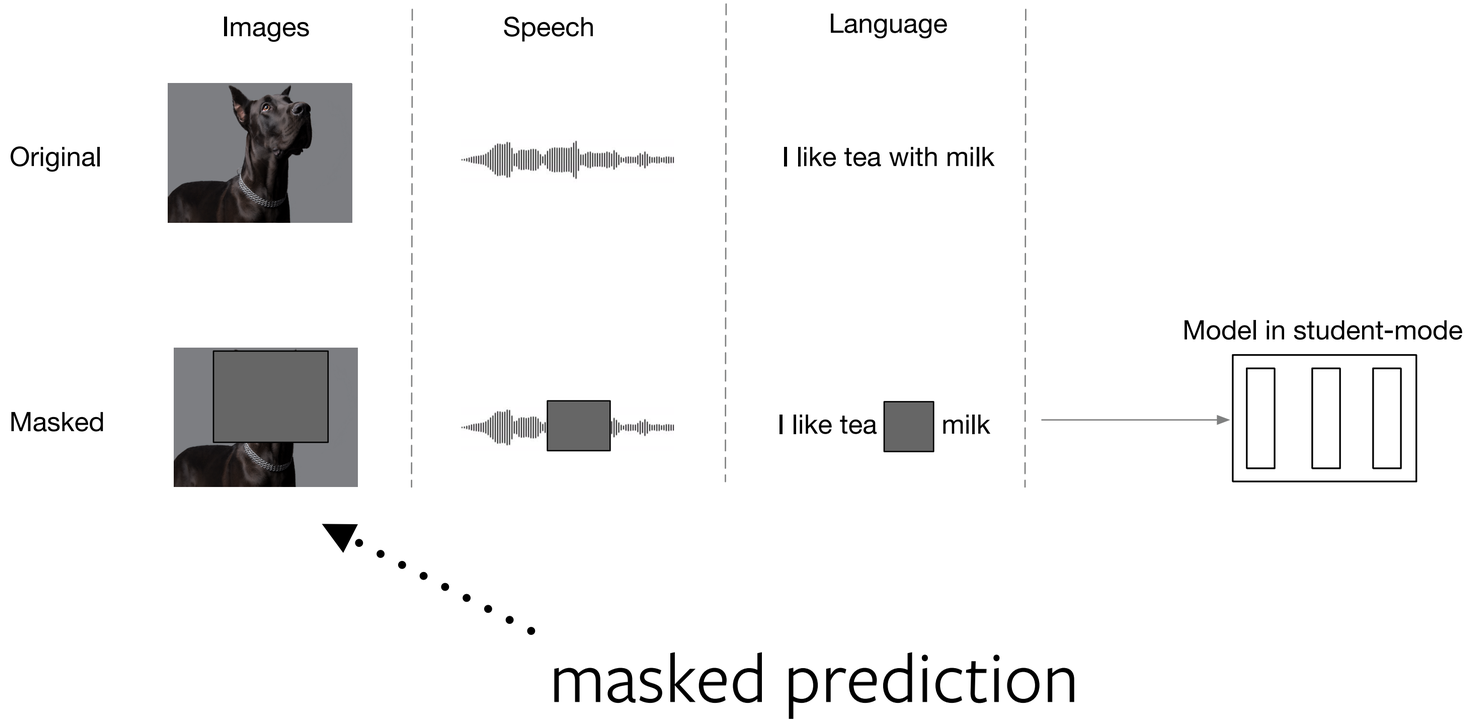
data2vec

- General algorithm that works very well across modalities.
- Same learning objective for each modality.
- How: self-distillation of contextualized representations in a masked prediction setup.

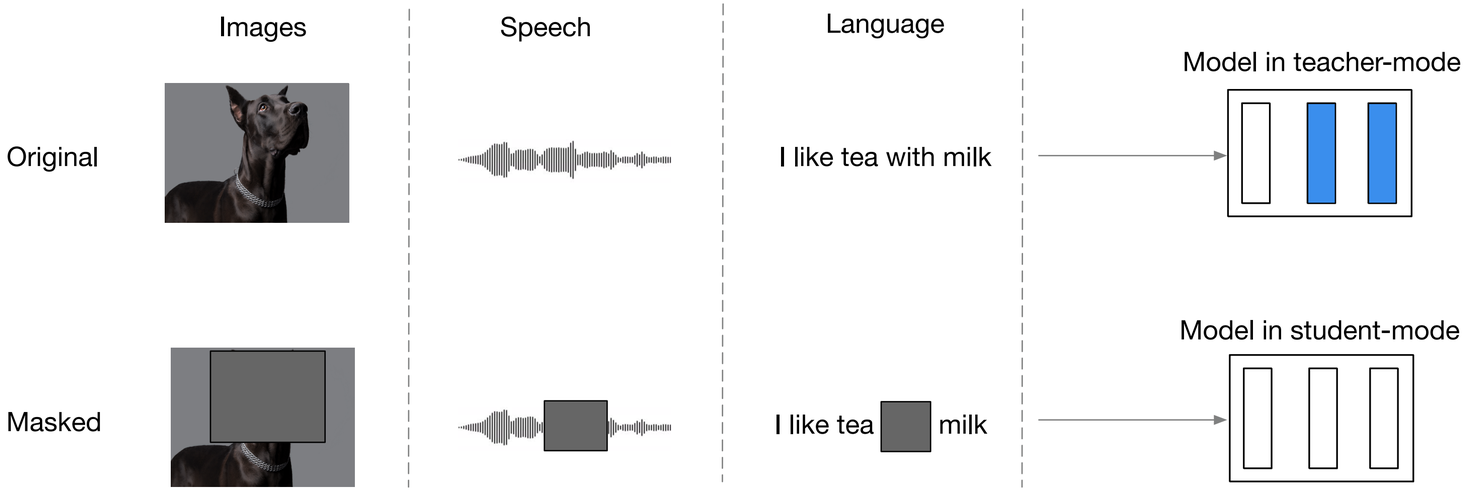
data2vec


Images
Speech
Language
Original
I like tea with milk
Masked
I like tea
milk

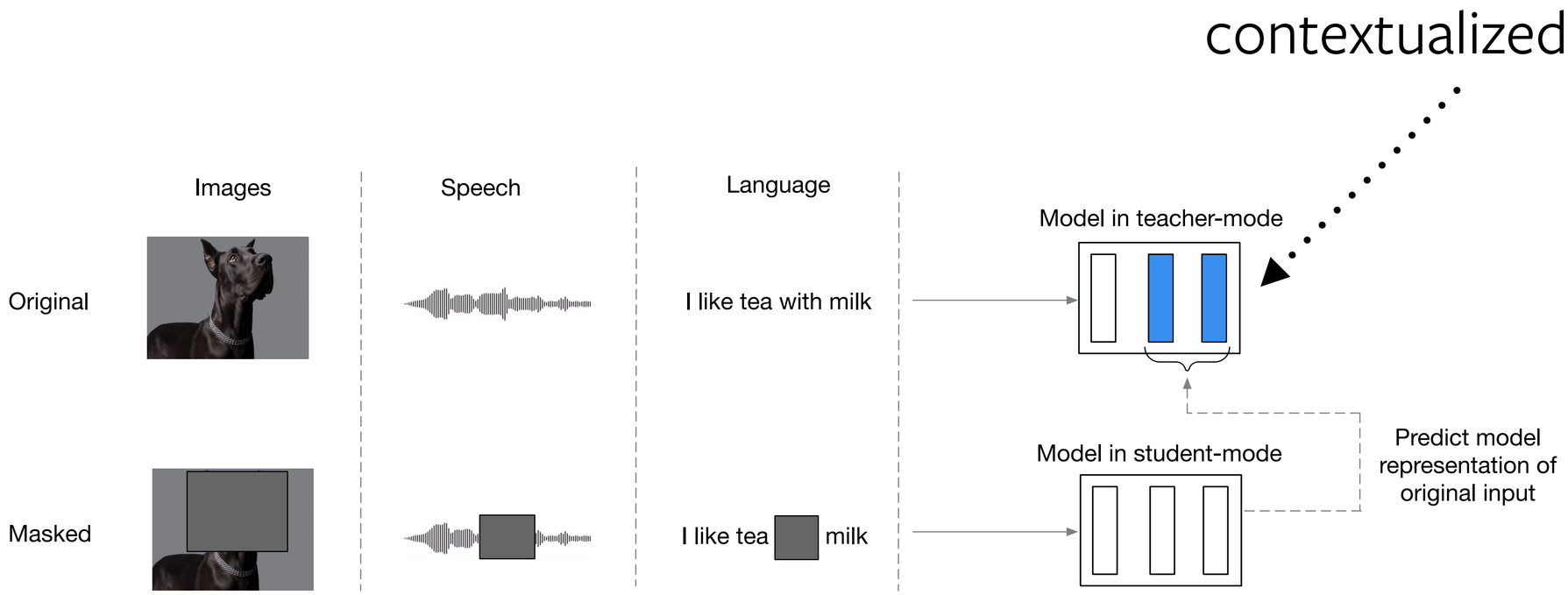
data2vec


Images
Speech
Language
Original
I like tea with milk
Masked
I like tea
milk
Model in student-mode
masked prediction

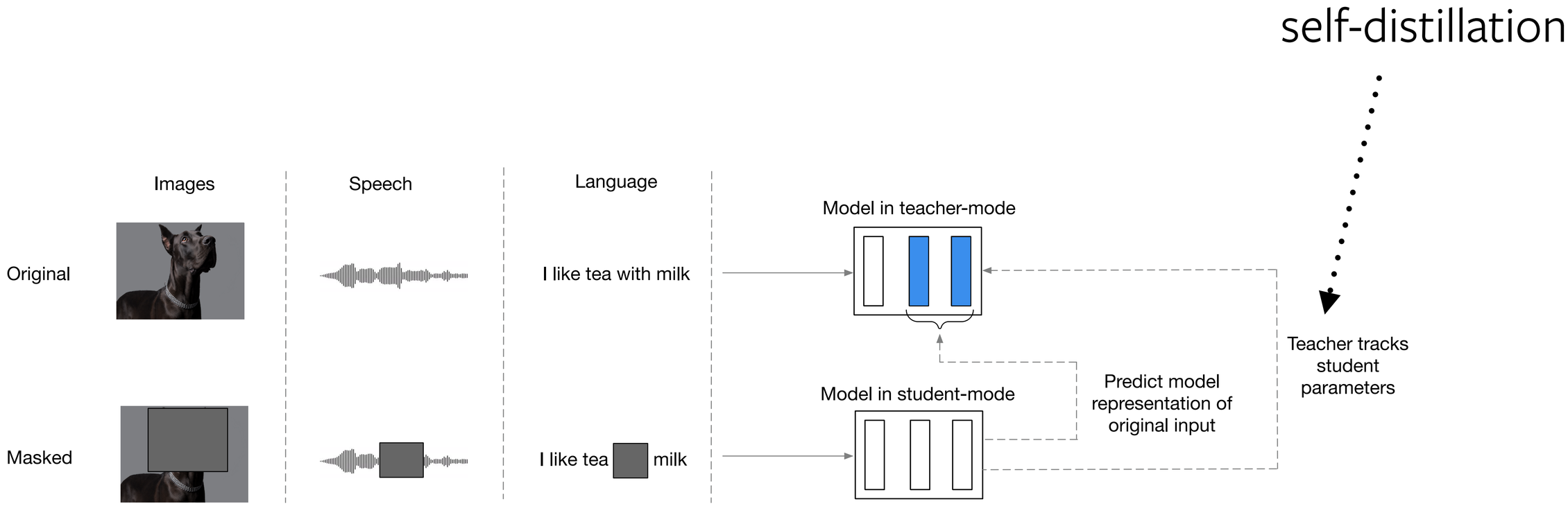
data2vec


Images
Speech
Language
Original
I like tea with milk
Model in teacher-mode
Masked
I like tea
milk
Model in student-mode

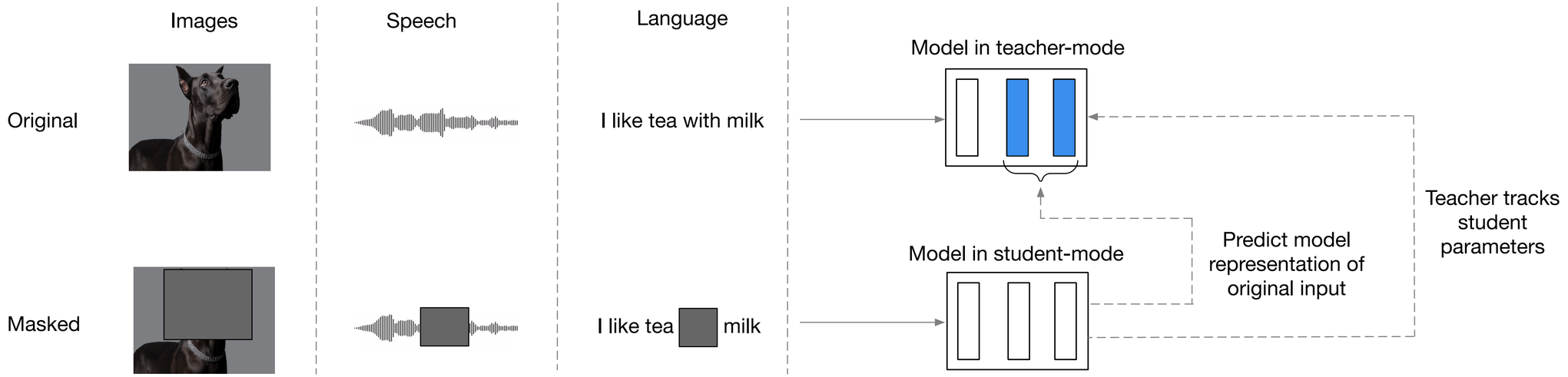
data2vec


contextualized targets
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

data2vec


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

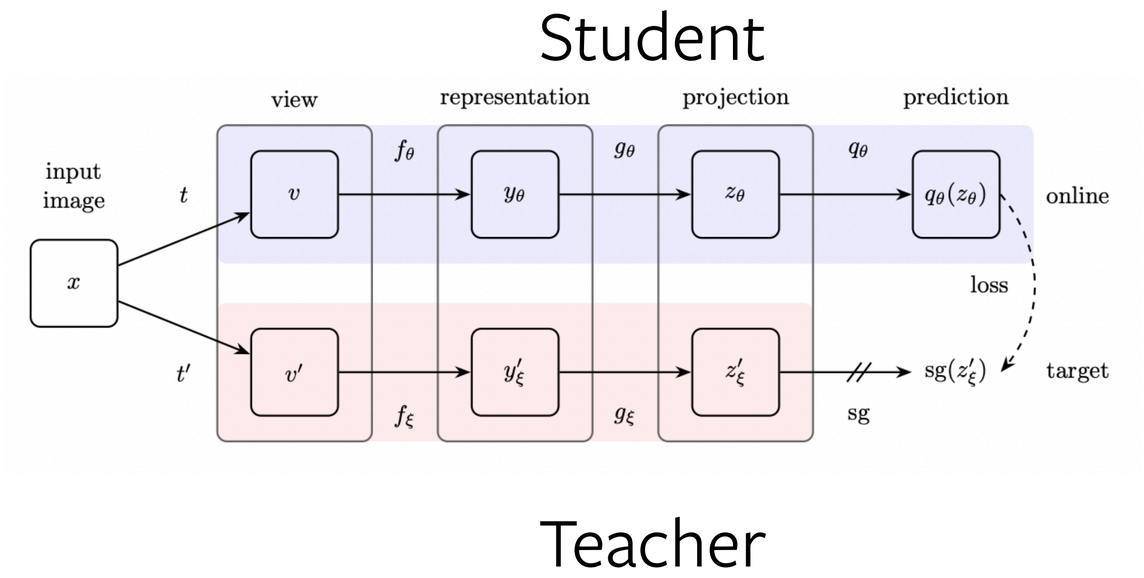
data2vec



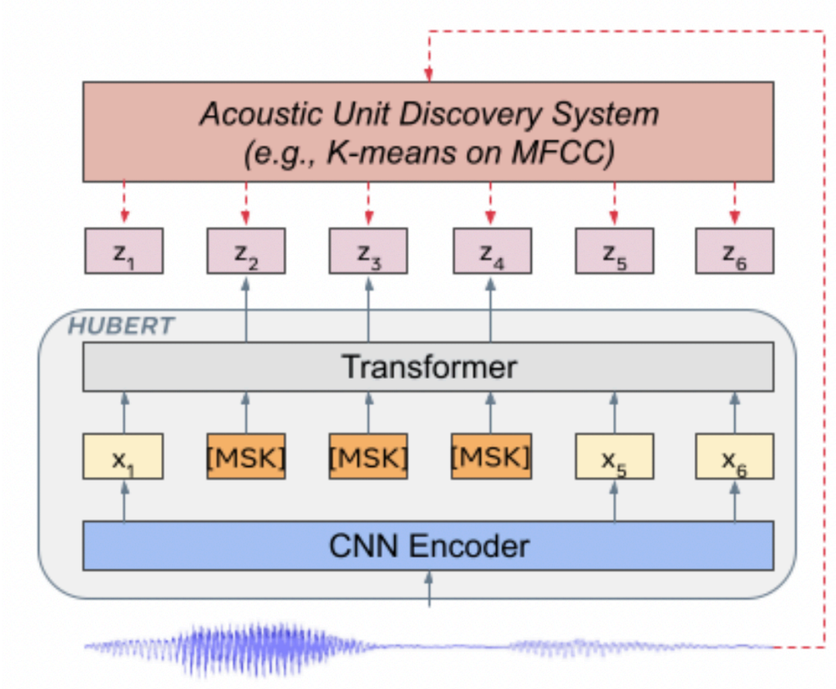
- Modality specific feature encoder (CNN, embedding table, patch mapping)
- Common masking policy, but modality/dataset specific parameterization
- Identical context encoder (Transformer)
- Identical learning task

Related Work

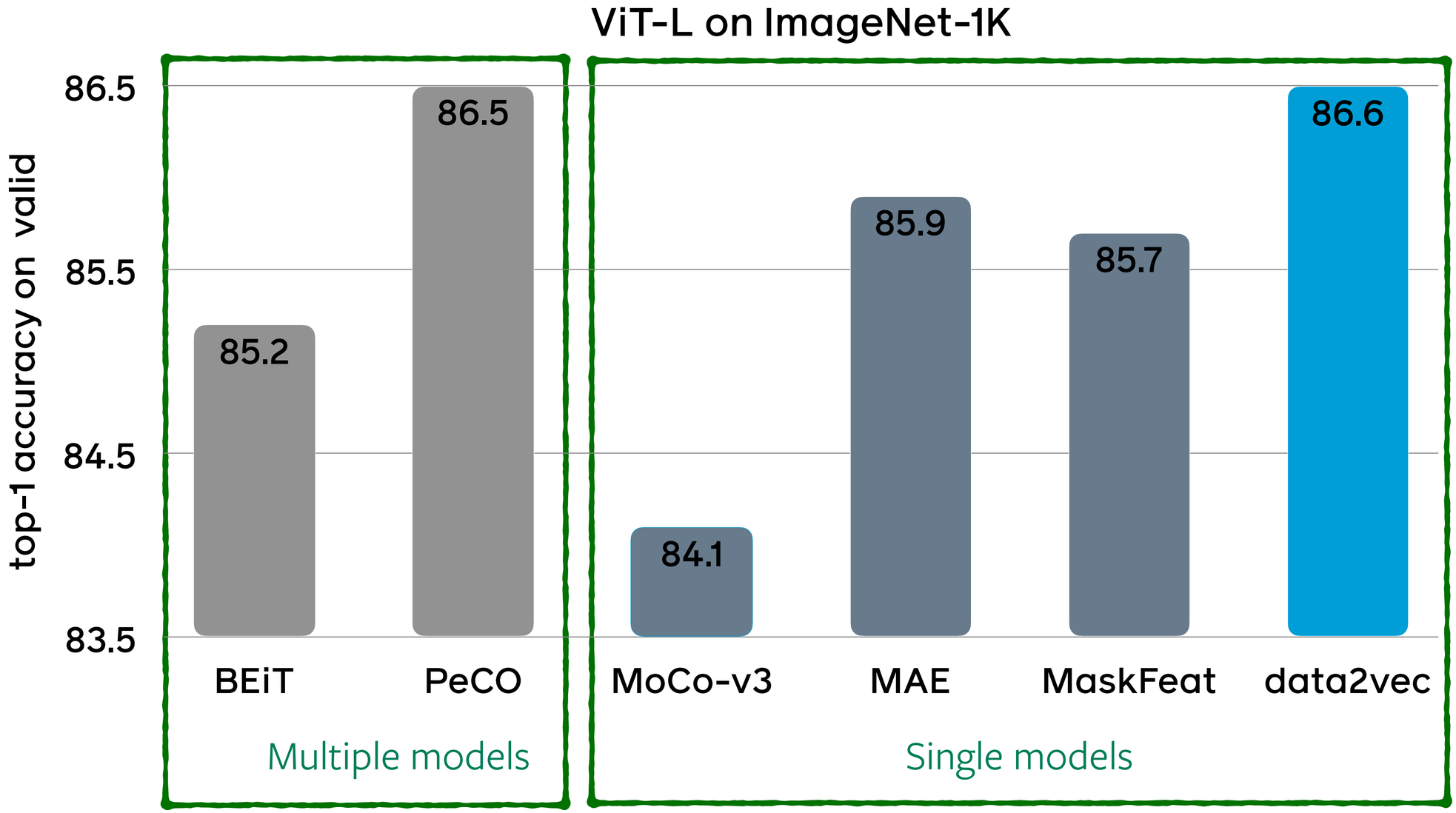
- Momentum teacher
  (Grill et al., ′20, Caron et al., ′21)



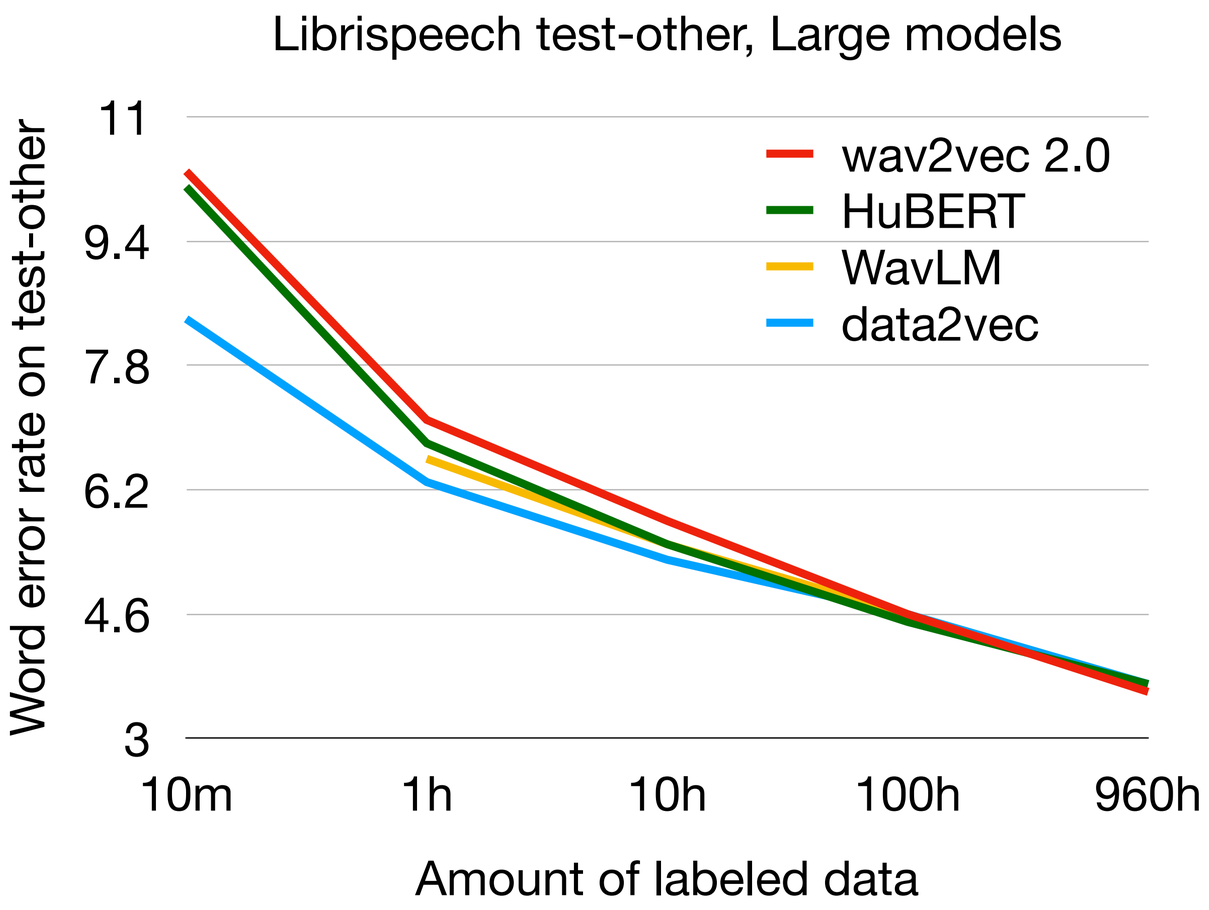
- Contextualized targets
  (Hsu et al., ′21)

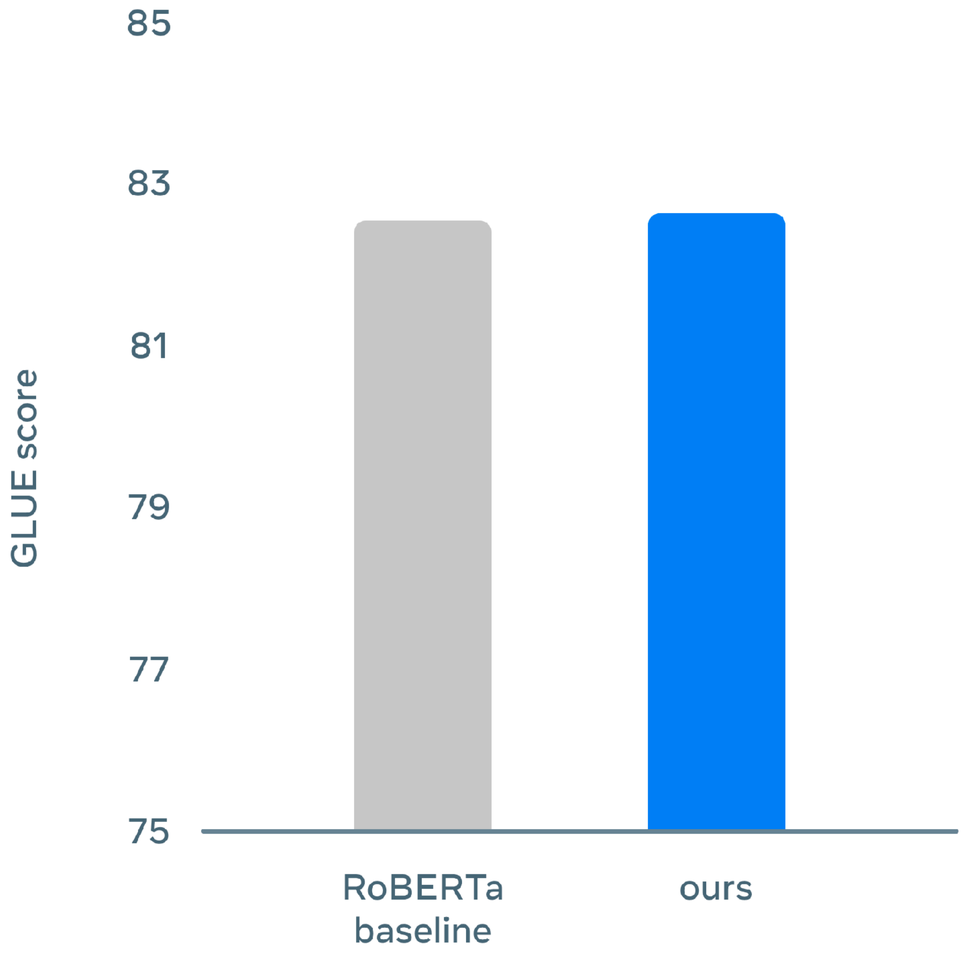


Vision Results

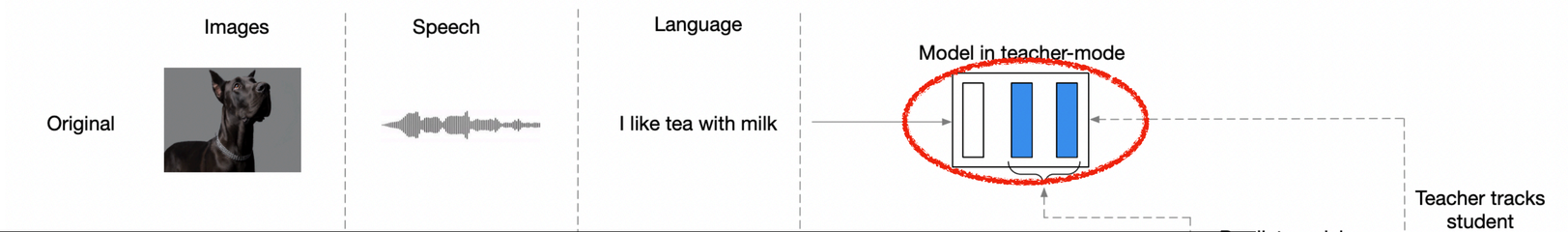


Speech & NLP Results

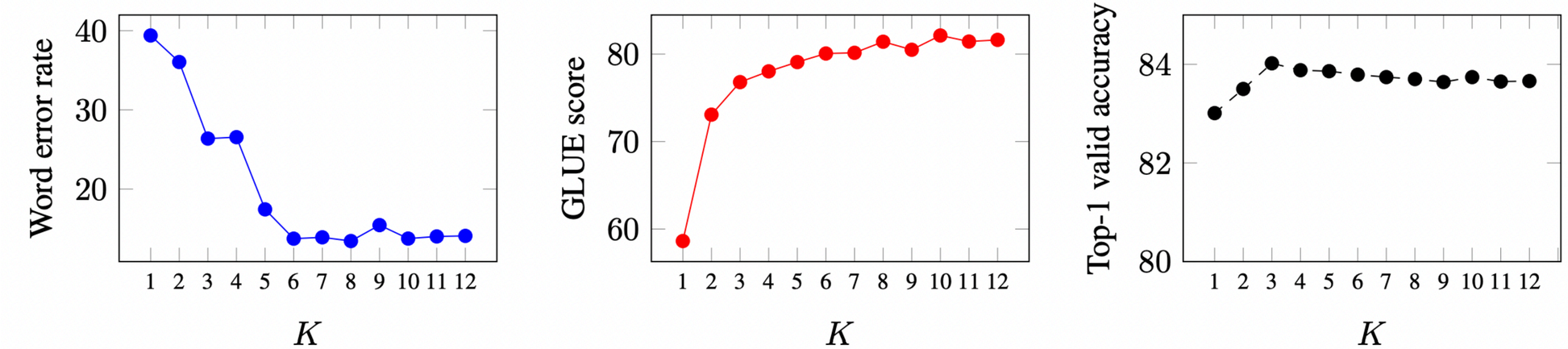




Teacher Representation Construction

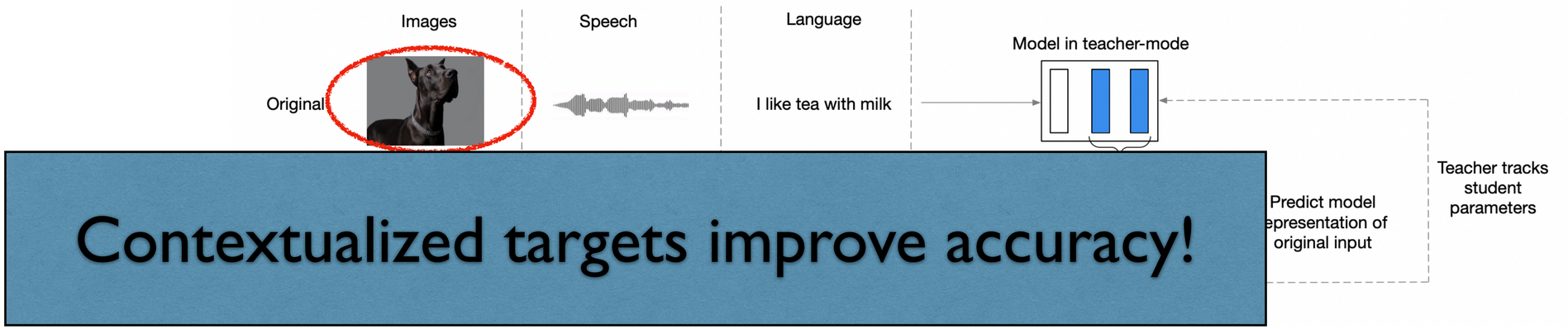


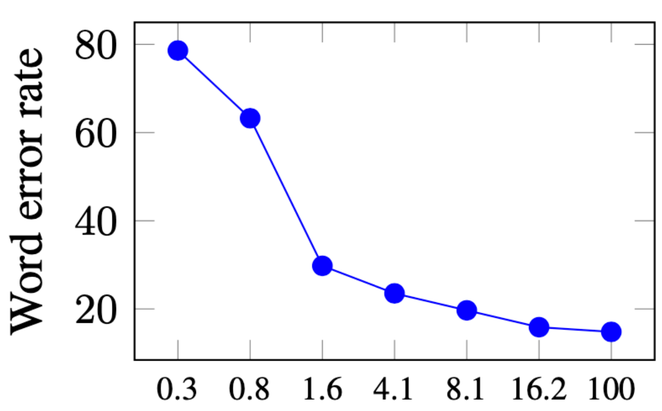
Averaging multiple layers helps!



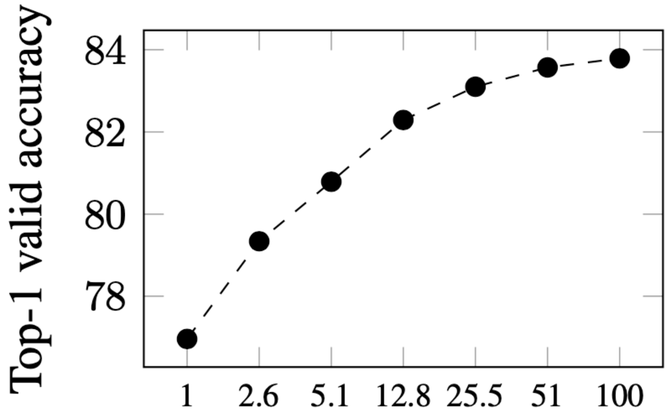
(a) Speech (b) NLP (c) Vision

Target Context Size





(a) Speech



(b) Vision

Limitations

- Modality specific feature encoder -> Perceiver work!
- Requires two forward-passes -> data2vec 2.0

Efficient Self-supervised Learning

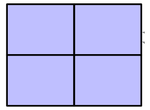
data2vec 2.0

- MAE: Do not encode masked time-steps.
- Multi-masking: Learn from different views & share target representation.
  - Amortizes the cost of the teacher.
- Result: train with less compute, fewer epochs & smaller batch size.

Example

Images

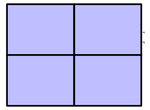
Teacher encoder



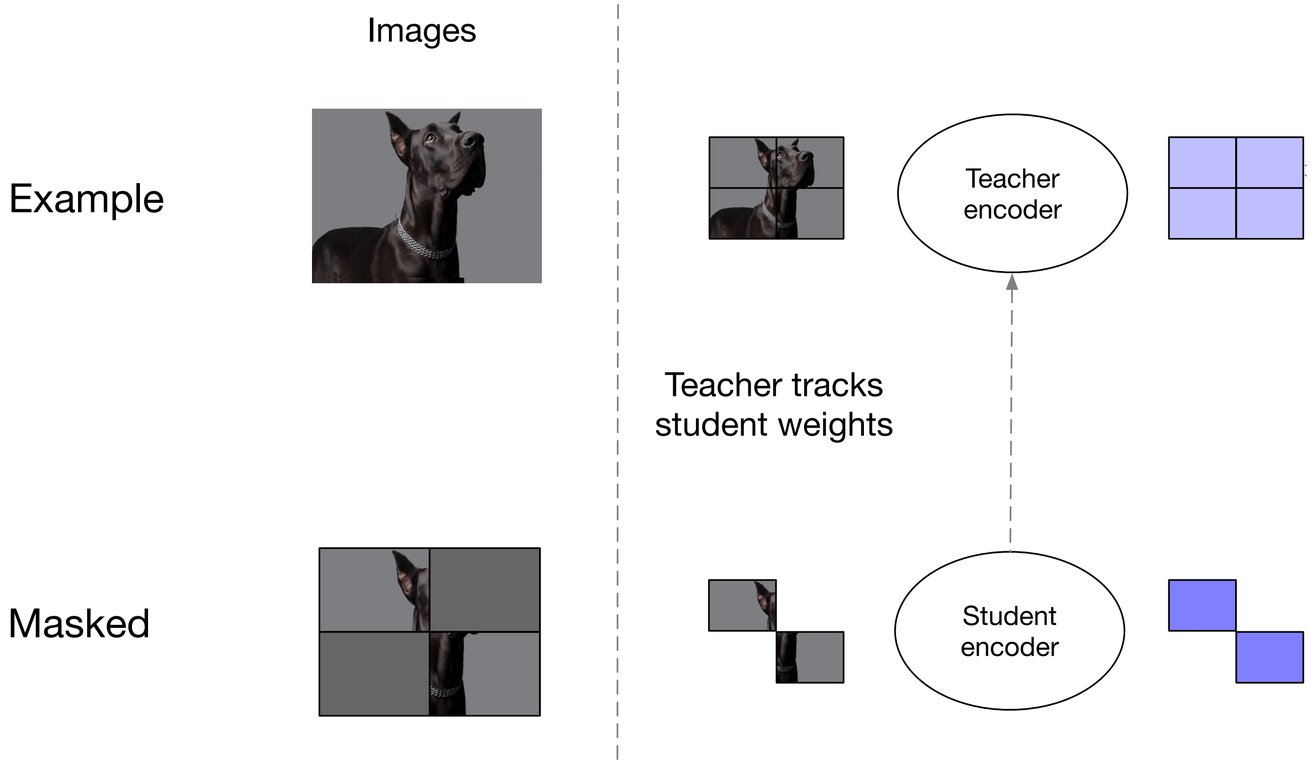
Images

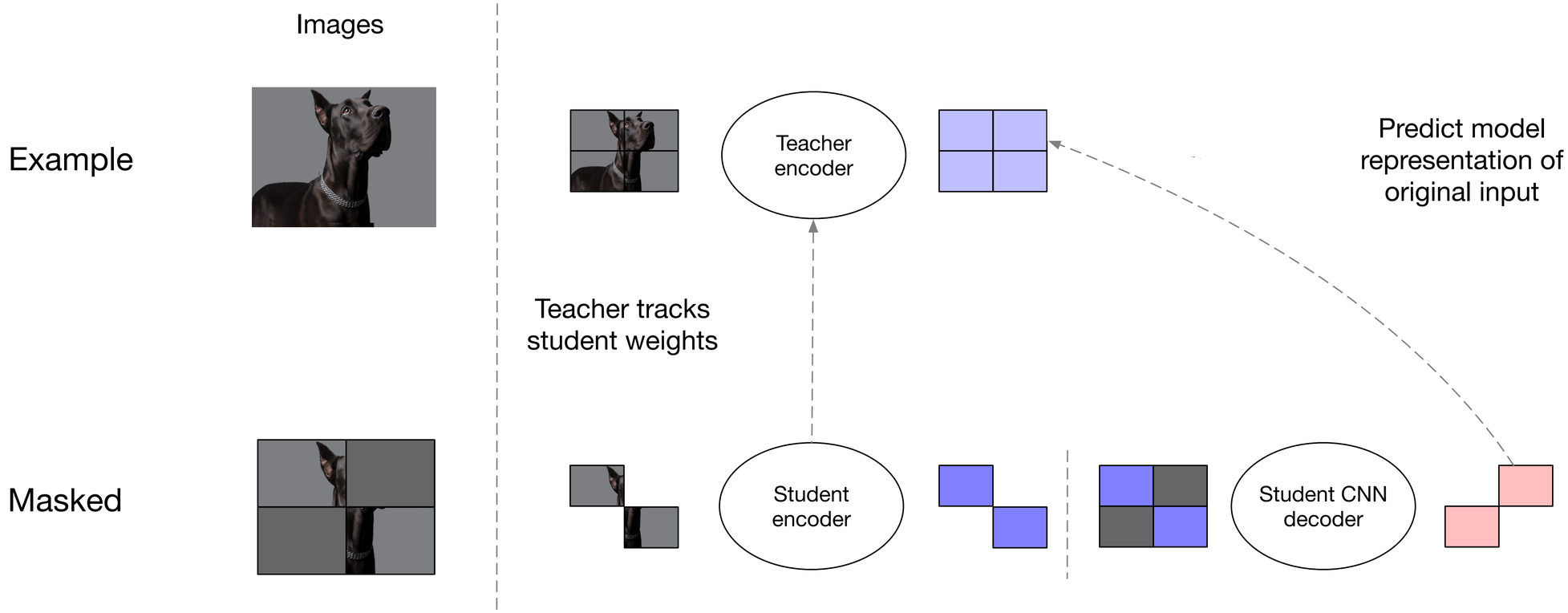
Example

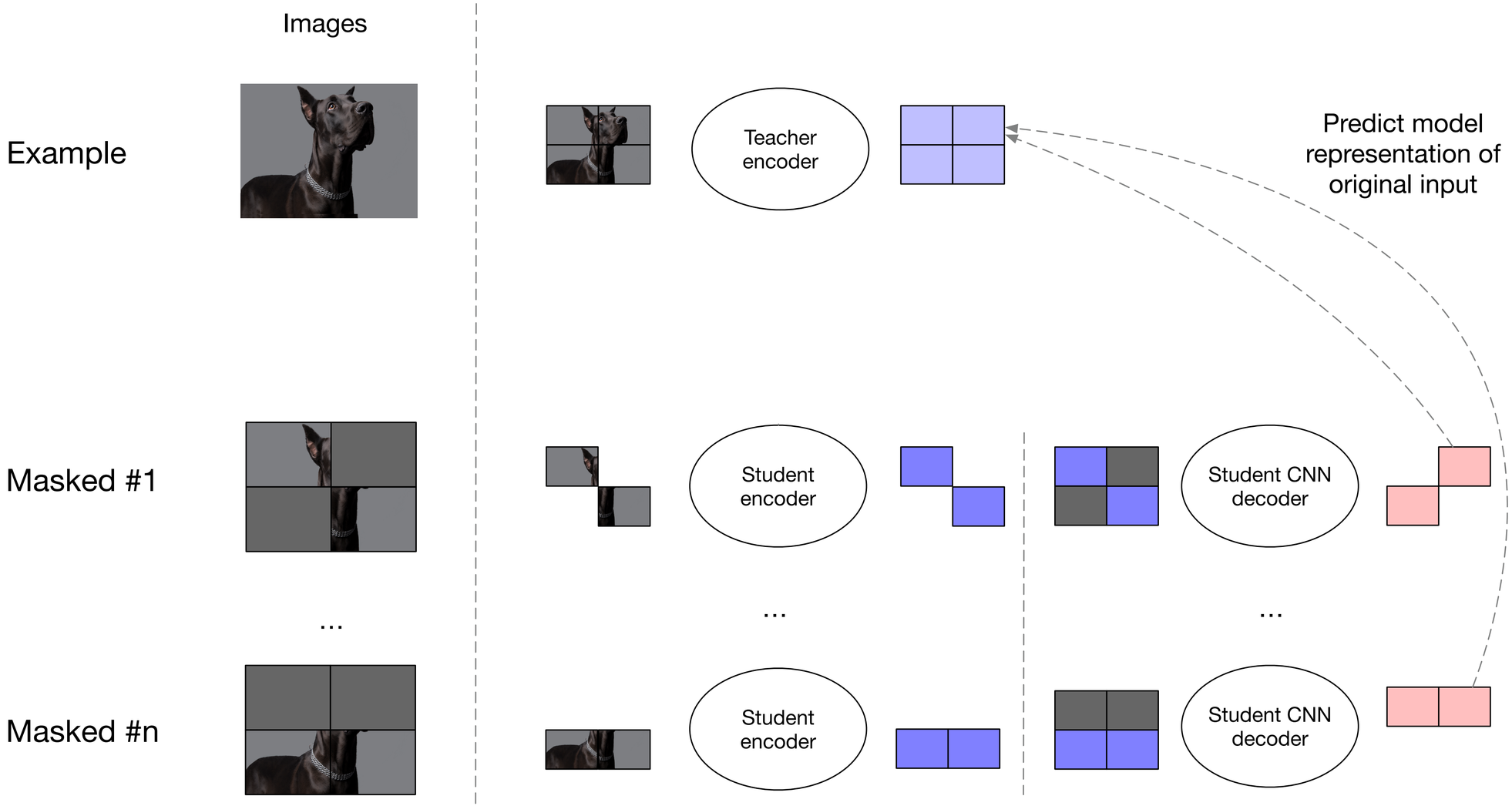
Teacher encoder

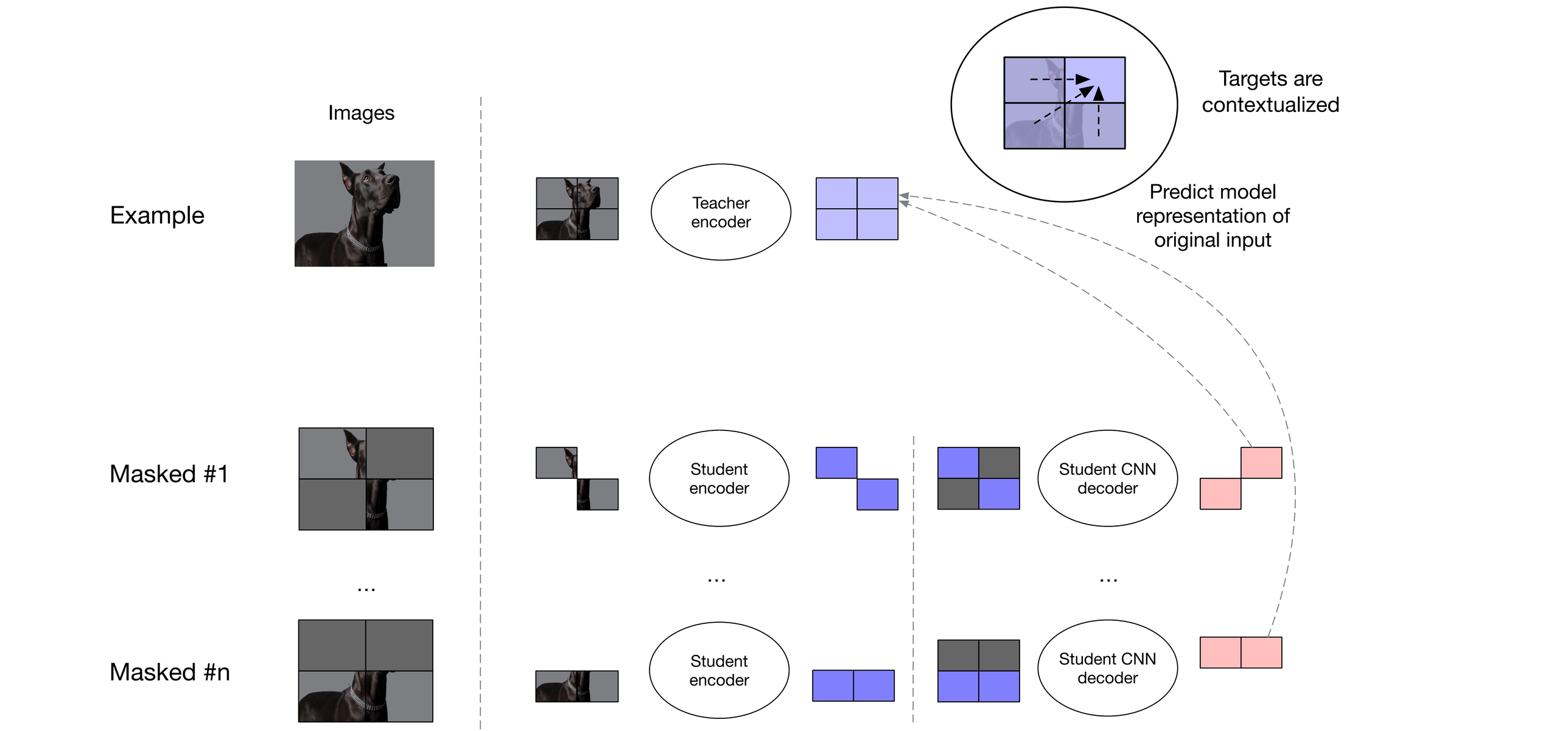


Masked

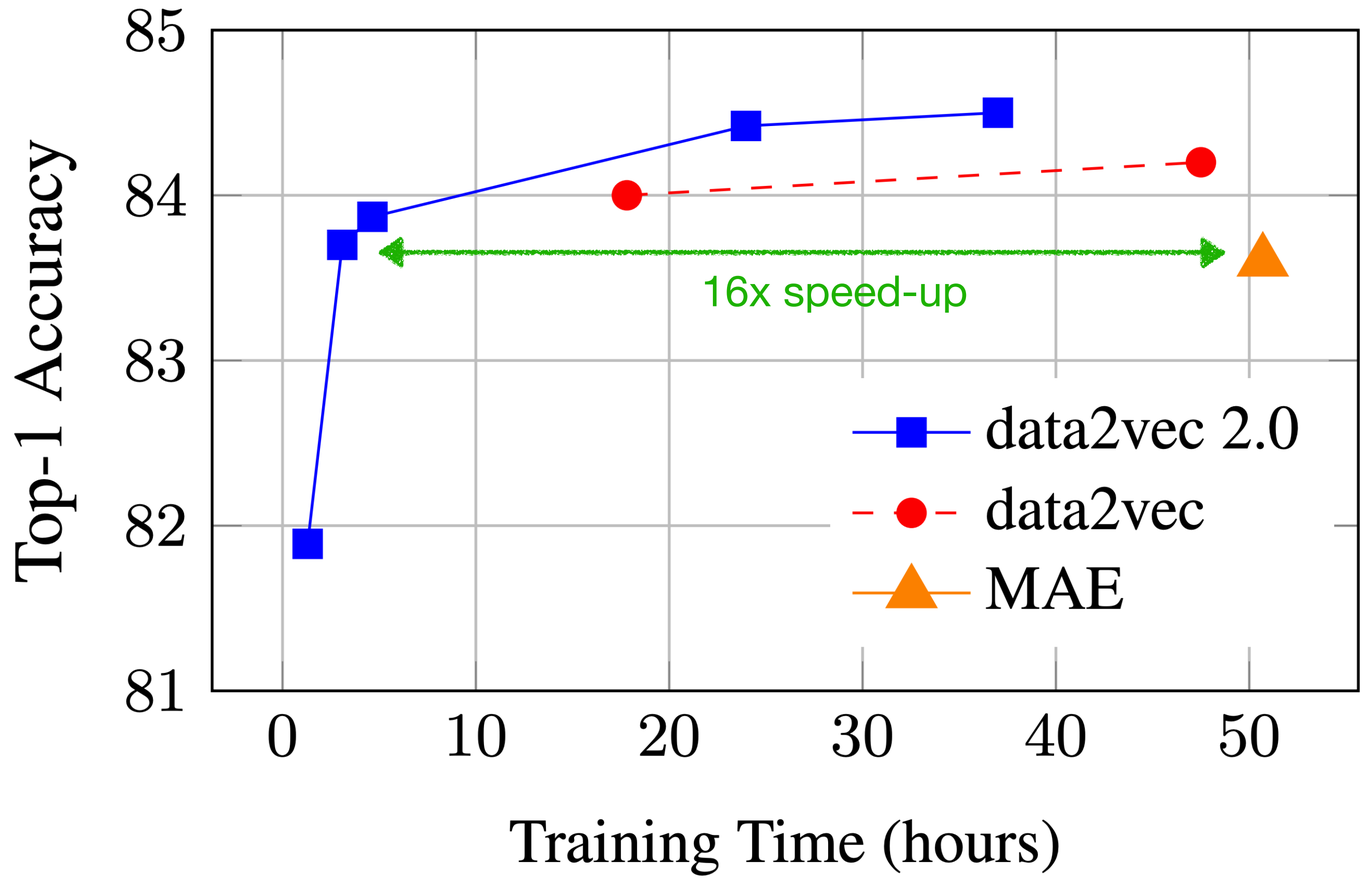

Images
Example
Masked
Teacher
encoder
Teacher tracks
student weights
Student
encoder


Images
Example
Masked
Teacher encoder
Predict model representation of original input
Teacher tracks student weights
Student encoder
Student CNN decoder


Images
Example
Masked #1
...
Masked #n
Teacher encoder
Predict model representation of original input
Student encoder
Student CNN decoder
...
...
Student encoder
Student CNN decoder


Images
Example
Masked #1
Masked #n
Teacher encoder
Student encoder
Student encoder
Student CNN decoder
Student CNN decoder
Targets are contextualized
Predict model representation of original input

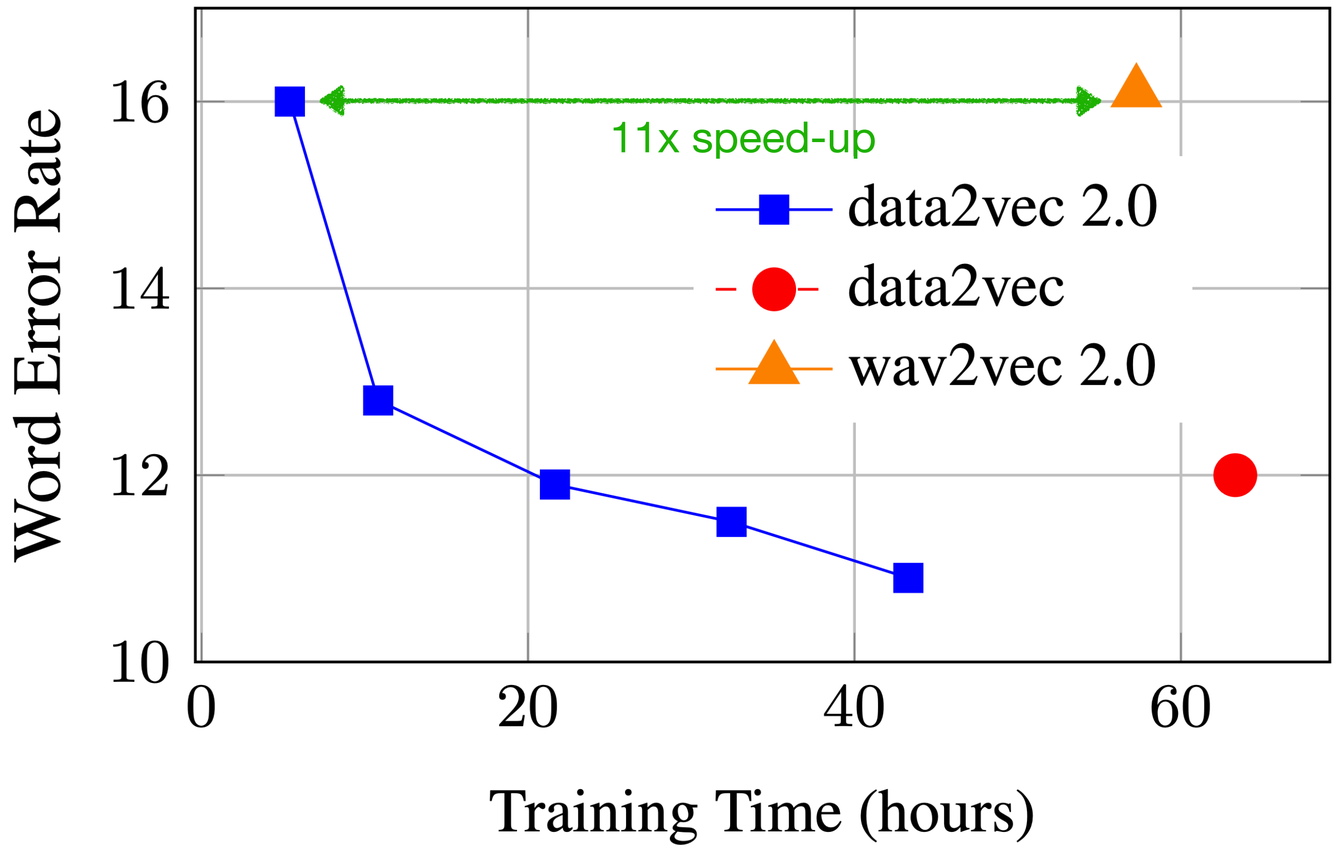
Compute Efficiency in Vision



	MAE	data2vec 2.0
Train time (h)	50.7	3.1
Epochs	1600	20
Batch size	4,096	512
Accuracy	83.6	83.7

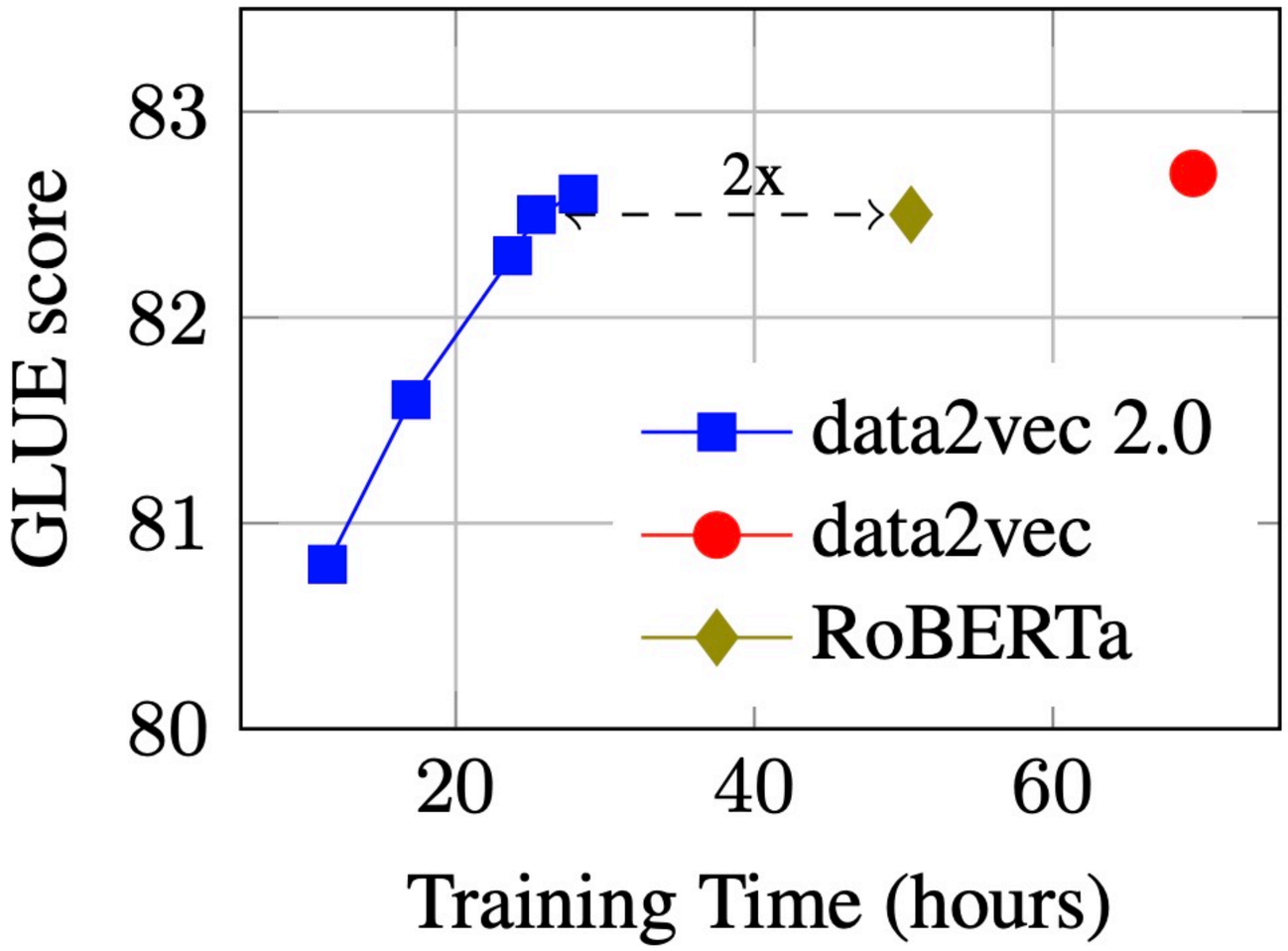
ViT-B, pre-train and fine-tune on ImageNet-1K, eval on dev
All training times are for 32 A100 GPUs

Compute Efficiency in Speech

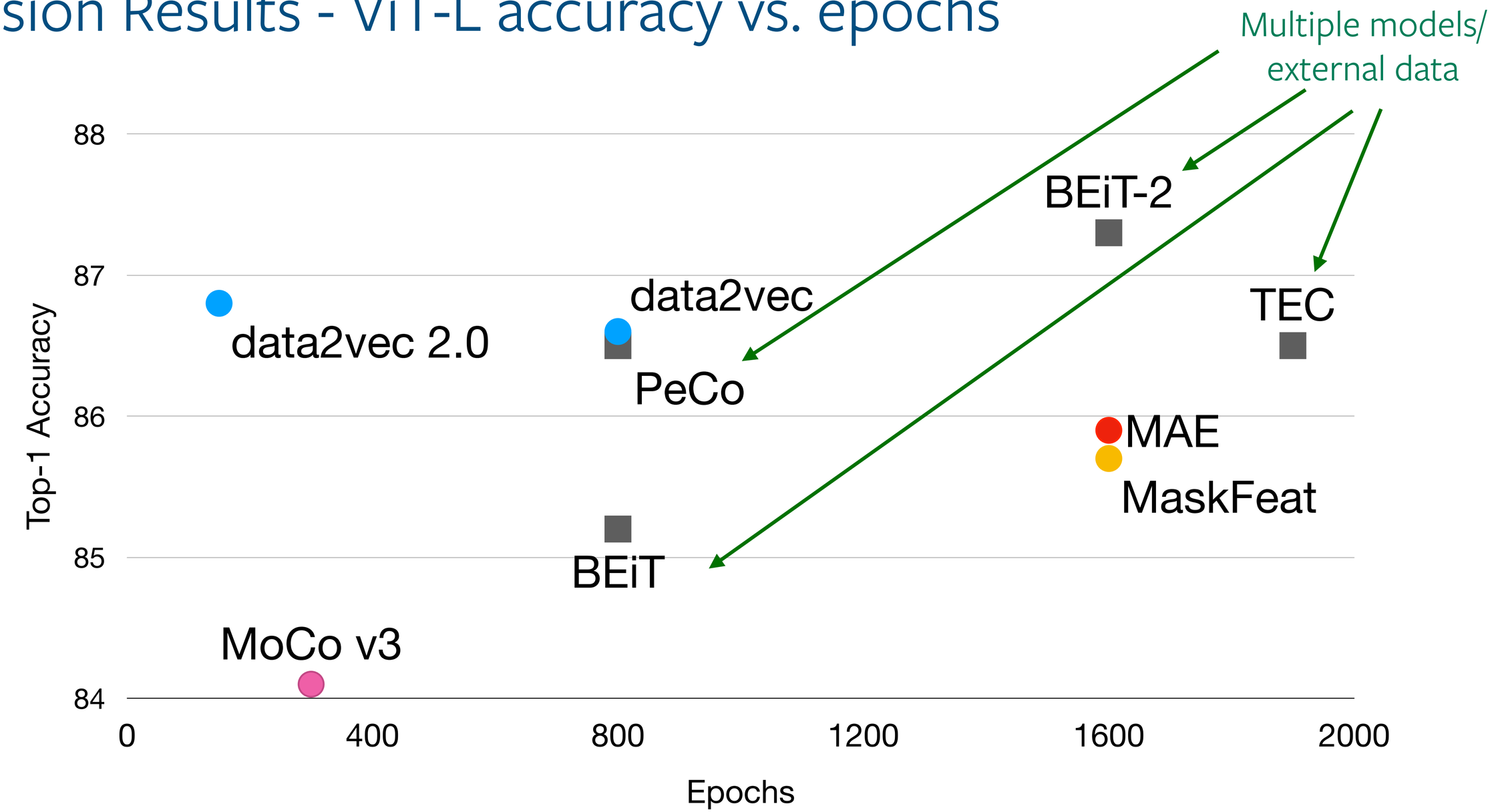


Transformer Base, pre-train on Librispeech,
fine-tune on Libri-light 10h, eval on dev-other, no language model

Compute Efficiency in NLP



Vision Results - ViT-L accuracy vs. epochs



Conclusion

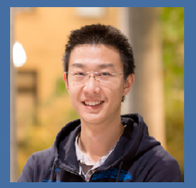
- A single learning objective can perform very well compared to the best modality-specific algorithms for vision/speech/NLP.
- Contextualized targets lead to a rich learning task which enables efficient training.
- Think about multiple modalities from the outset.

Thank you


Arun Babu


Alexis Conneau


Steffen Schneider


Henry Zhou


Abdelrahman Mohamed


Jiatao Gu

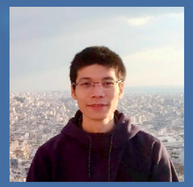

Naman Goyal


Wei-Ning Hsu


Alexei Baevski


Michael Auli


Kushal Lakhotia


Andros Tjandra


Kritika Singh


Yatharth Saraf

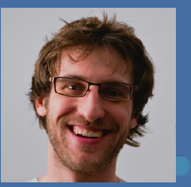

Geoffrey Zweig


Qiantong Xu


Tatiana Likhomanenko


Paden Tomasello


Ronan Collobert


Gabriel Synnaeve