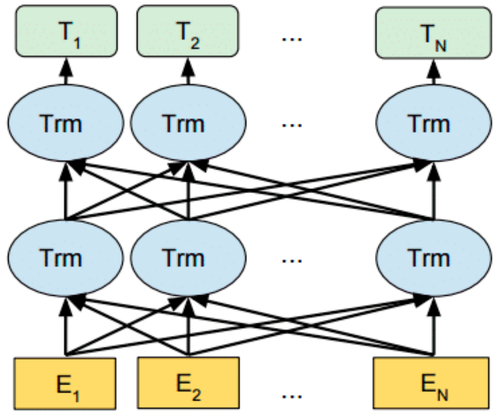
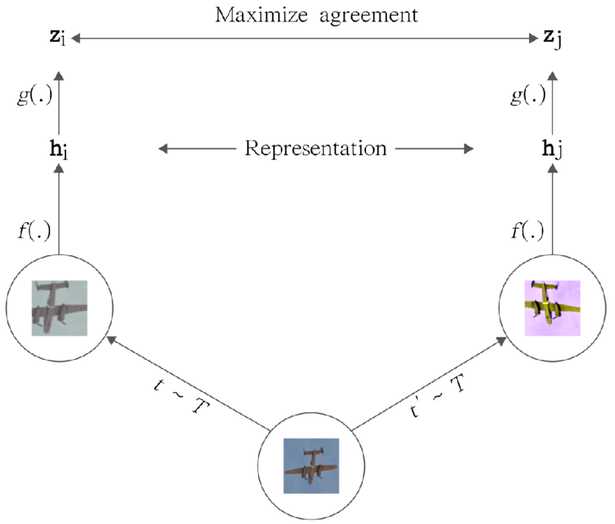
data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language

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

Current state of self-supervised learning

- NLP: masked prediction (BERT)
- Vision: contrastive, data aug, input reconstruction (MoCo, SimCLR, BYOL, DINO, MAE, ...)
- Speech: similar to vision (wav2vec, CPC, Hubert, TERA, ...)





Current state of self-supervised learning

- Many different algorithms
- Most algorithms developed for particular modality
- Little focus on algorithms that generalize across modalities

data2vec

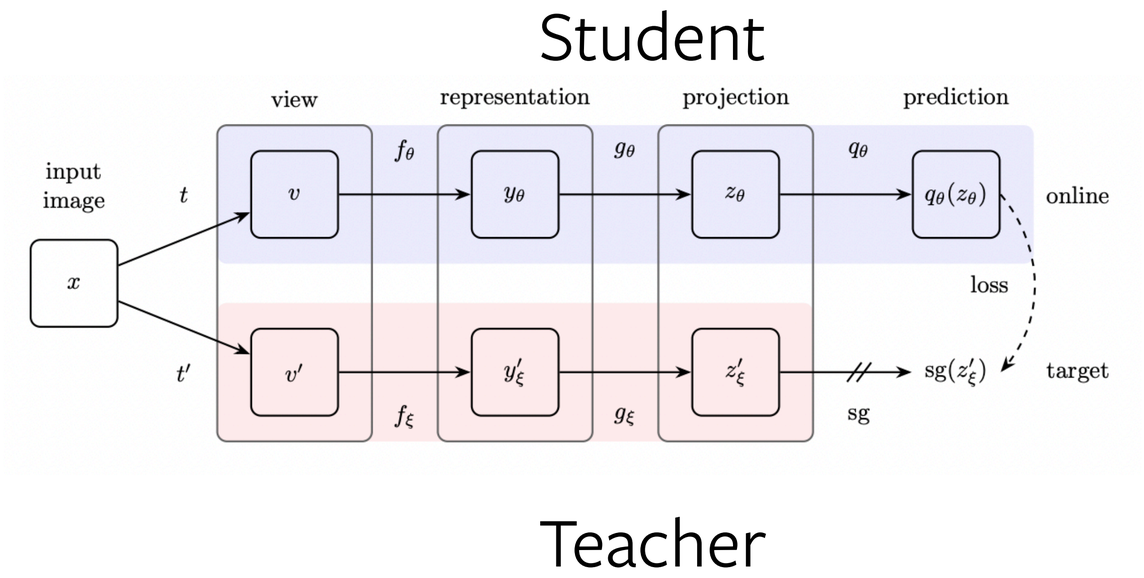
- General algorithm that works very well across modalities
  (Outperforms best algorithms in speech/vision and competitive in NLP)

- Same learning objective for each modality

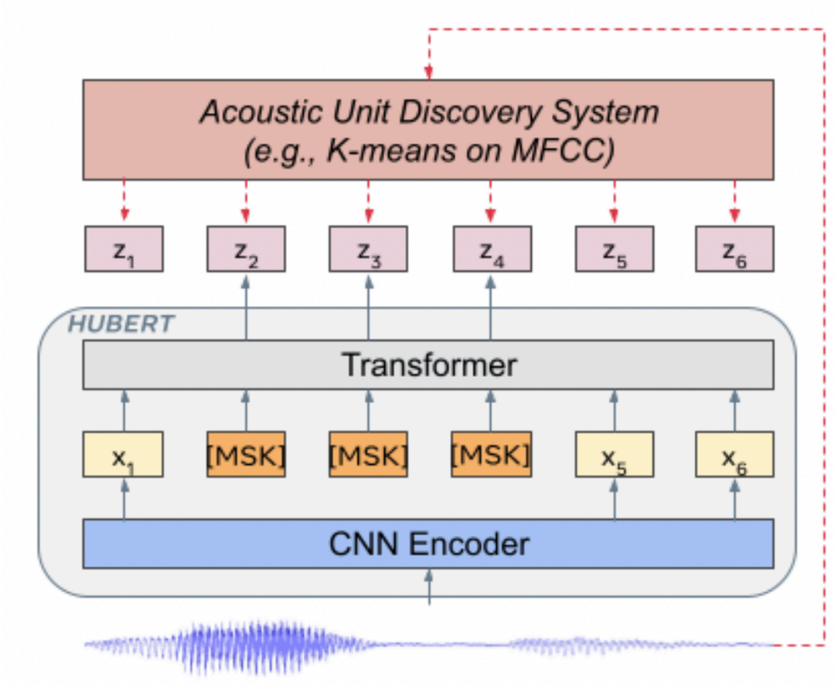
- Idea: self-distillation of contextualized representations in a masked prediction setup

Related work

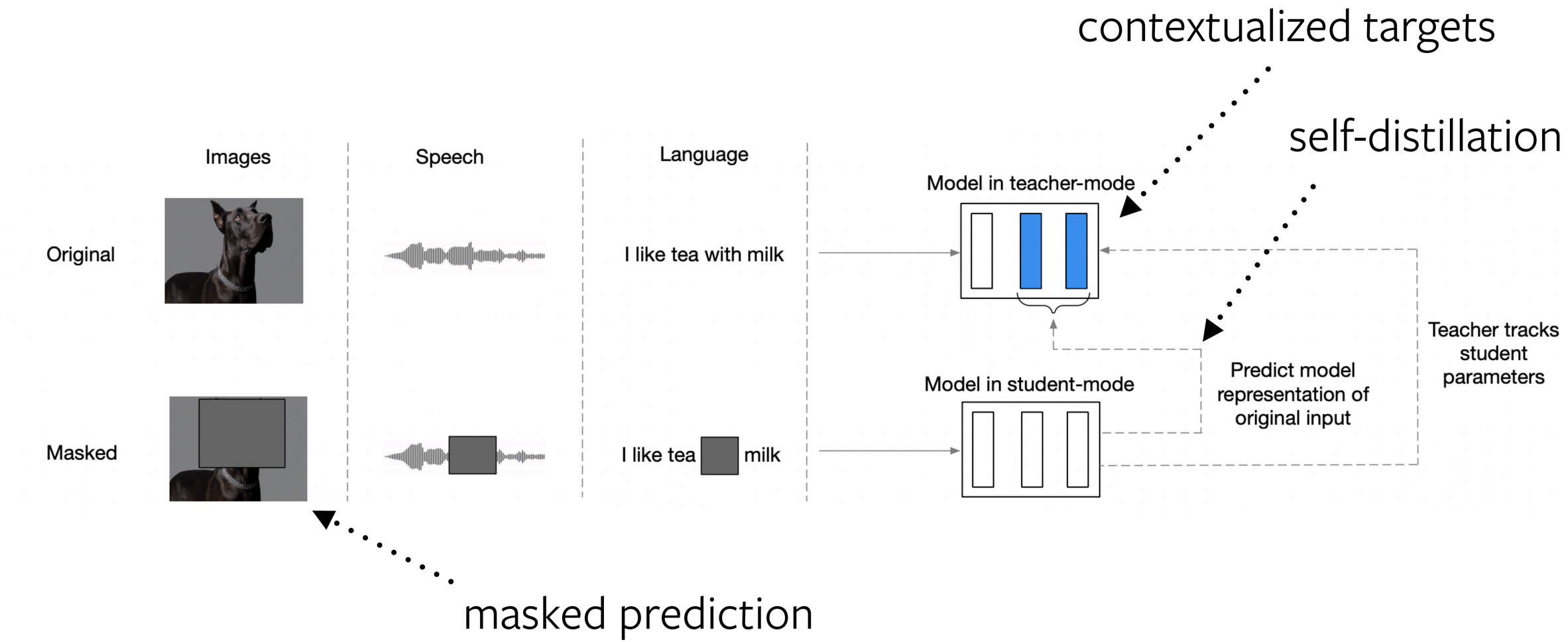
- Momentum teacher - BYOL, DINO (Grill et al., ′20, Caron et al., ′21)



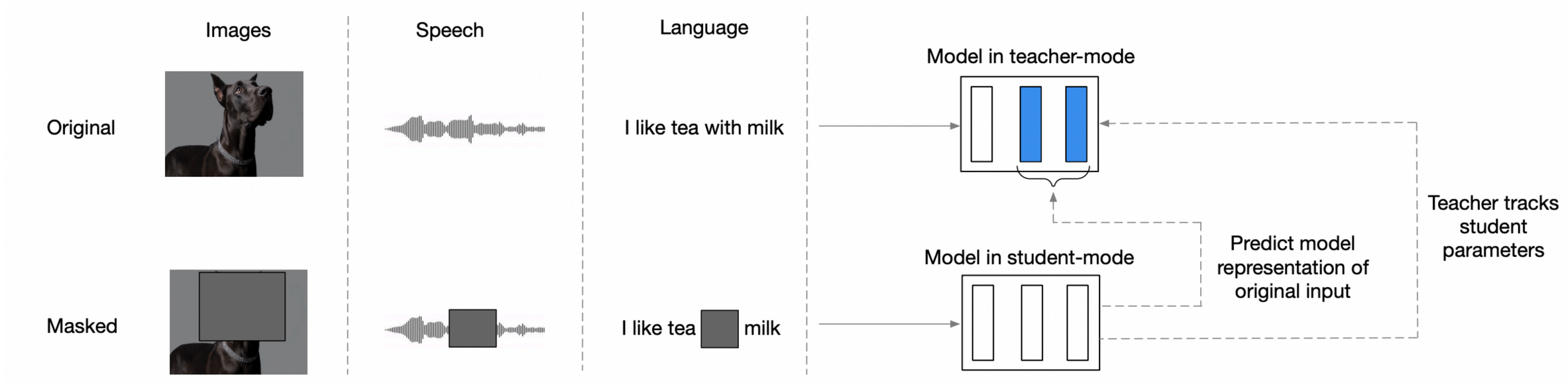
- Contextualized targets - HuBERT (Hsu et al., ′21)



data2vec

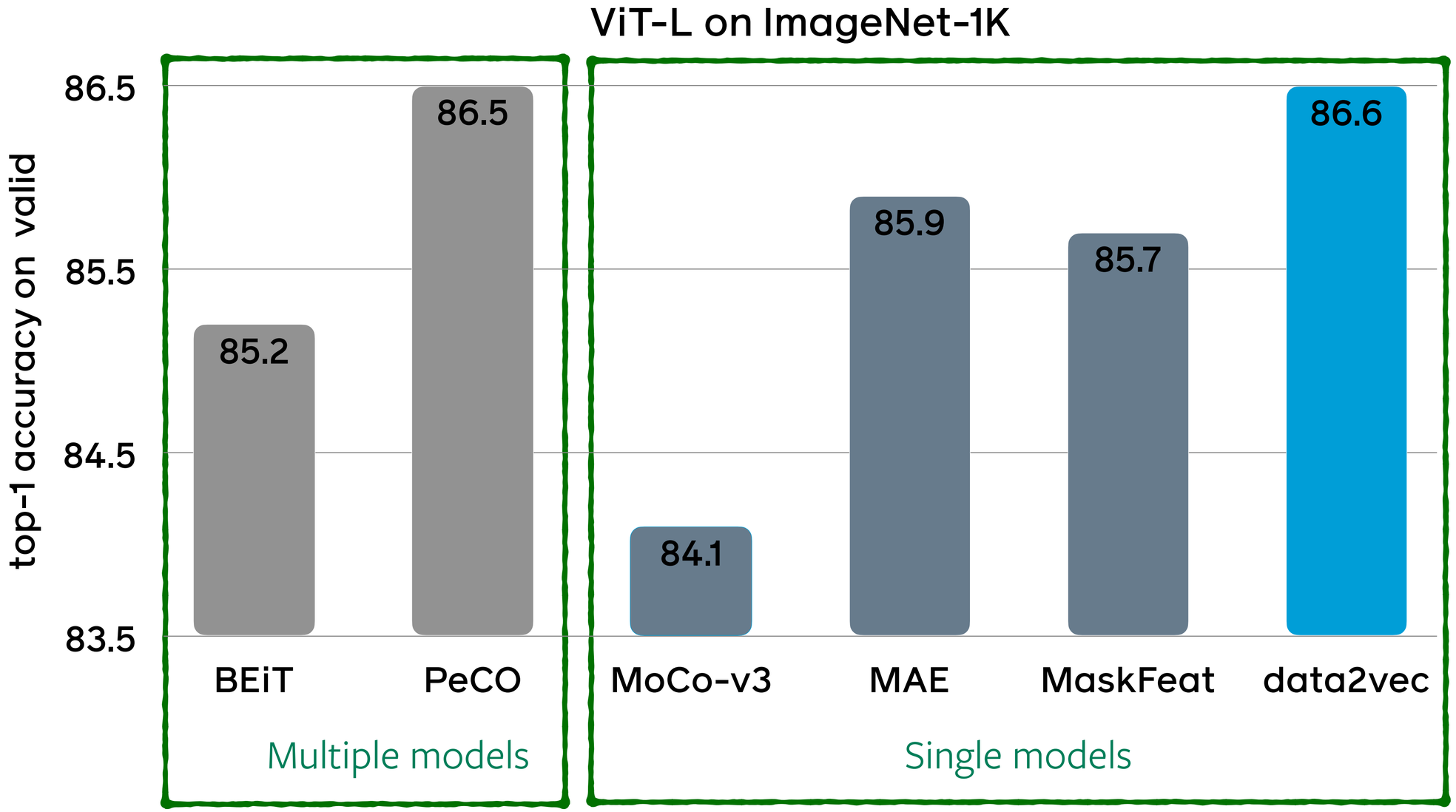

contextualized targets
self-distillation
masked prediction
Images
Speech
Language
Original
Masked
I like tea with milk
I like tea
milk
Model in teacher-mode
Model in student-mode
Predict model representation of original input
Teacher tracks student parameters

data2vec

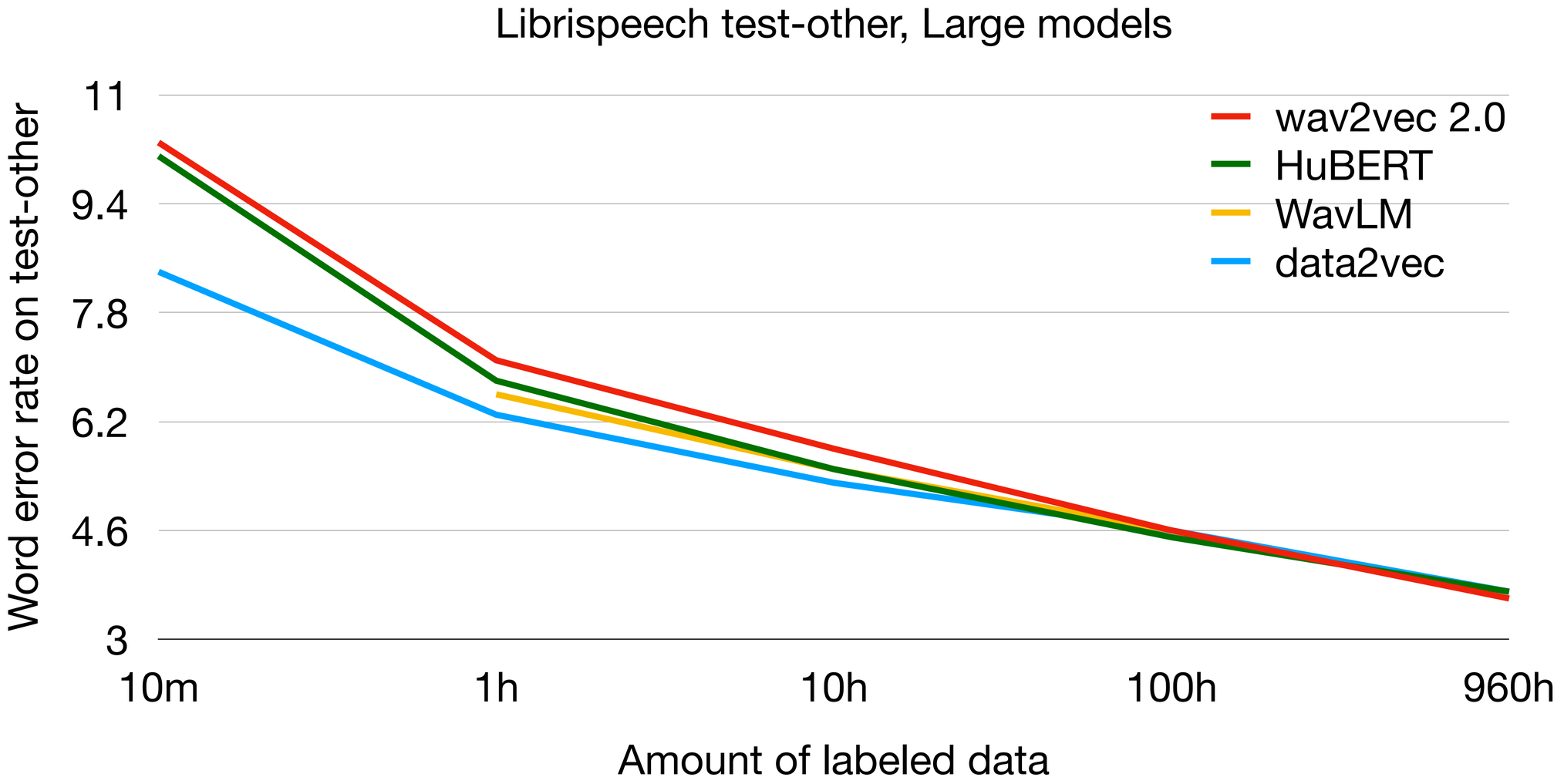


- Modality specific feature encoder (CNN, embedding table, patch mapping)
- Common masking policy, but modality/dataset specific parameterization
- Identical context encoder (Transformer)
- Identical learning task

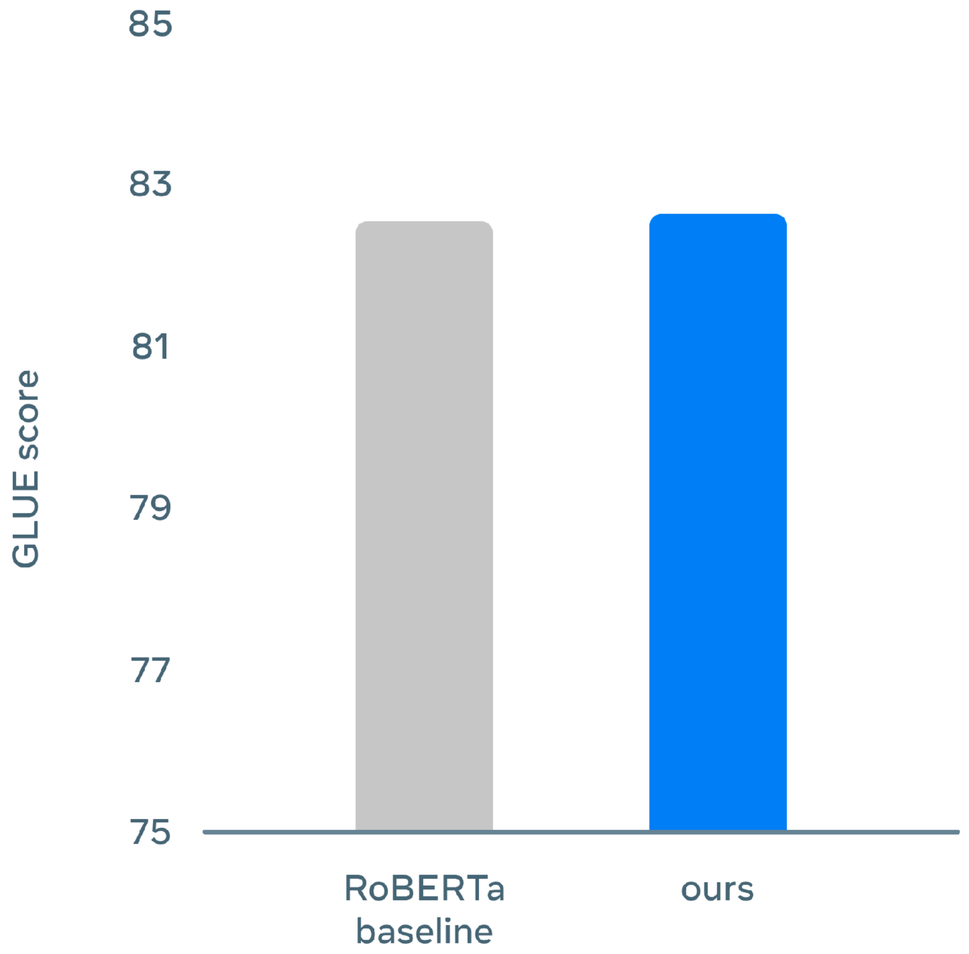
Vision Results



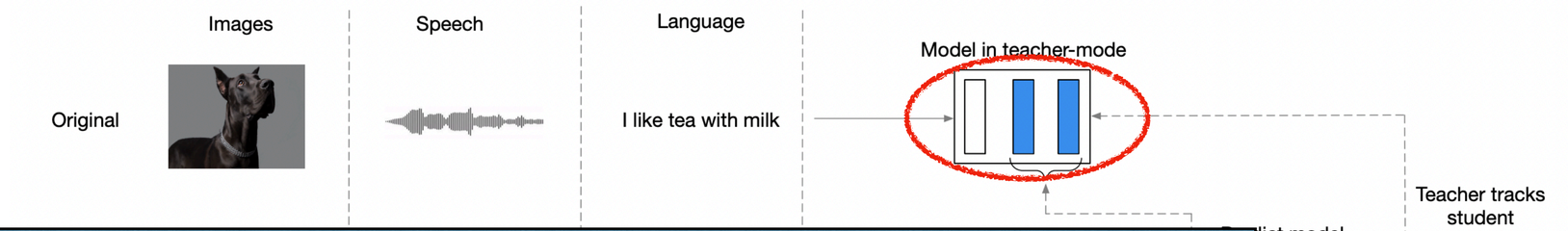
Speech Results



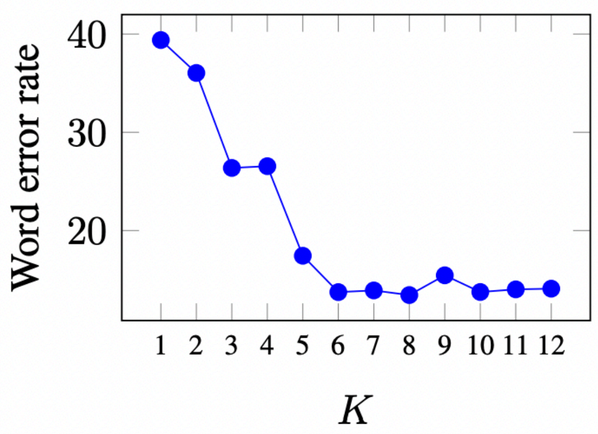
NLP Results



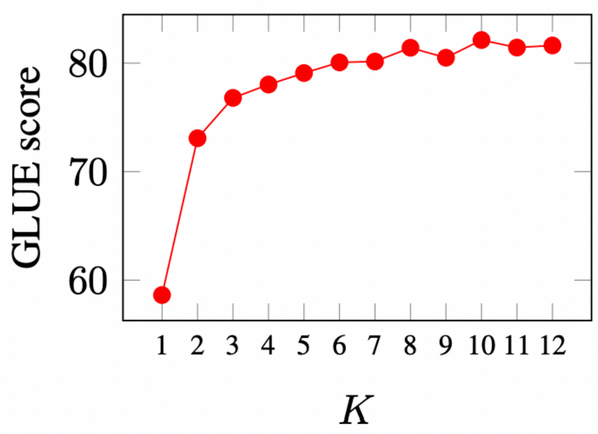
Teacher representation construction



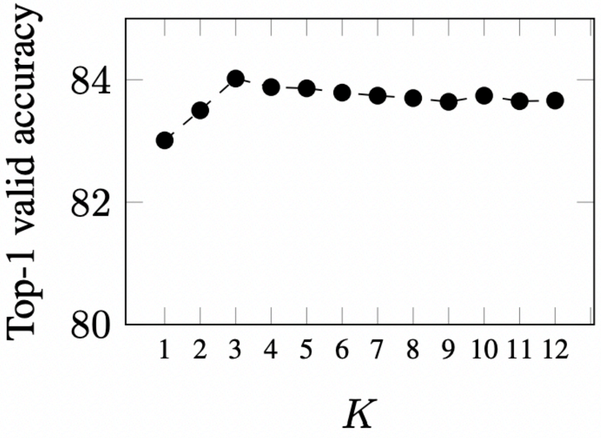
Averaging multiple layers helps!



(a) Speech

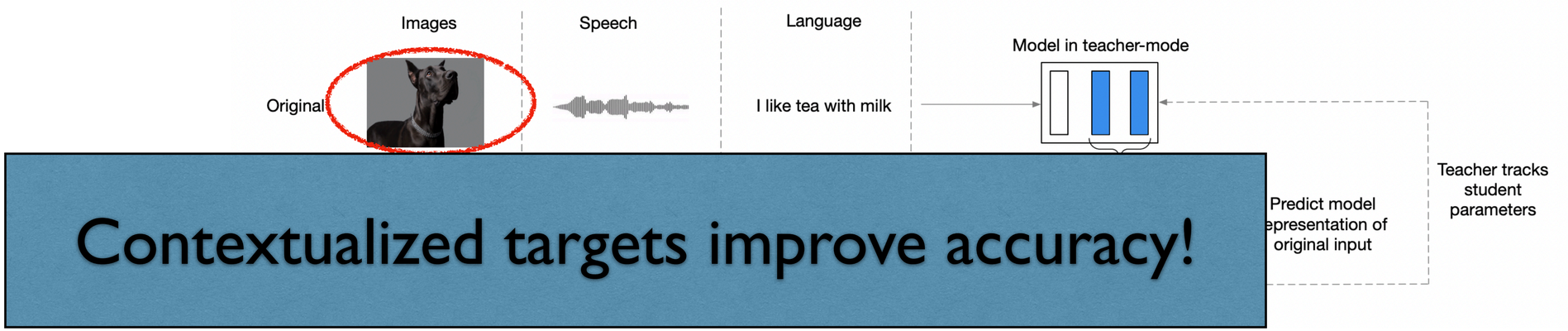


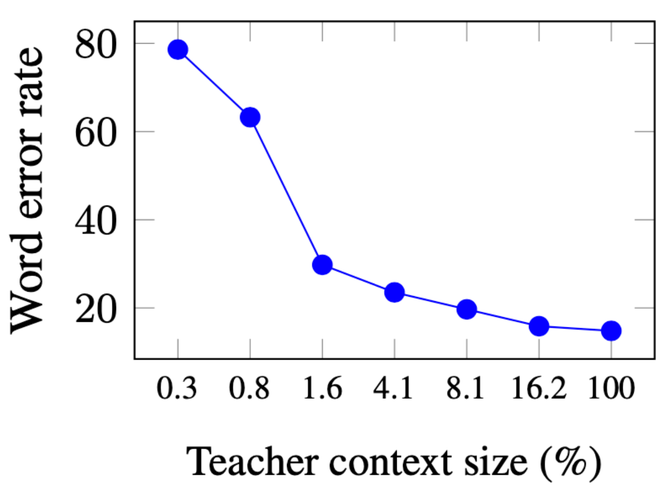
(b) NLP



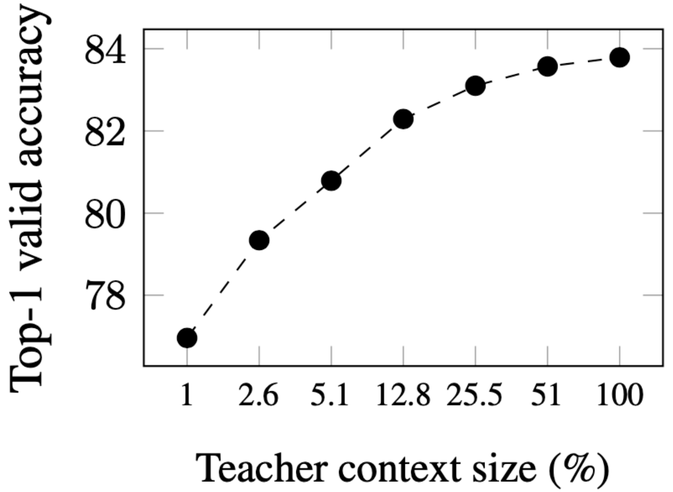
(c) Vision

Target context size





(a) Speech



(b) Vision

Limitations

- Modality specific feature encoder and masking parameters
- Requires two forward-passes

Conclusion

- A single learning objective can outperform the best modality-specific algorithms for vision/speech while being competitive on NLP
- Target representations based on large context windows and from multiple layers lead to a richer SSL task and improve performance
- We hope future work will continue to devise learning algorithms that work across multiple modalities, rather than focusing on individual settings