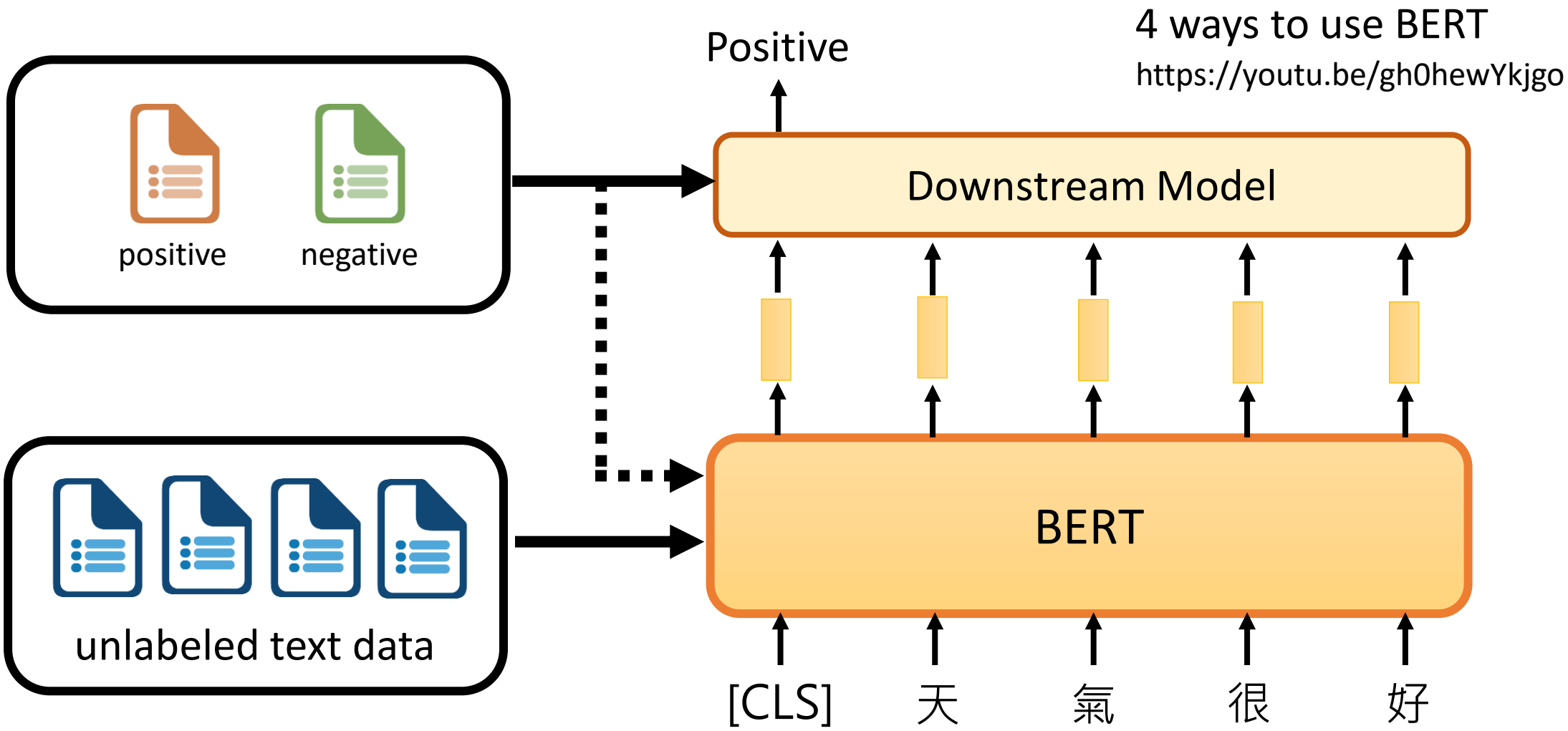
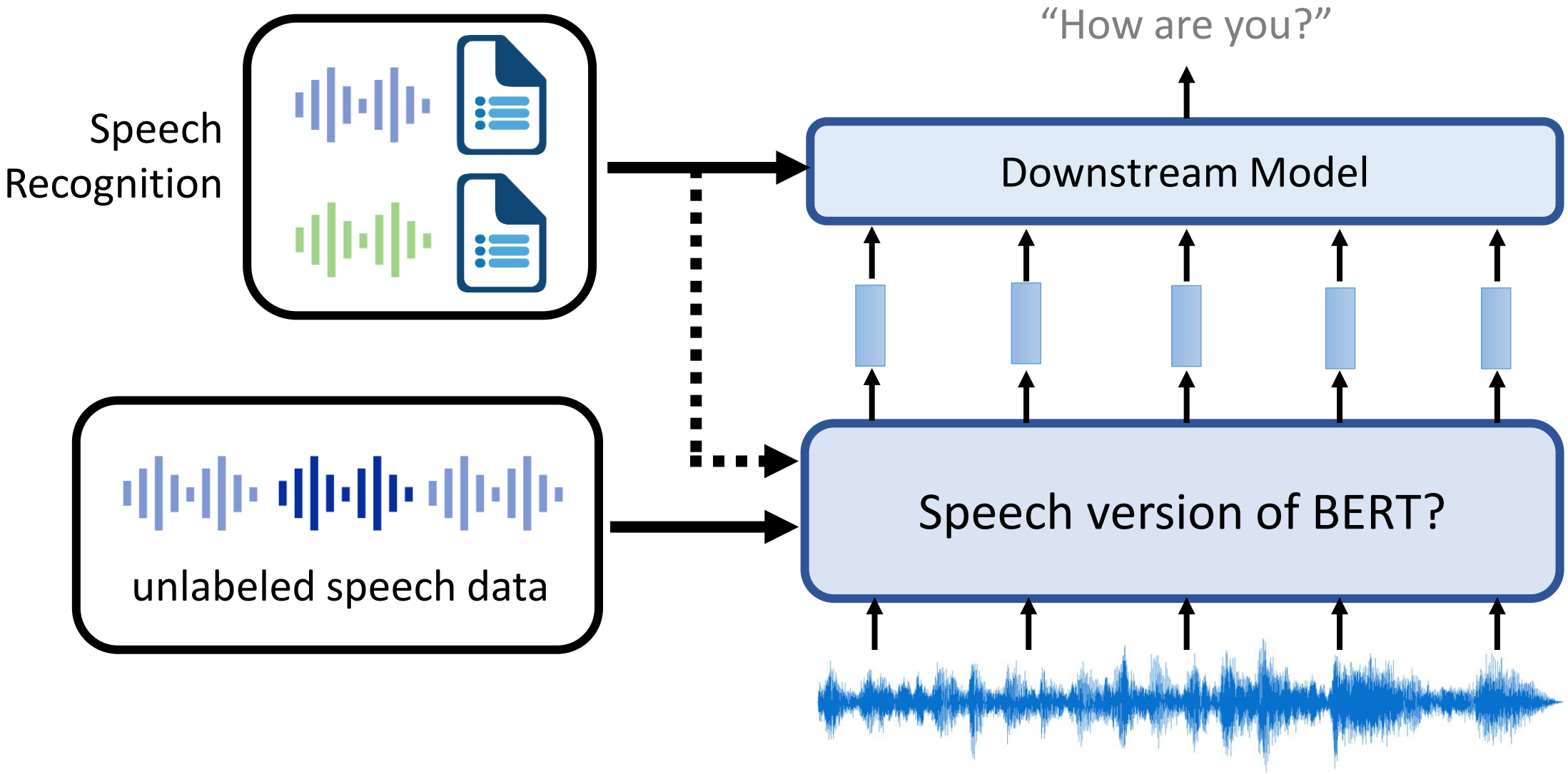
Self-supervised Learning for Speech and Image

Hung-yi Lee

Review: Self-supervised Learning for **Text**

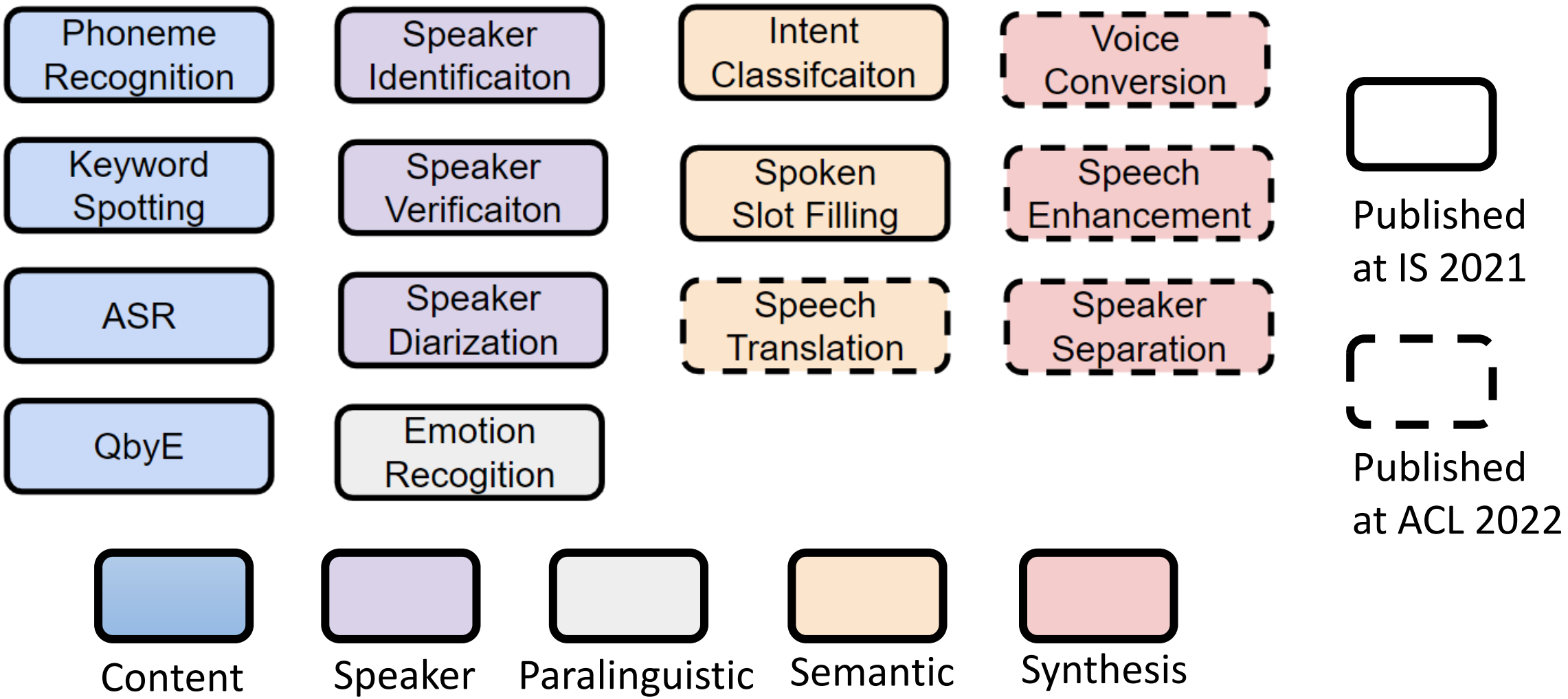


Self-supervised Learning for **Speech**



Speech processing Universal PERformance Benchmark (SUPERB)

https://superbbenchmark.org/



SUPERB: Speech processing Universal PERformance Benchmark

Shu-wen Yang[1], Po-Han Chi[1], Yung-Sung Chuang[1*], Cheng-I Jeff Lai[2*], Kushal Lakhotia[3*], Yist Y. Lin[1*], Andy T. Liu[1*], Jiatong Shi[4*], Xuankai Chang[6], Guan-Ting Lin[1], Tzu-Hsien Huang[1], Wei-Cheng Tseng[1], Ko-tik Lee[1], Da-Rong Liu[1], Zili Huang[4], Shuyan Dong[5†], Shang-Wen Li[5†], Shinji Watanabe[6], Abdelrahman Mohamed[3], Hung-yi Lee[1]*

Presented at INTERSPEECH 2021

https://arxiv.org/abs/2105.01051

SUPERB−SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities

Hsiang-Sheng Tsai[1*], Heng-Jui Chang[1*], Wen-Chin Huang[2*], Zili Huang[3*], Kushal Lakhotia[4*], Shu-wen Yang[1], Shuyan Dong[5], Andy T. Liu[1], Cheng-I Lai[6], Jiatong Shi[7], Xuankai Chang[7], Phil Hall[8], Hsuan-Jui Chen[1], Shang-Wen Li[5], Shinji Watanabe[7], Abdelrahman Mohamed[5], Hung-yi Lee[1]

To be appeared at ACL 2022

https://arxiv.org/abs/2203.06849

Speech processing Universal PERformance Benchmark (SUPERB)

- To learn more:

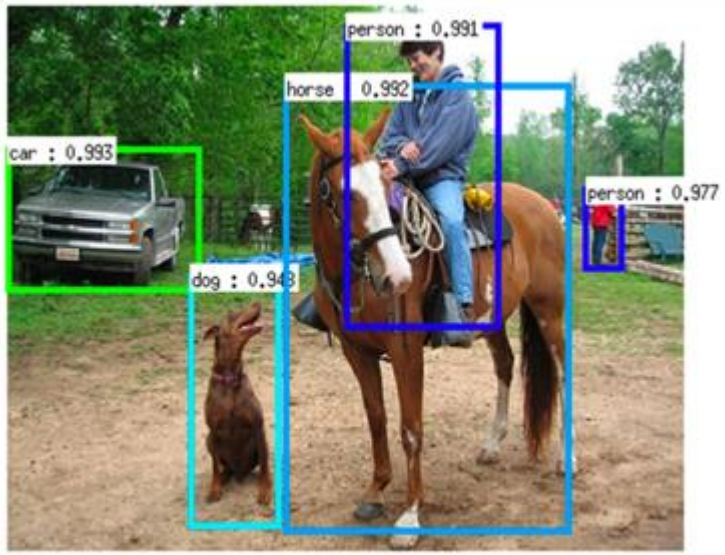
https://youtu.be/MpsVE60iRLM
(Mandarin version)

https://youtu.be/GTjwYzFG54E
(English version)

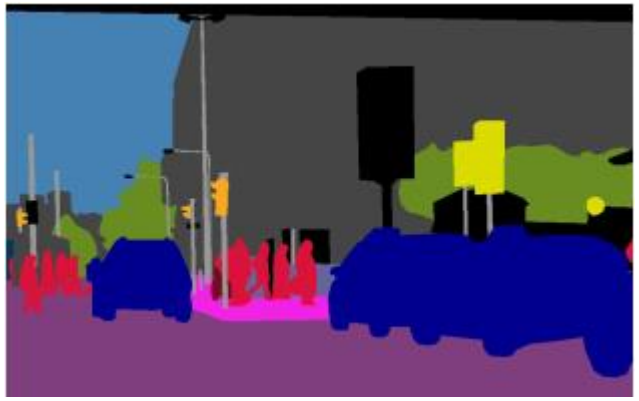
- Toolkit – S3PRL: https://github.com/s3prl/s3prl

Self-supervised Learning for **Image**

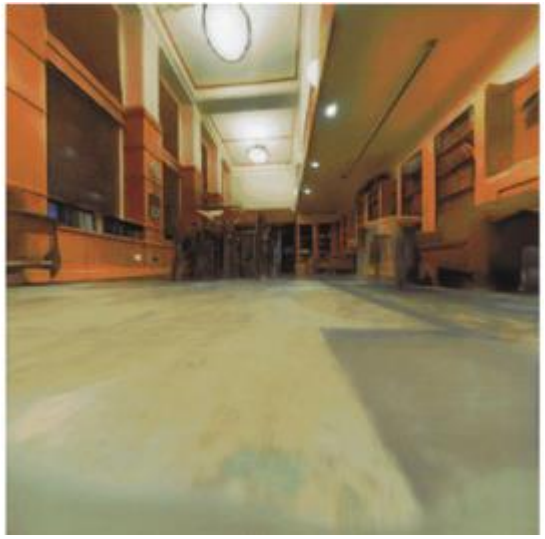
Image Recognition



Object Detection

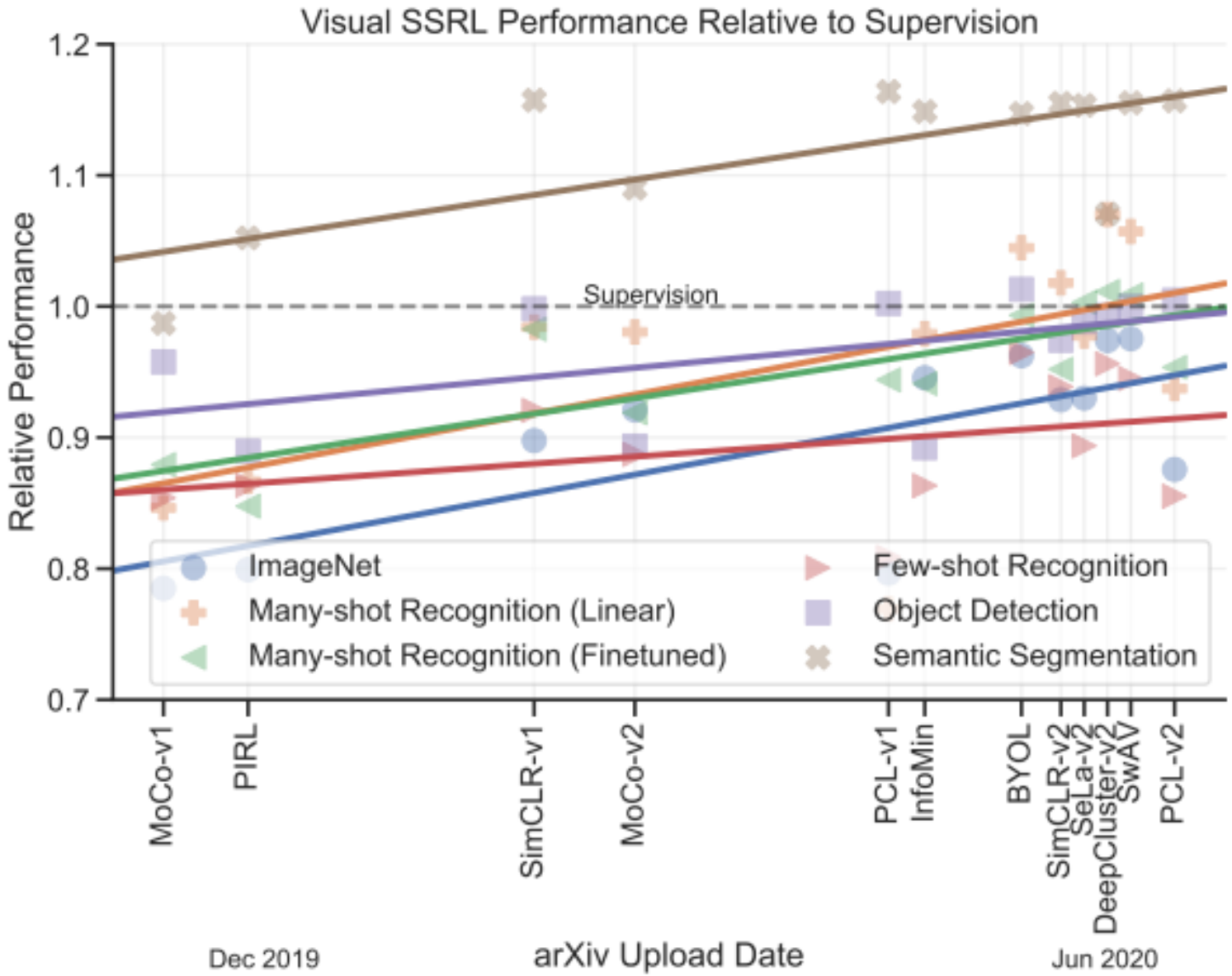


Semantic Segmentation



Visual Navigation

- How Well Do Self-Supervised Models Transfer? https://arxiv.org/abs/2011.13377
- Scaling and Benchmarking Self-Supervised Visual Representation Learning https://arxiv.org/abs/1905.01235



Source of image: https://arxiv.org/abs/2110.09327

BERT series

GPT series

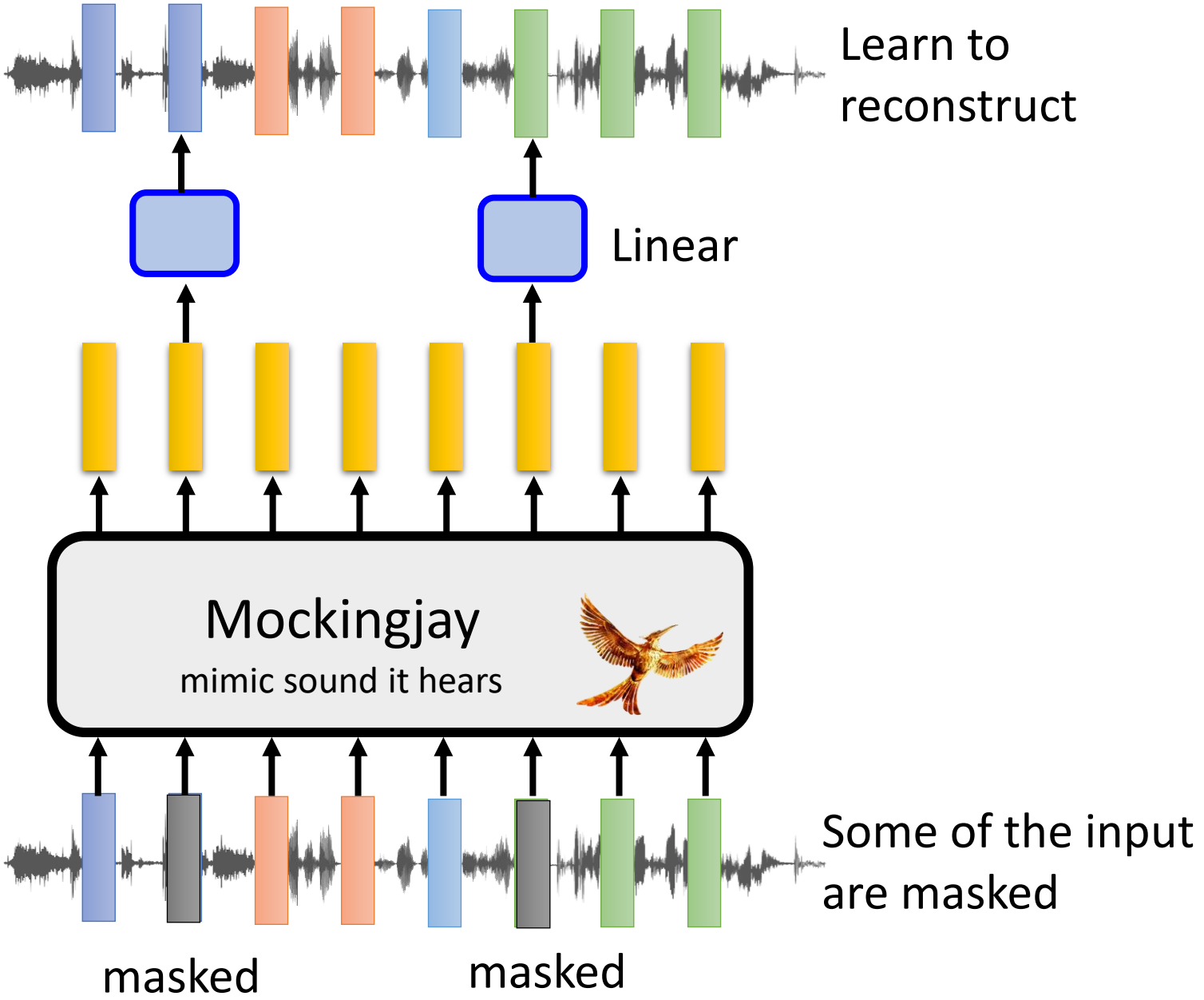
1. Generative Approaches

Masking

BERT series

How about **speech**?

https://arxiv.org/abs/1910.12638

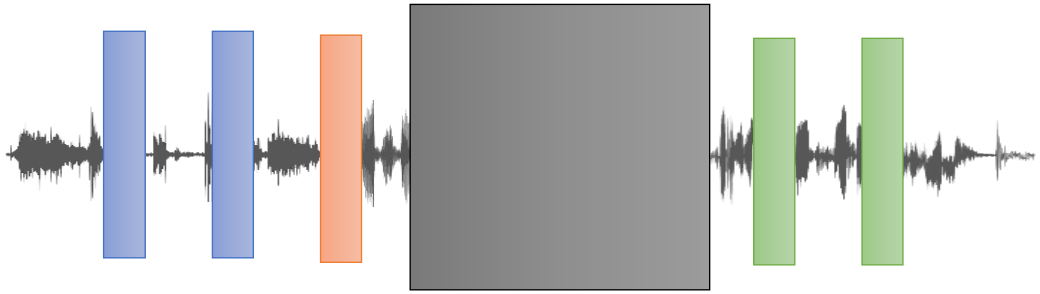


Masking

- Smoothness of acoustic features

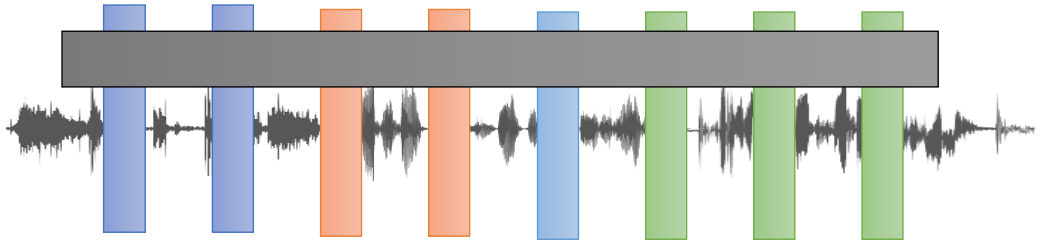
  https://arxiv.org/abs/1910.12638

Masking consecutive features



- Masking strategies for speech

Masking specific dimensions



Learn more speaker information in this way

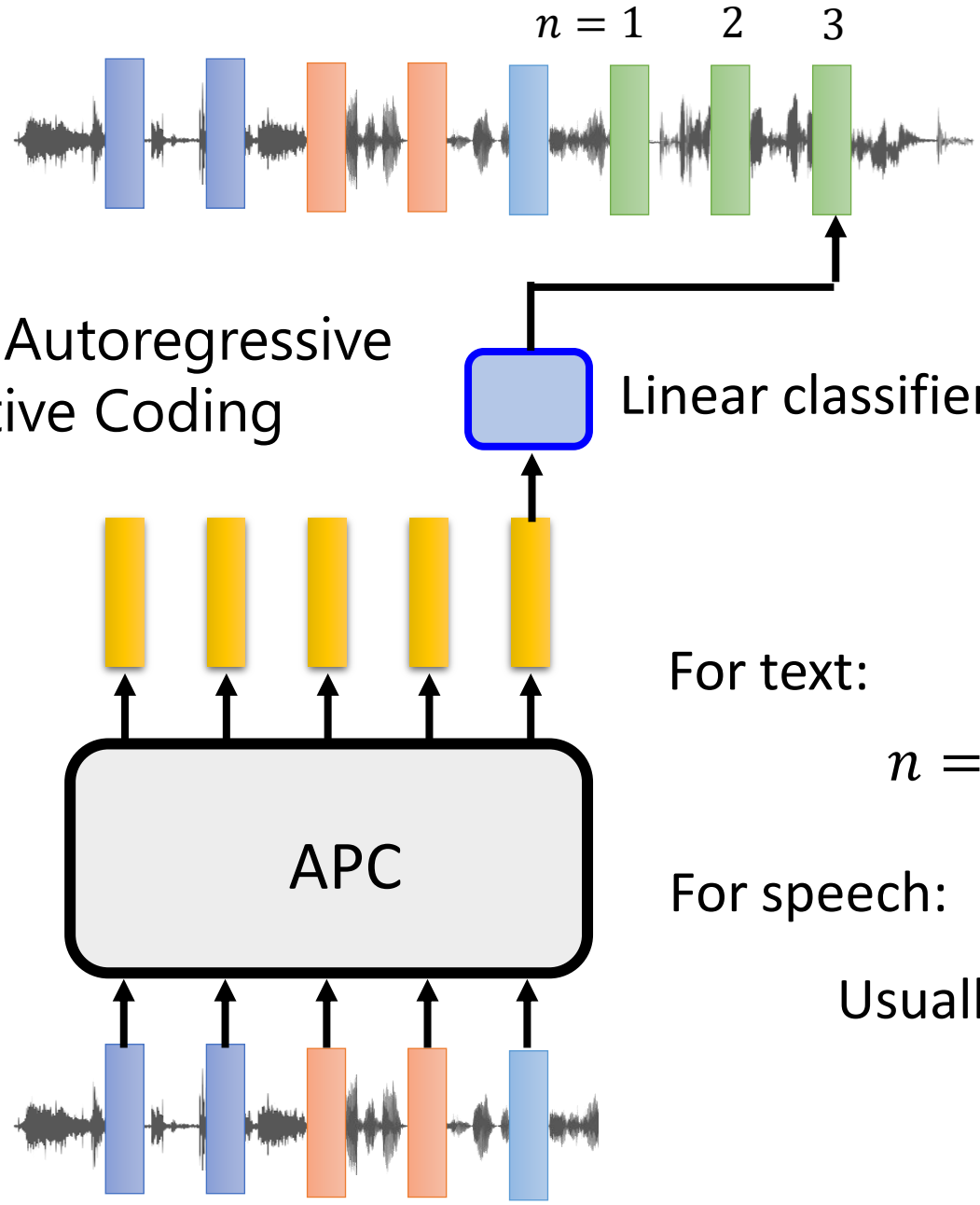
TERA: https://arxiv.org/abs/2007.06028

Predicting Future

GPT series

How about **speech**?

https://arxiv.org/abs/1910.12607



For text:

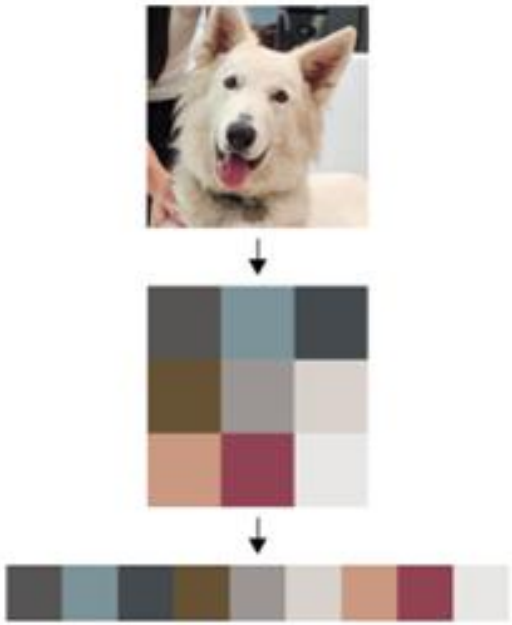
$n = 1$

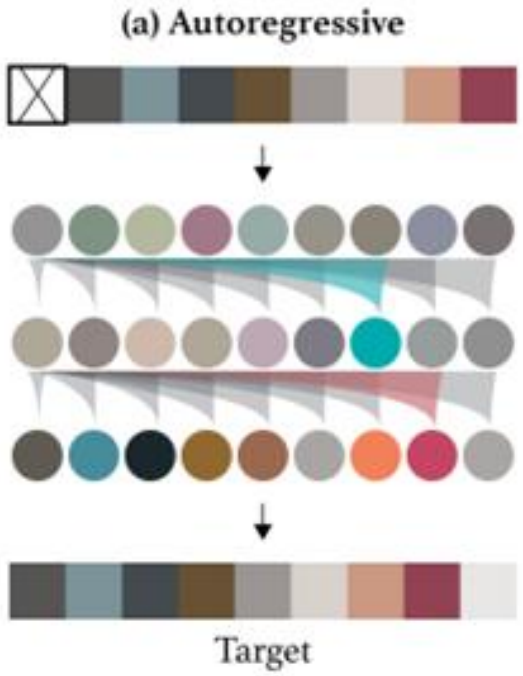
For speech:

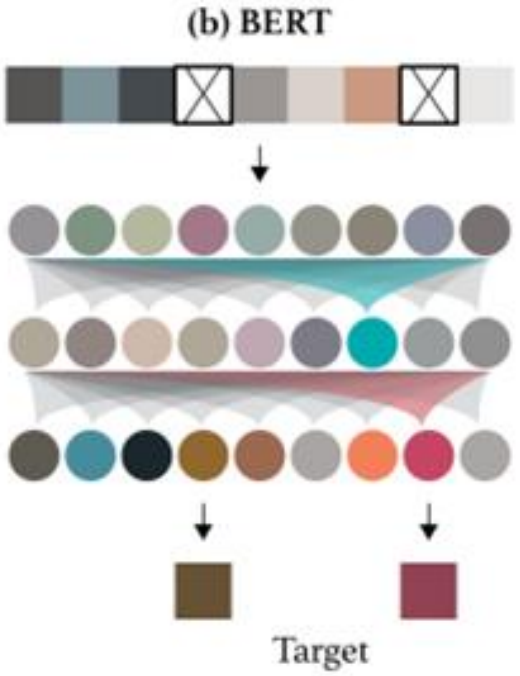
Usually $n > 3$

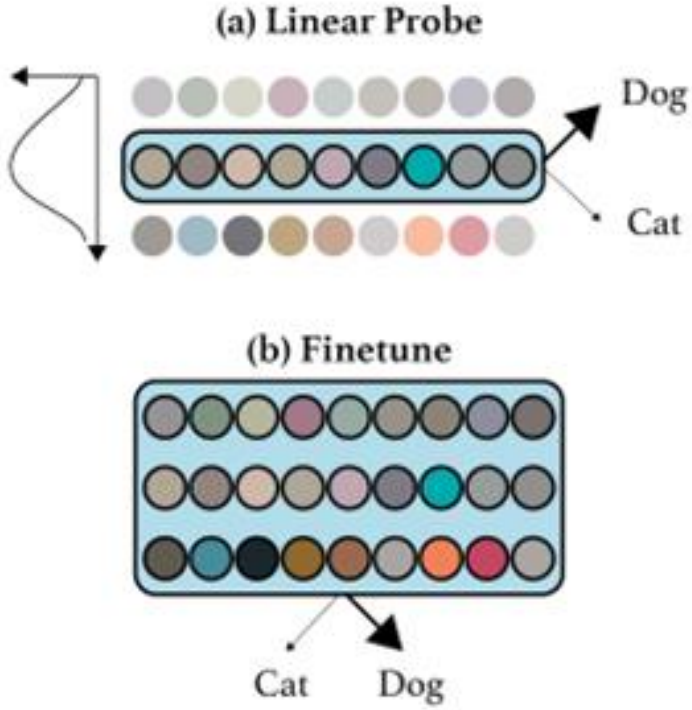
https://openai.com/blog/image-gpt/

How about **image**?









Speech and images contain many details that are difficult to generate.

Can a model learn without generation?

2. Predictive Approach

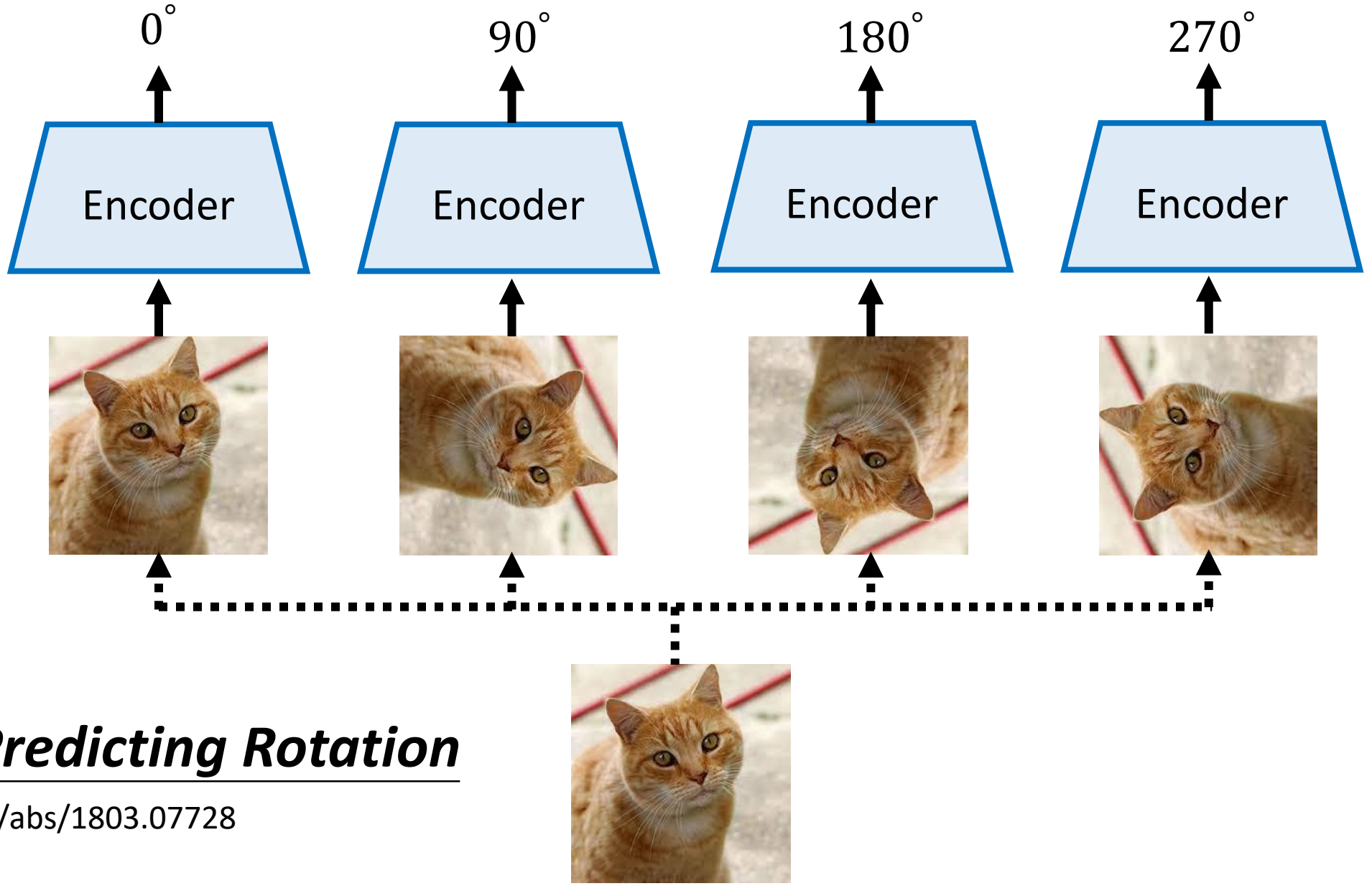
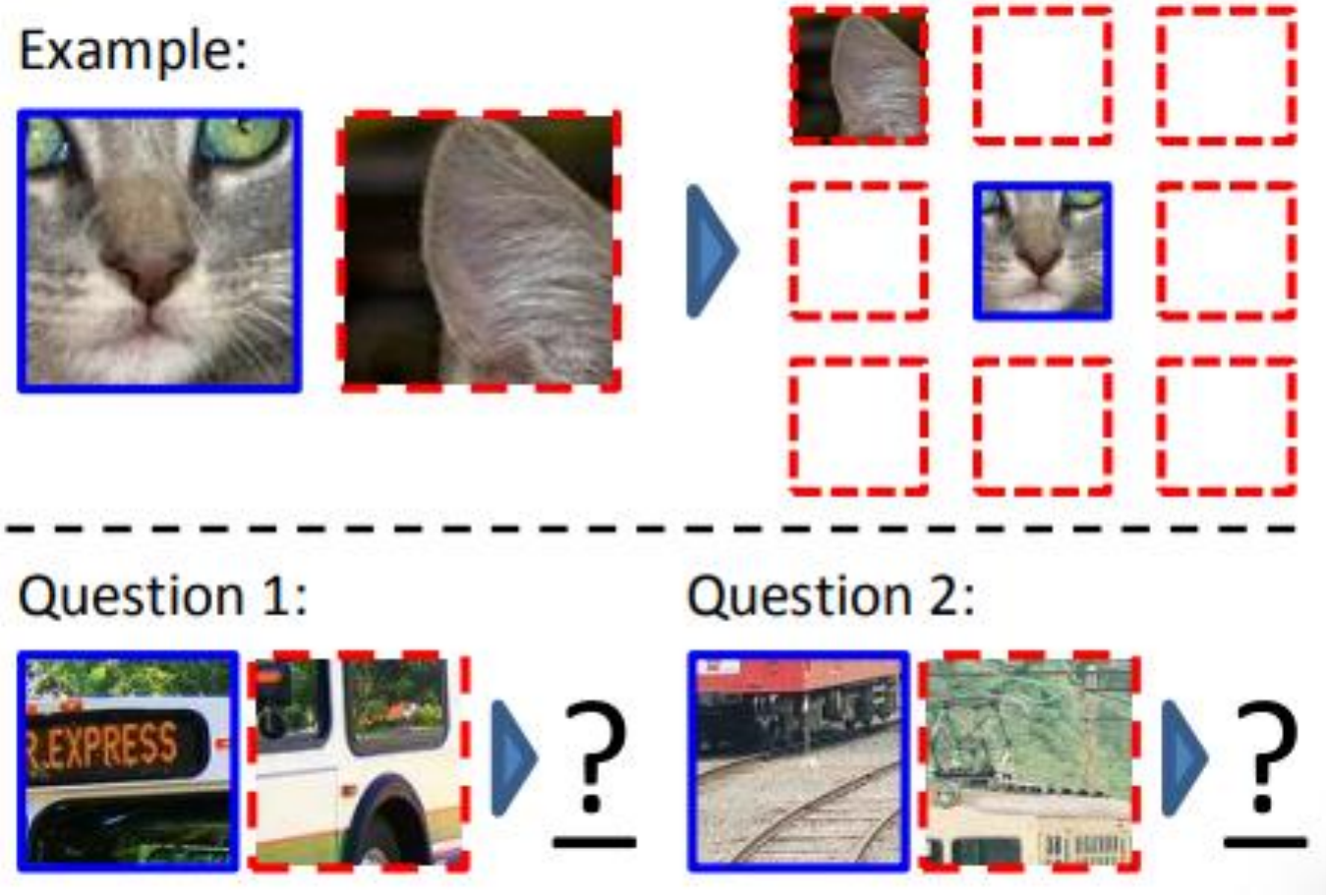


Image - Predicting Rotation

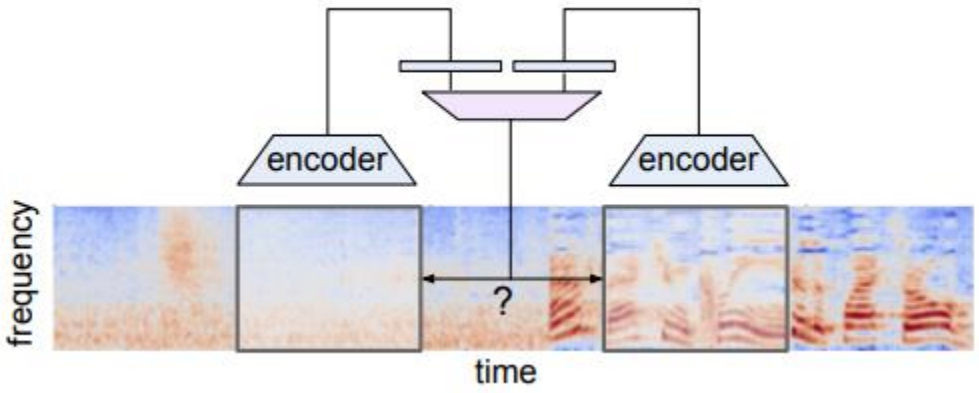
https://arxiv.org/abs/1803.07728

Image – Context Prediction

https://arxiv.org/abs/1505.05192



*Similar idea on **Speech***



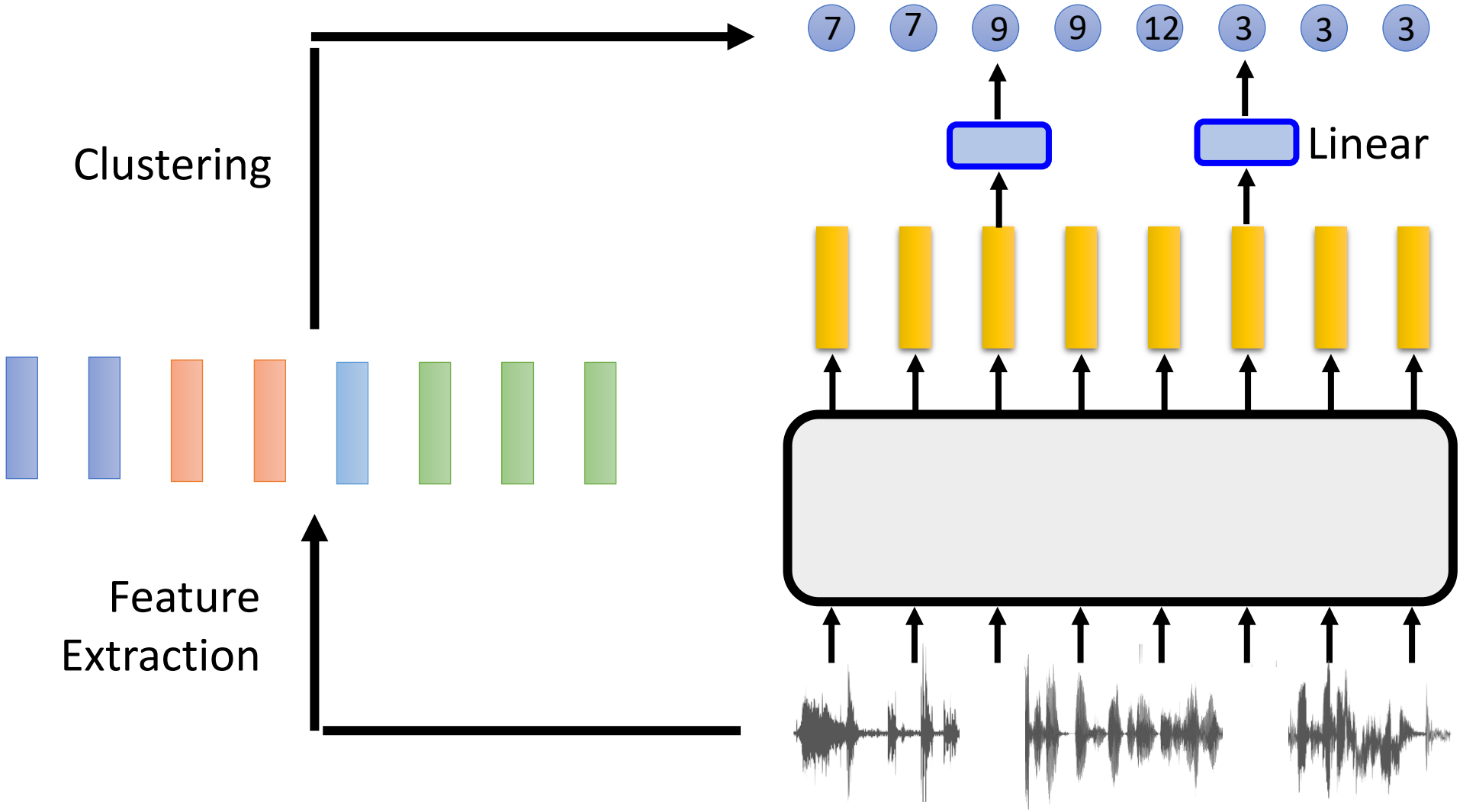
https://ieeexplore.ieee.org/document/9060816

Predict Simplified Objects

Speech HuBERT https://arxiv.org/abs/2106.07447

BEST-RQ https://arxiv.org/abs/2202.01855

Image DeepCluster https://arxiv.org/abs/1807.05520

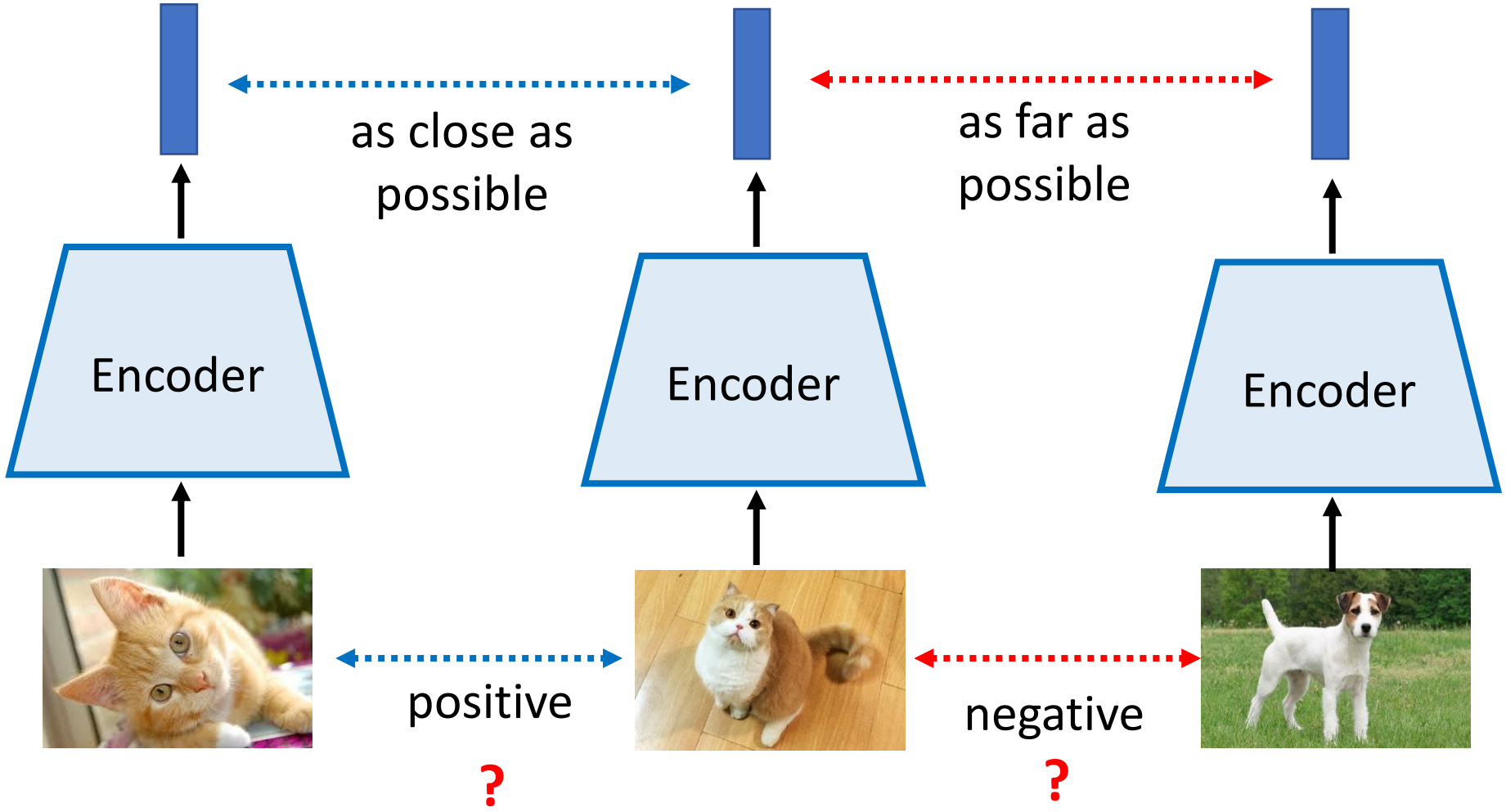


Speech and images contain many
details that are difficult to generate.

Can a model learn without generation?

3. Contrastive Learning

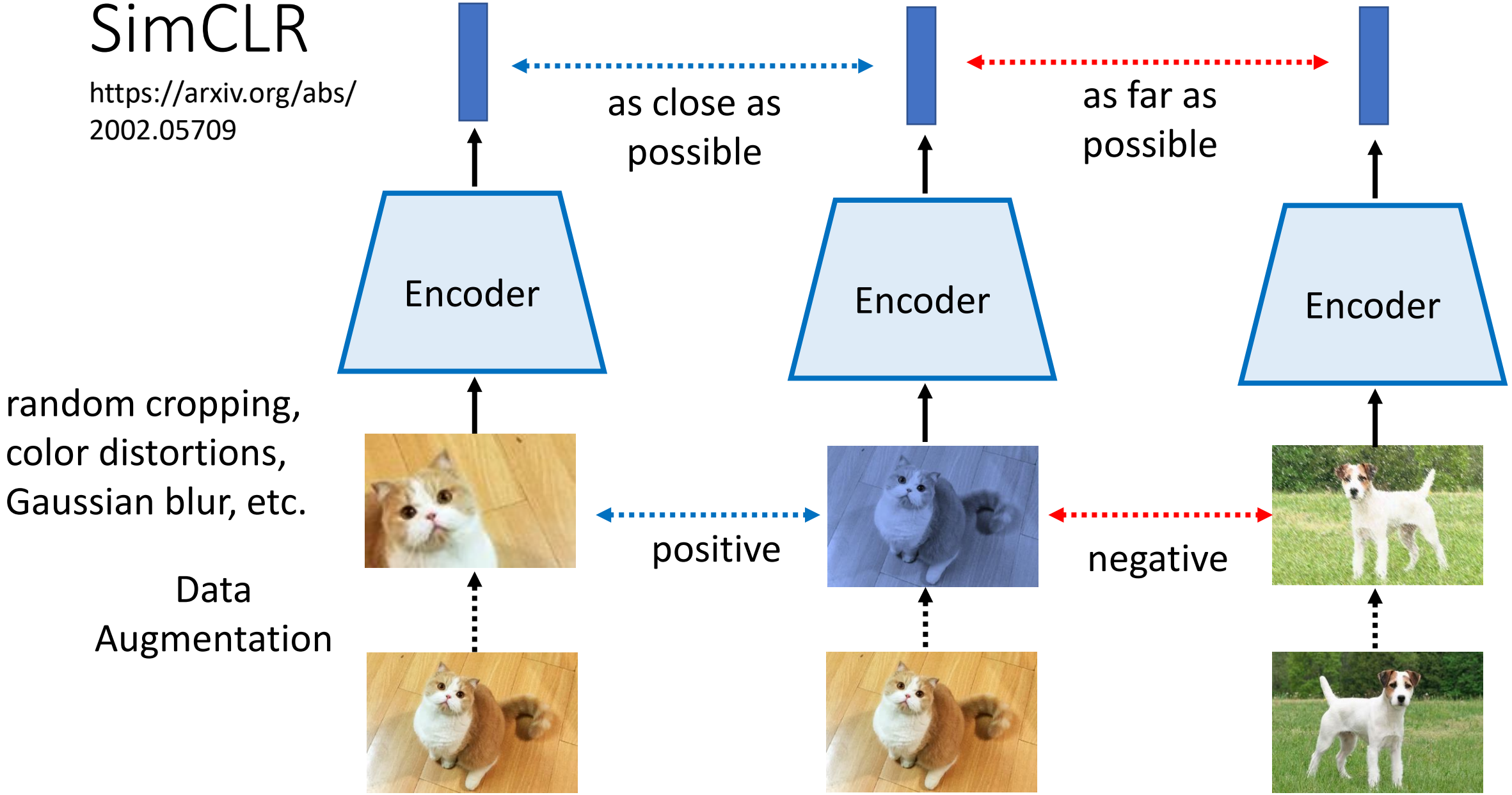
Basic Idea of Contrastive Learning



Speech SimCLR: https://arxiv.org/abs/2010.13991

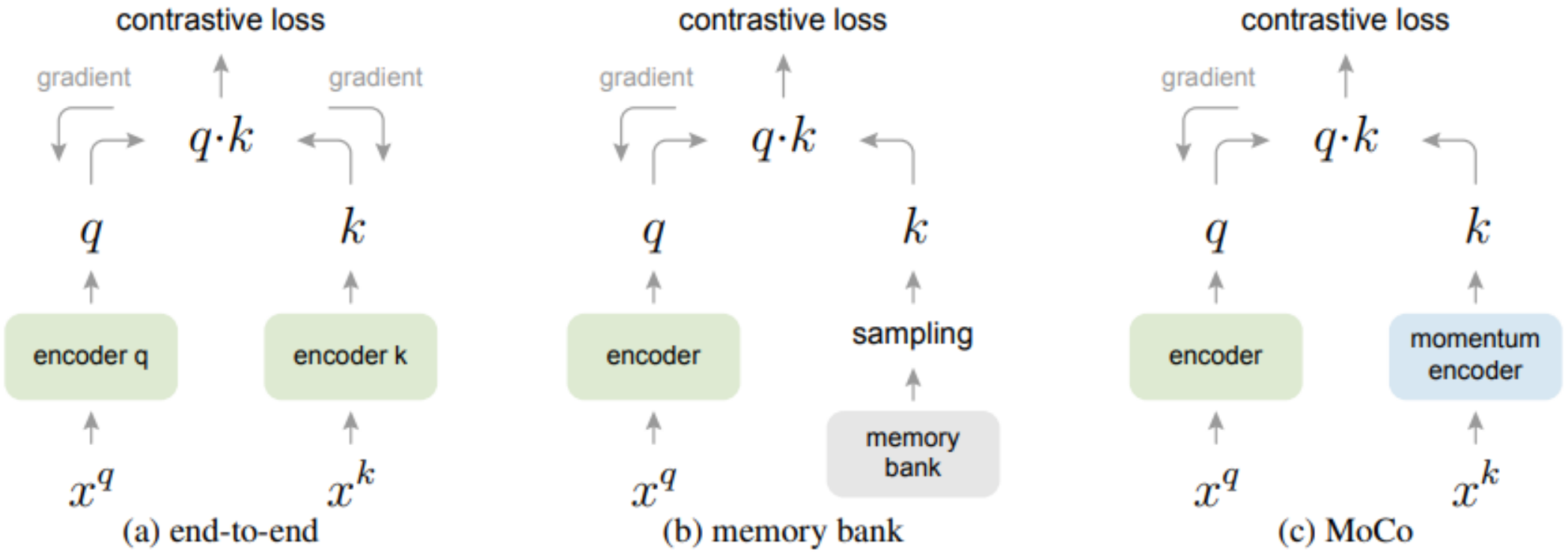
SimCLR

https://arxiv.org/abs/2002.05709



MoCo

https://arxiv.org/abs/1911.05722



MoCo v2 https://arxiv.org/abs/2003.04297

Contrastive Learning for **Speech**

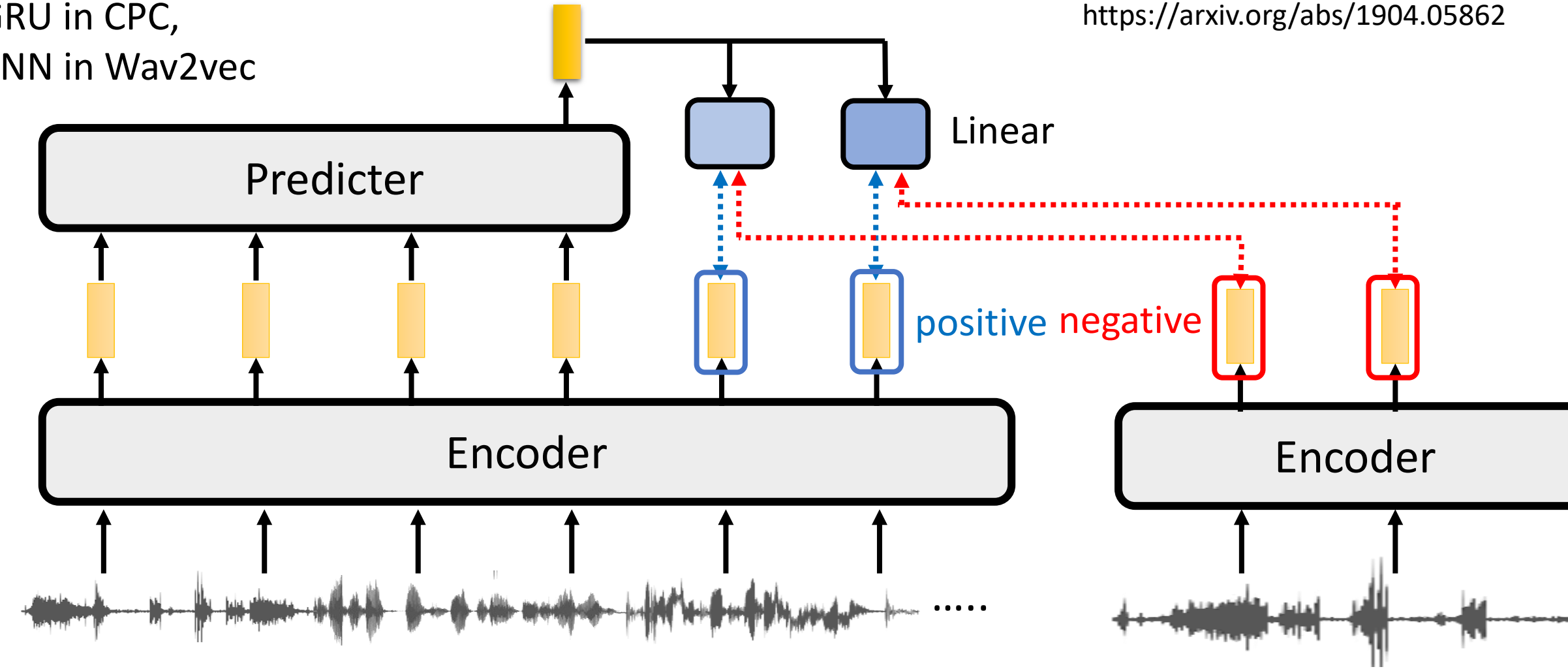
CPC

https://arxiv.org/abs/1807.03748

Wav2vec

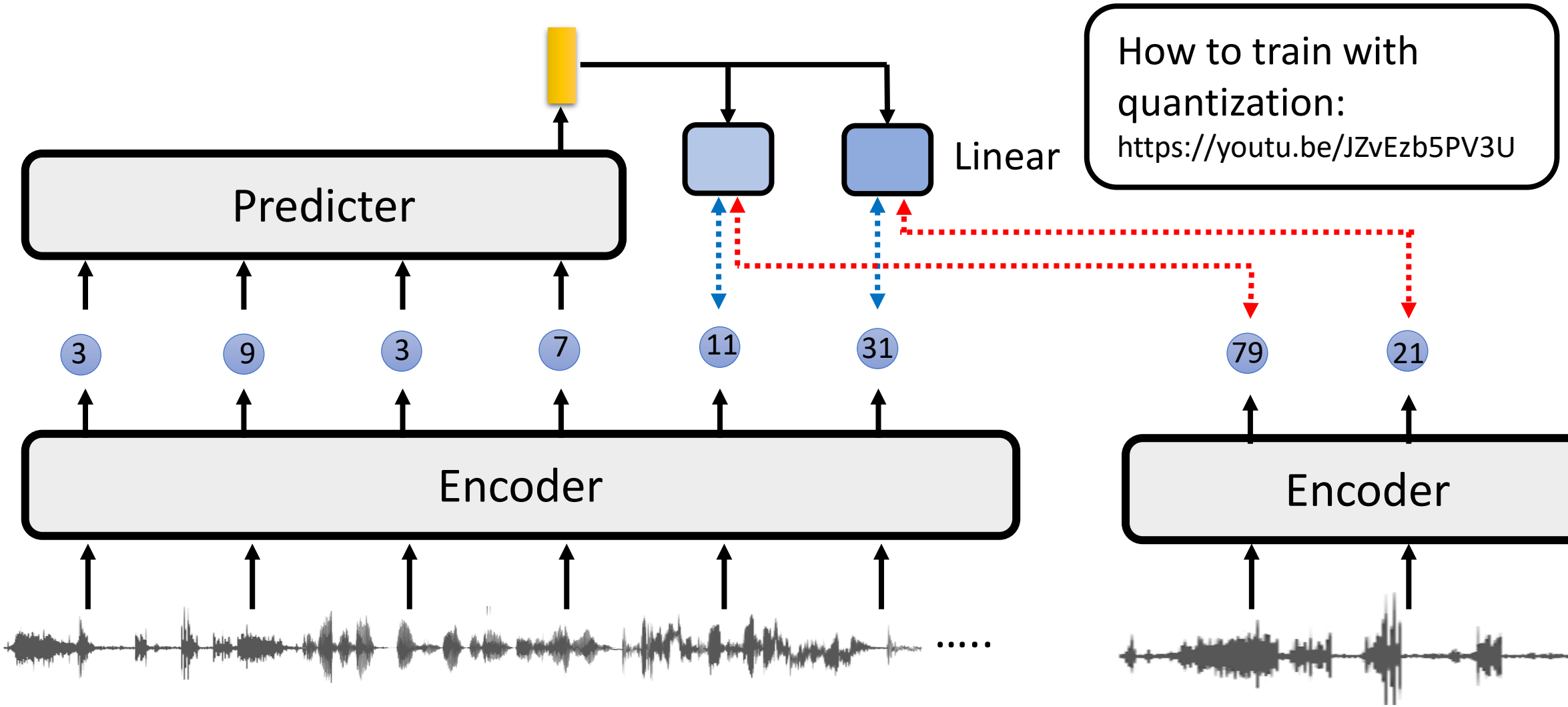
https://arxiv.org/abs/1904.05862

GRU in CPC,
CNN in Wav2vec



Contrastive Learning for **Speech**

VQ-wav2vec
https://arxiv.org/abs/1910.05453



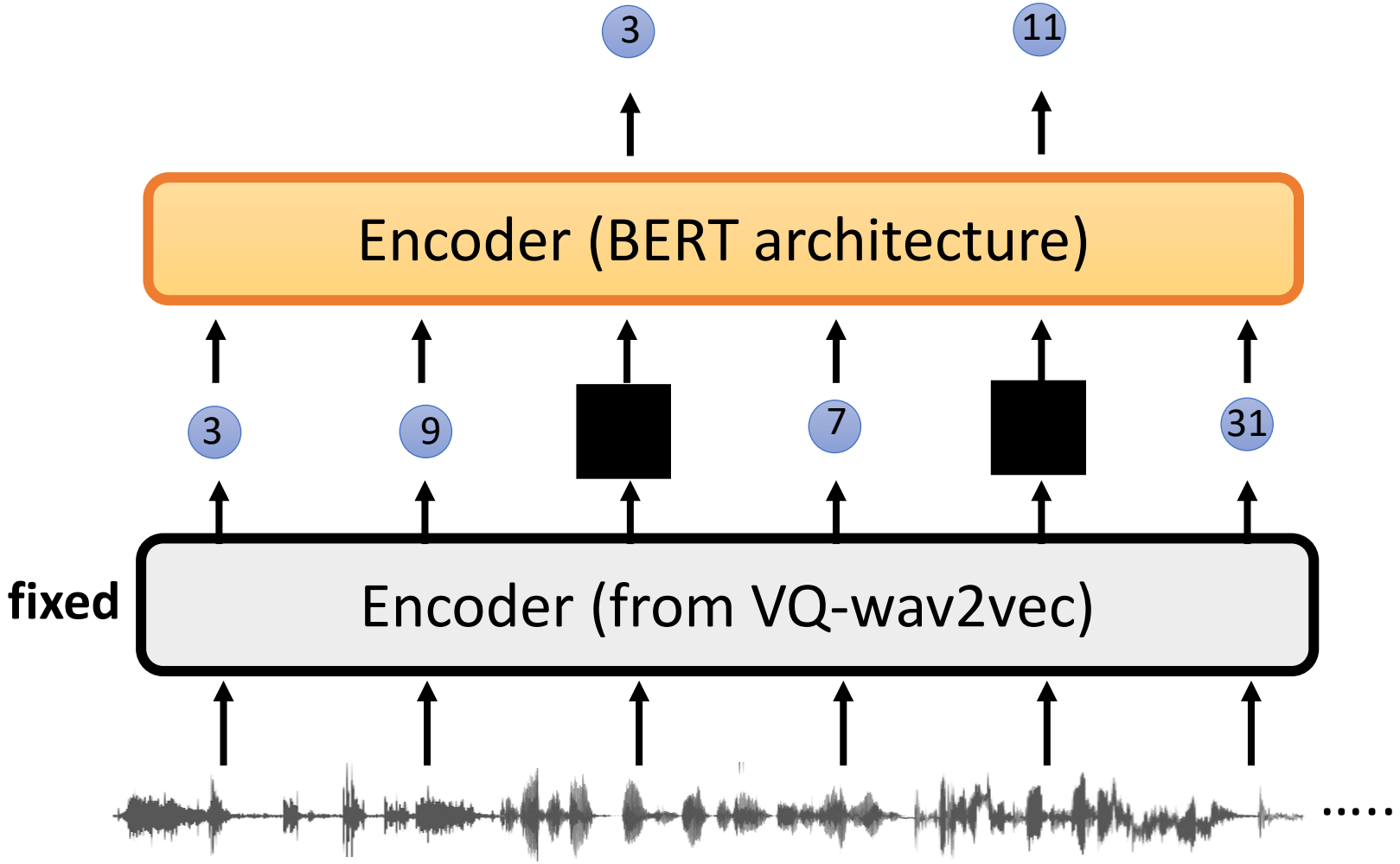
Contrastive Learning for **Speech**

VQ-wav2vec + BERT

https://arxiv.org/abs/1910.05453

Discrete BERT

https://arxiv.org/abs/1911.03912



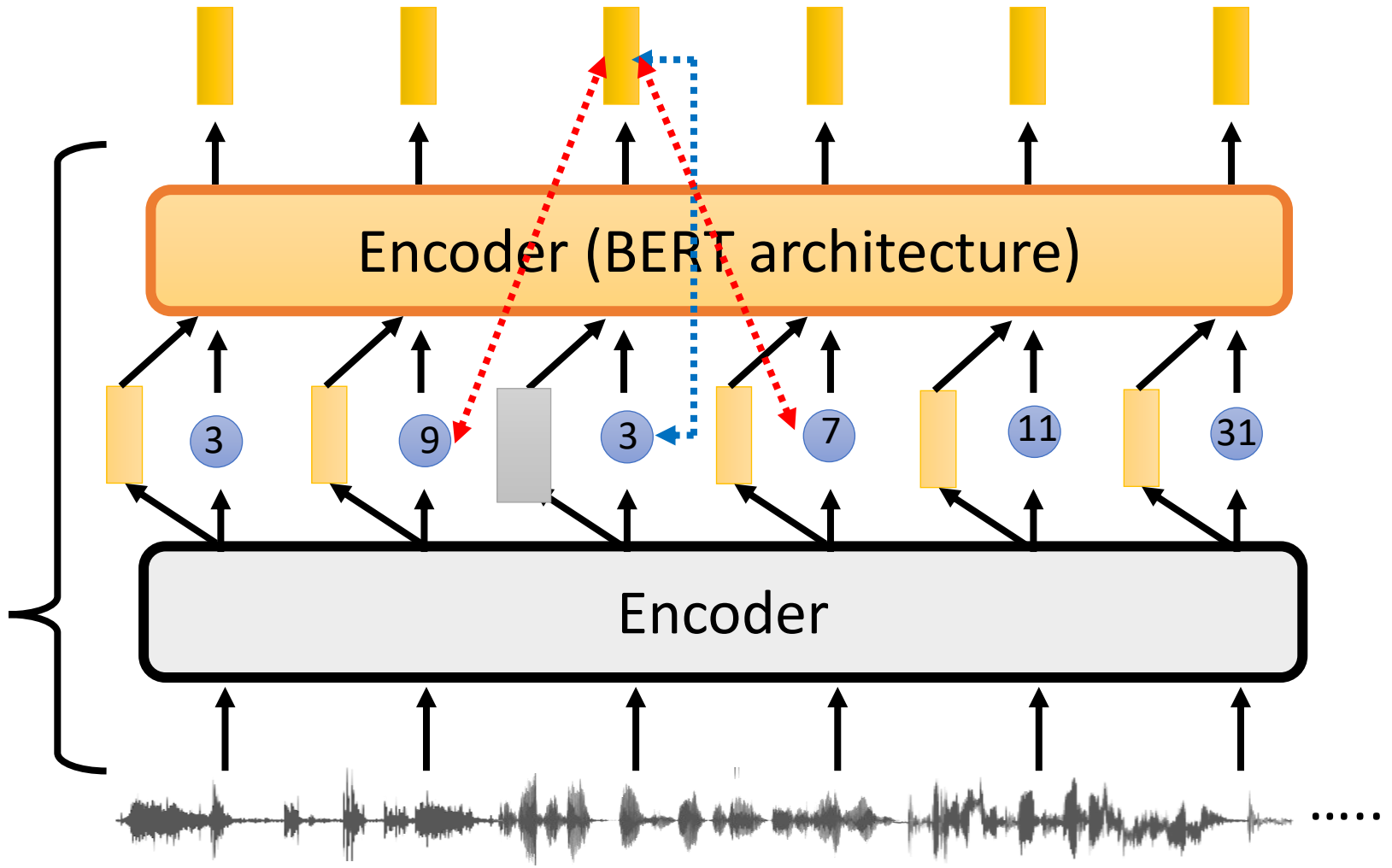
Contrastive Learning for **Speech**

Wav2vec 2.0

https://arxiv.org/abs/2006.11477

Continuous input is critical

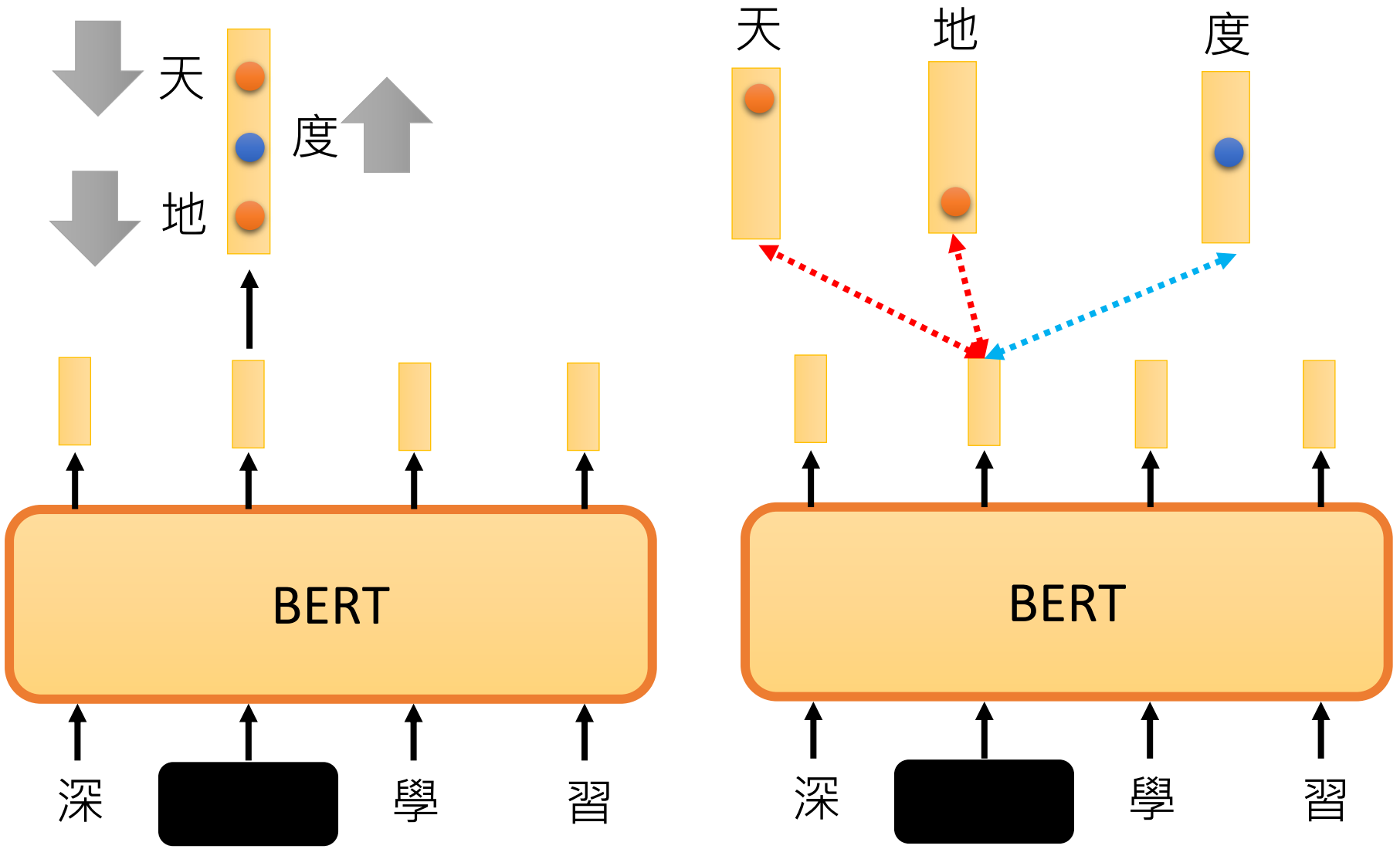
Quantized target improves performance



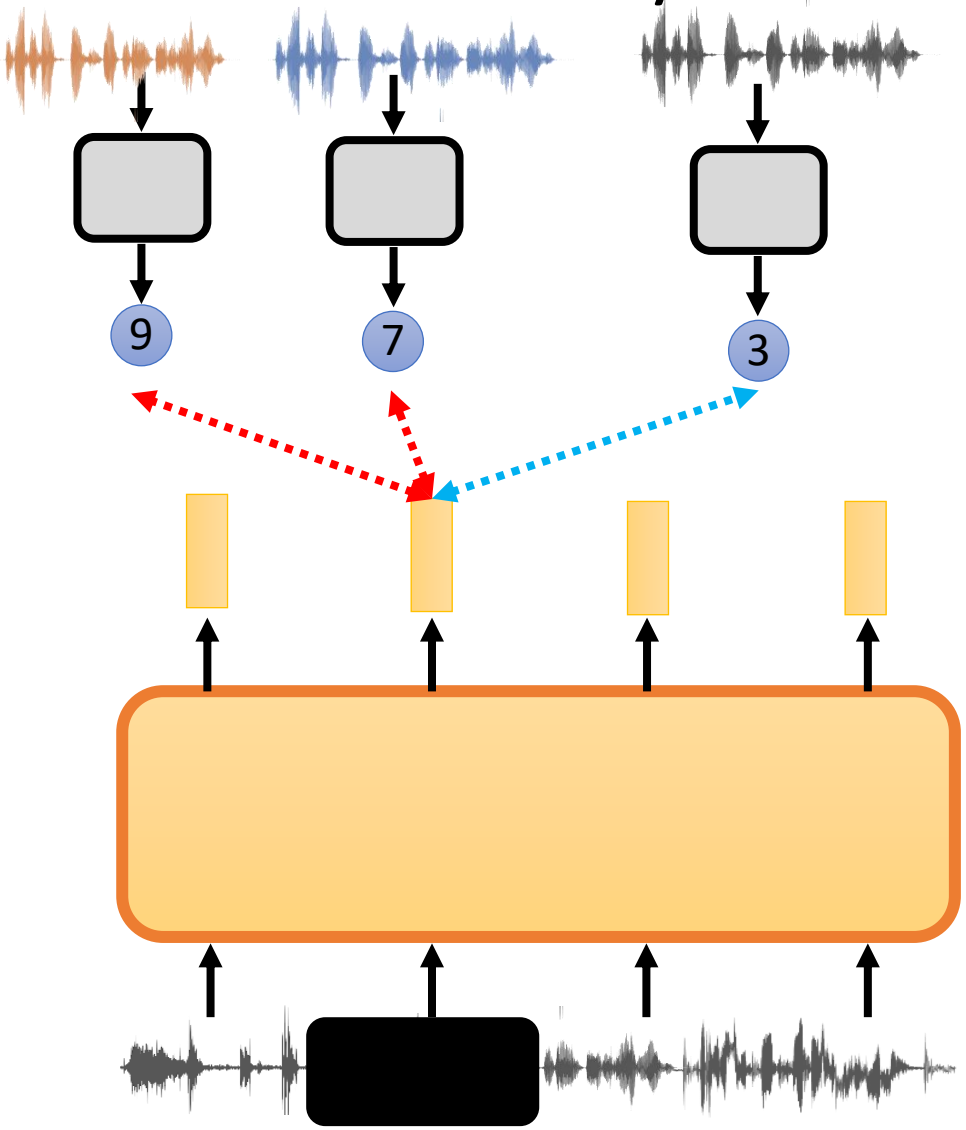
Why not formulated as typical classification?

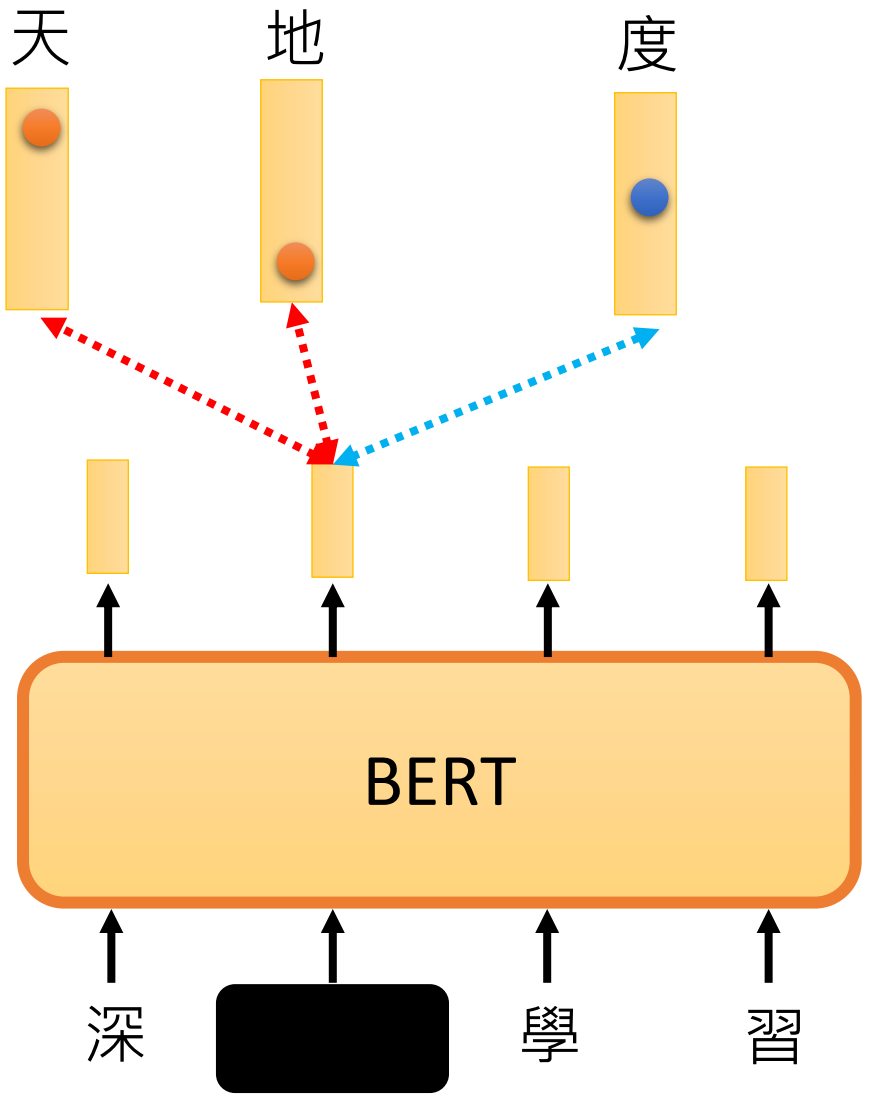
Alterative way to understand Wav2vec 2.0

Is BERT contrastive learning?

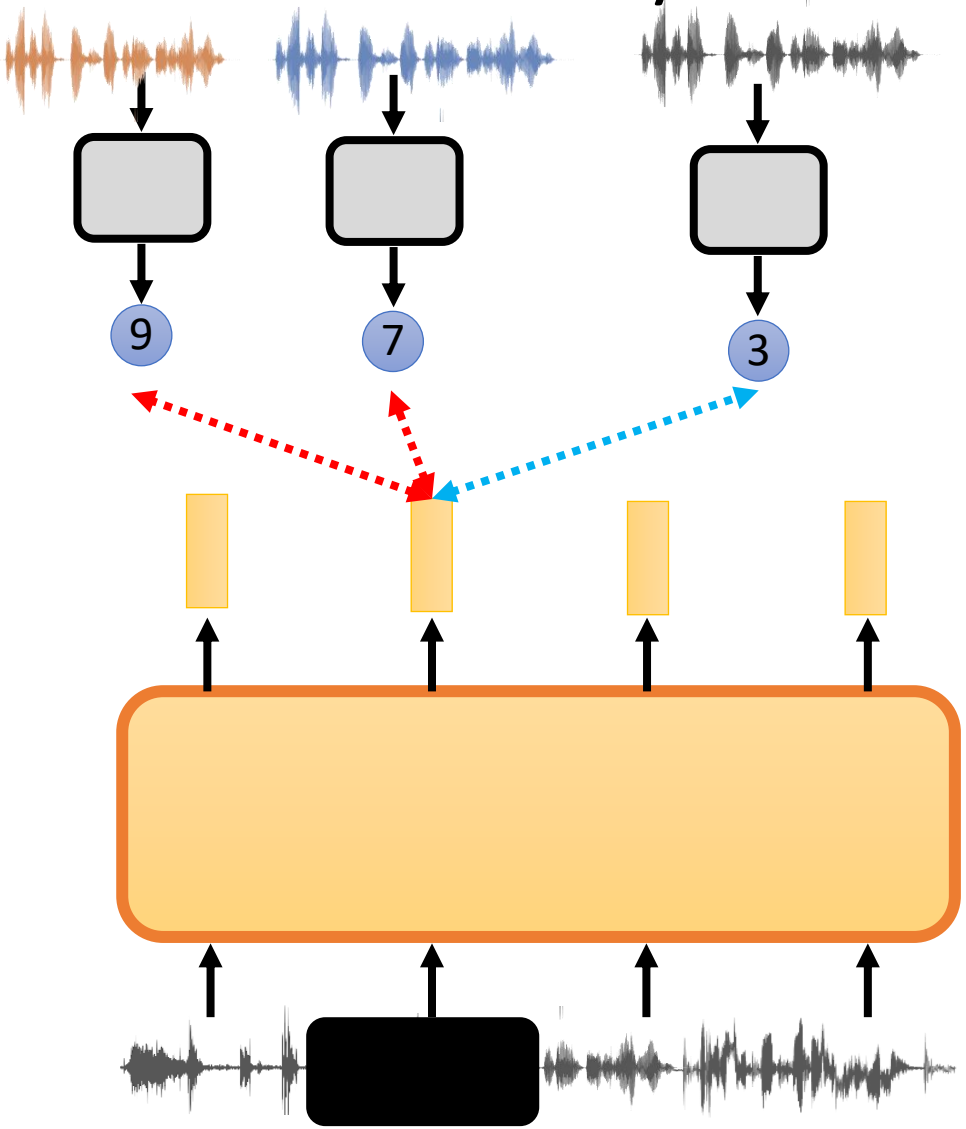


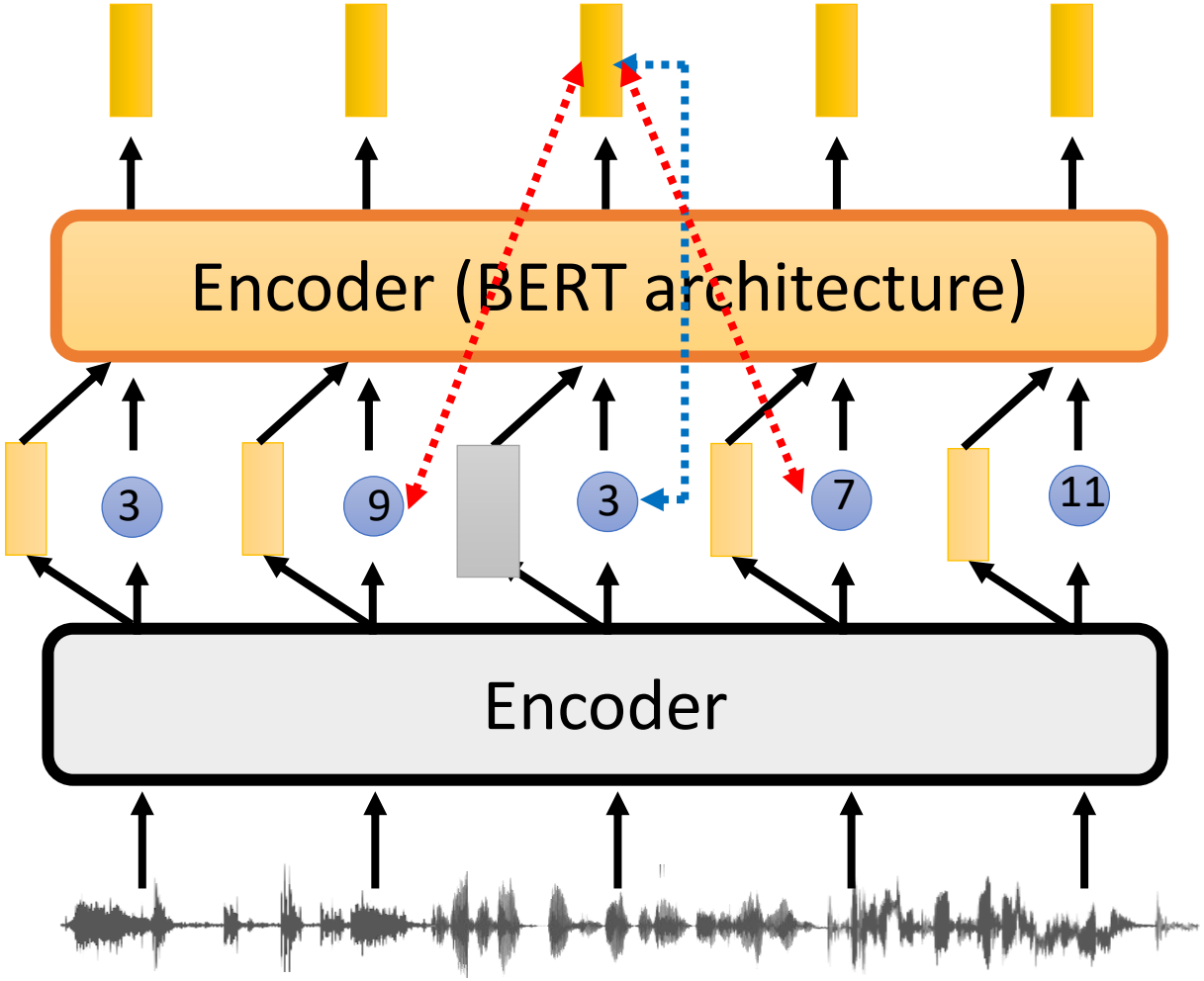
Alterative way to understand Wav2vec 2.0





Alterative way to understand Wav2vec 2.0





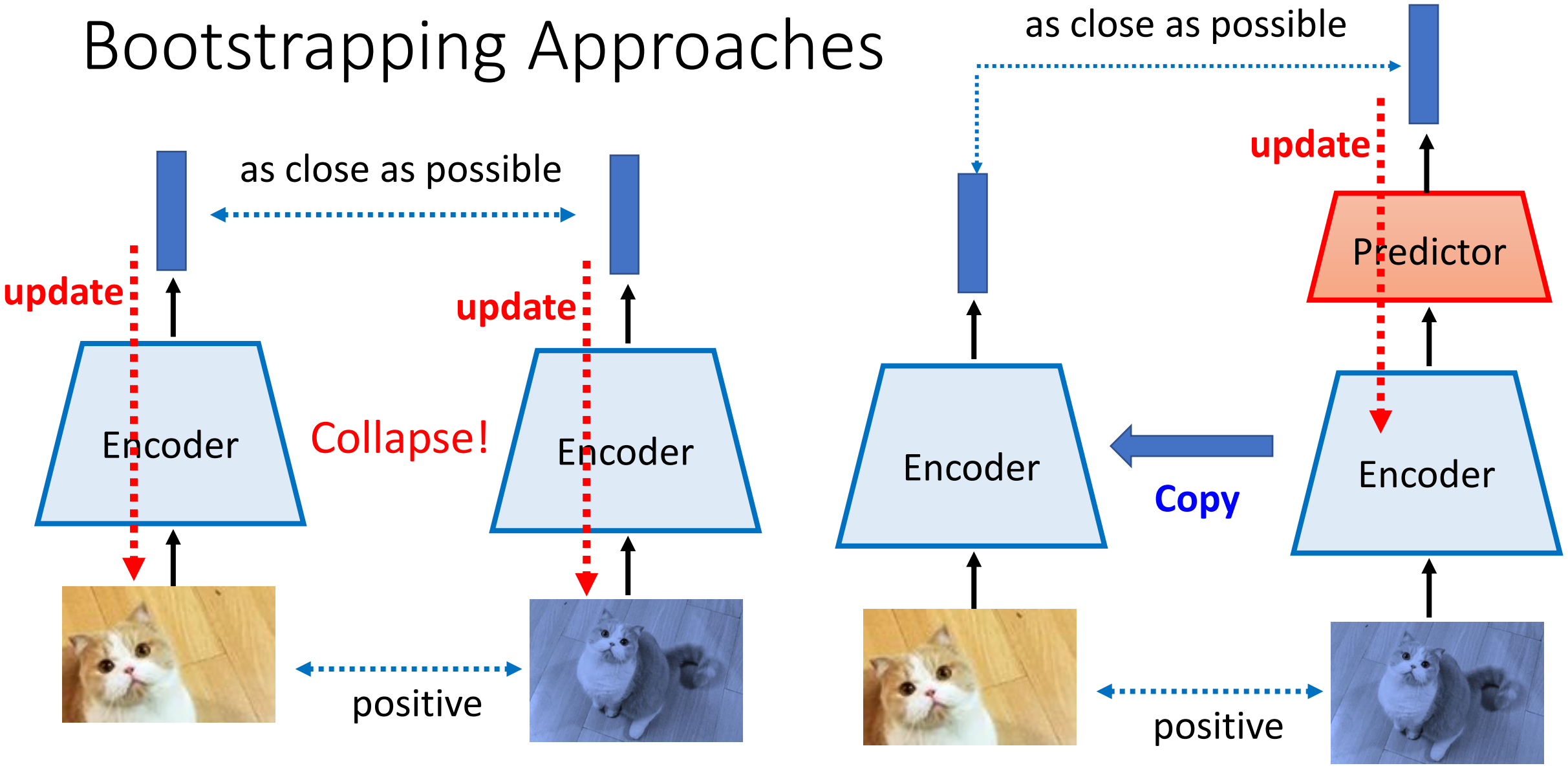
Selecting Negative Examples is not trivial …

- The negative examples should be hard enough. But cannot be too hard …

Learning without negative examples

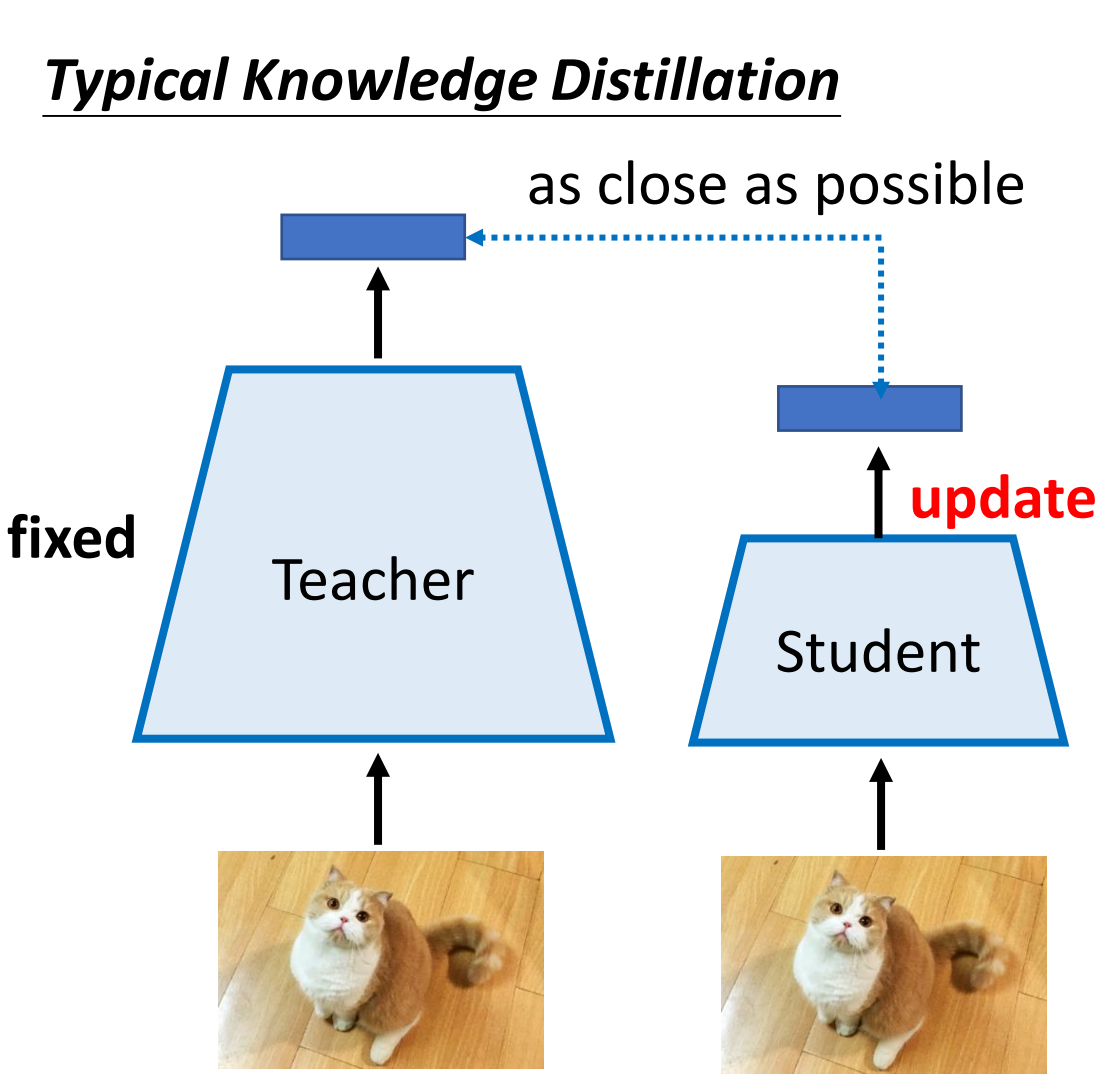
4. Bootstrapping Approaches

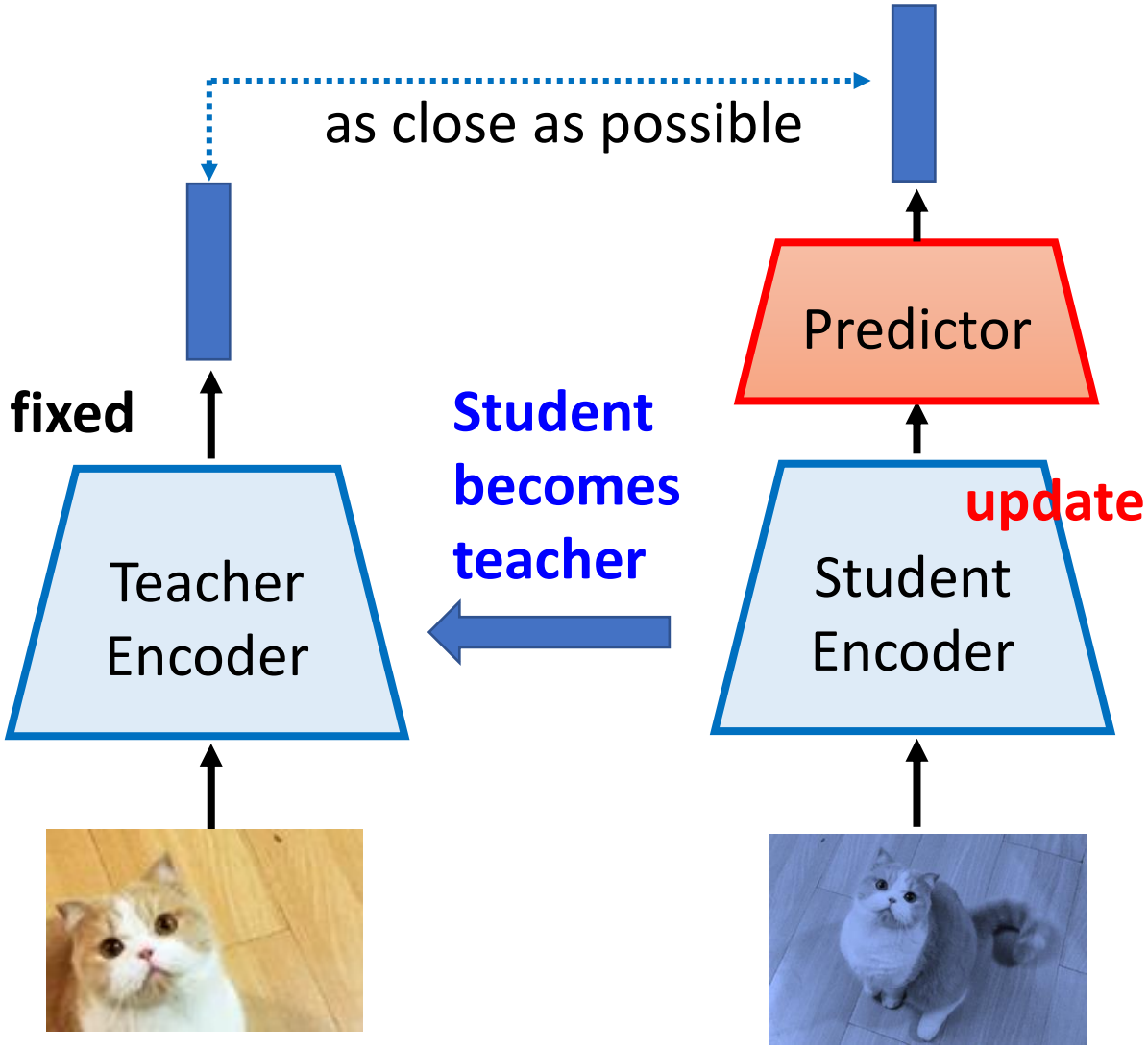
Bootstrapping Approaches



Alterative way to understand Bootstrapping

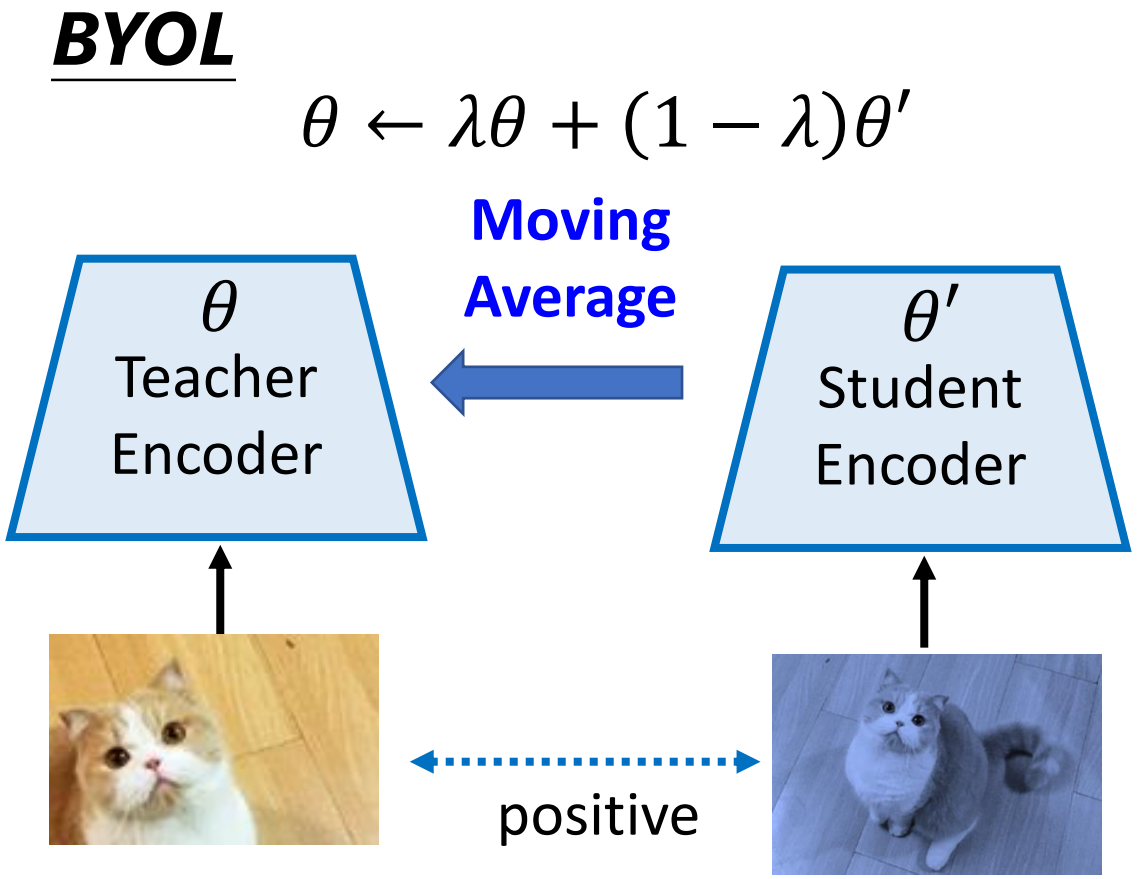
Typical Knowledge Distillation





Bootstrapping Approaches

- Image
  - Bootstrap your own latent (BYOL)
    - https://arxiv.org/abs/2006.07733
  - Simple Siamese (SimSiam)
    - https://arxiv.org/abs/2011.10566
- Speech
  - Data2vec: the student learns from multiple layers of the teacher
    - https://arxiv.org/abs/2202.03555



Learning without negative examples

5. Simply Extra Regularization

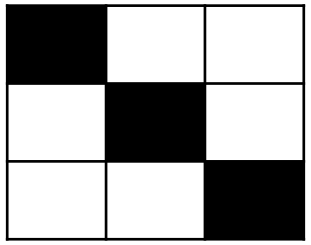
Barlow Twins https://arxiv.org/abs/2103.03230

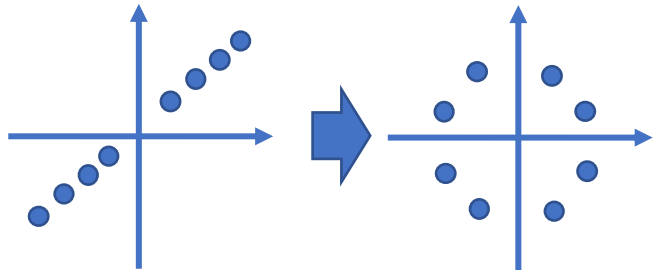
Variance-Invariance-Covariance Regularization (VICReg)

https://arxiv.org/abs/2105.04906

Covariance

Off-diagonal elements close to 0





Variance

Variance lager than a threshold

Prevent collapse

Invariance

as close as possible

Encoder

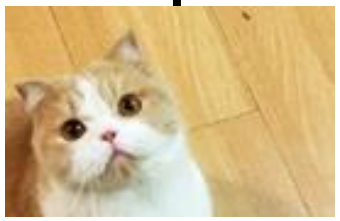
Encoder

Encoder

Encoder

Encoder

positive



Concluding Remarks

- Generative
- Contrastive
- No negative examples
  - Predictive
  - Bootstrapping
  - Regularization