Self-supervised Learning for Speech and Image

Hung-yi Lee
Review: Self-supervised Learning for Text

4 ways to use BERT
https://youtu.be/gh0hewYkjgo
Self-supervised Learning for **Speech**

Speech Recognition

unlabeled speech data

“How are you?”

Downstream Model

Speech version of BERT?
Speech processing Universal PERformance Benchmark (SUPERB)

Published at IS 2021
Published at ACL 2022
SUPERB: Speech processing Universal PERformance Benchmark

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

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, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy T. Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, Hung-yi Lee

To be appeared at ACL 2022
https://arxiv.org/abs/2203.06849
Speech processing **Universal PERformance Benchmark (SUPERB)**

• To learn more:

  ![QR Code](https://youtu.be/GTjwYzFG54E)
  (English version)

  ![QR Code](https://youtu.be/MpsVE60iRLM)
  (Mandarin version)

• Toolkit – S3PRL: https://github.com/s3prl/s3prl
Self-supervised Learning for Image Recognition

- Image Recognition
- Object Detection
- Semantic Segmentation
- Visual Navigation

1. Generative Approaches

BERT series

GPT series
Masking

BERT series

How about speech?

Mockingjay
mimic sound it hears

Learn to reconstruct

Some of the input are masked

masked
masked
Masking

• Smoothness of acoustic features

• Masking strategies for speech

  Learn more speaker information in this way

Predicting Future

APC = Autoregressive Predictive Coding

Linear classifier

For text:

For speech:

How about speech?

How about image?

https://openai.com/blog/image-gpt/
2. Predictive Approach

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

Can a model learn without generation?
Image - Predicting Rotation

https://arxiv.org/abs/1803.07728
Image – Context Prediction

https://arxiv.org/abs/1505.05192

Similar idea on Speech

**Predict Simplified Objects**

Speech

Image
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

Encoder

Encoder

Encoder

positive

negative

as close as possible

as far as possible
SimCLR


Encoder

as close as possible

Encoder

as far as possible

Data Augmentation

random cropping, color distortions, Gaussian blur, etc.

Encoder

positive

Encoder

negative

MoCo
https://arxiv.org/abs/1911.05722

MoCo v2
Contrastive Learning for Speech

CPC
https://arxiv.org/abs/1807.03748

Wav2vec

GRU in CPC, CNN in Wav2vec

Predicter

Encoder

...
Contrastive Learning for Speech

VQ-wav2vec

How to train with quantization:
https://youtu.be/JZvEzb5PV3U
Contrastive Learning for Speech

VQ-wav2vec + BERT

Discrete BERT
Contrastive Learning for **Speech**

**Wav2vec 2.0**

Continuous input is critical
Quantized target improves performance

Jointly trained

Why not formulated as typical classification?
Alternative way to understand Wav2vec 2.0

Is BERT contrastive learning?
Classification vs. Contrastive

**Classification**

= 

**Contrastive**

Last layer output

Embedding for each class
Alternative way to understand Wav2vec 2.0

Infinite negative examples

Limited numbers of negative examples

BERT
Alternative way to understand Wav2vec 2.0

Encoder (BERT architecture)
Alternative way to understand Wav2vec 2.0

Encoder (BERT architecture)

Encoder
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

Encoder

as close as possible

Encoder

Collapse!

Encoder

Predictor

Update

Encoder

Copy

positive

positive
Alternative way to understand Bootstrapping

**Typical Knowledge Distillation**

- **Teacher**
- **Student**
- **Predictor**

The diagram illustrates the process of knowledge distillation and bootstrapping. The teacher model is fixed and serves as a source of knowledge. The student model updates its parameters to be as close as possible to the teacher's predictions. Over time, the student becomes the new teacher, and the process is repeated. This cycle helps in transferring knowledge from a teacher model to a student model, allowing the student to learn from the teacher's experience and improve its performance.
Bootstrapping Approaches

• Image
  • Bootstrap your own latent (BYOL)
  • Simple Siamese (SimSiam)

• Speech
  • Data2vec: the student learns from multiple layers of the teacher
    • https://arxiv.org/abs/2202.03555

\[
\theta \leftarrow \lambda \theta + (1 - \lambda) \theta'
\]

**BYOL**

Teacher Encoder

\[ \theta \]

Student Encoder

\[ \theta' \]

Moving Average

positive
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

**Invariance**

as close as possible

**Variance**

Variance larger than a threshold

Prevent collapse

Audio: DeLoRes
https://arxiv.org/abs/2203.13628

Encoder

Encoder

Encoder

Encoder

Encoder
Concluding Remarks

<table>
<thead>
<tr>
<th>Generative</th>
<th>GPT for image</th>
<th>Speech / Audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive</td>
<td>Rotation Prediction, etc.</td>
<td>HuBERT</td>
</tr>
<tr>
<td>Contrastive</td>
<td>SimCLR, MoCo</td>
<td>CPC, Wav2vec series</td>
</tr>
<tr>
<td>Bootstrapping</td>
<td>BYOL, SimSiam</td>
<td>Data2vec</td>
</tr>
<tr>
<td>Regularization</td>
<td>Barlow Twins, VICReg</td>
<td>DeLoRes</td>
</tr>
</tbody>
</table>