Mockingjay: Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders

Andy T. Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, Hung-yi Lee

Speech Lab, National Taiwan University (NTU)
Introduction

NLP BERT: Language Representation Learning

Unsupervised pre-train on text

Unsupervised pre-train on text

Text tokens
Introduction

NLP BERT: Language Representation Learning

Usage:
Extracts features for downstream NLP models
(can also be fine-tuned)

Unsupervised pre-train on text

Text tokens
Introduction

NLP BERT: Language Representation Learning

Usage:
Extracts features for downstream NLP models (can also be fine-tuned)

Speech BERT: Speech Representation Learning

Unsupervised pre-train on text

Unsupervised pre-train on speech

Acoustic Frames
Introduction

NLP BERT: Language Representation Learning

Usage:
Extracts features for downstream NLP models (can also be fine-tuned)

Speech BERT: Speech Representation Learning

Usage:
Extracts features for downstream SLP models (can also be fine-tuned)

Unsupervised pre-train on text

Unsupervised pre-train on speech
A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

CPC
Phone / Speaker

[1] Representation Learning with Contrastive Predictive Coding
A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

[1] Representation Learning with Contrastive Predictive Coding

April, 2019
MIT

A View of Recent Unsupervised Speech Representation Learning Approaches

- **CPC** (July, 2018, DeepMind)
  - Phone / Speaker
  - [1] Representation Learning with Contrastive Predictive Coding

- **APC** (April, 2019, MIT)
  - Phone / Speaker

**Common heuristic:**
They both encode past information and predict information about future frames.
A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

CPC
Phone / Speaker

[1] Representation Learning with Contrastive Predictive Coding

April, 2019
MIT

APC
Phone / Speaker


October, 2019
Ours

Mockingjay

Mockingjay
A View of Recent Unsupervised Speech Representation Learning Approaches

July, 2018
DeepMind

[1] Representation Learning with Contrastive Predictive Coding

CPC
Phone / Speaker

April, 2019
MIT


APC
Phone / Speaker

October, 2019
Ours

Mockingjay
Phone / Speaker / Sentiment

Mockingjay

Ref: [1] Representation Learning with Contrastive Predictive Coding
Pre-Training Task: Masked Acoustic Model

(A, B, C, D, E, F, G, H, I, J, K) Real Frames (Phoneme level, Spectrogram)
Pre-Training Task: Masked Acoustic Model

Real Frames (Phoneme level, Spectrogram)

Masking Probabilistic Policy

Masked Frames (Phoneme level, Spectrogram)
Pre-Training Task: Masked Acoustic Model

- **Phoneme level, Spectrogram**

- **Mockingjay** (Bidirectional, Self-Attention)
- **Transformer Encoders**
- **Mockingjay Representations**
- **Masked Frames** (Phoneme level, Spectrogram)

- A  B  C  D  E  F  G  H  I  J  K  Real Frames
Pre-Training Task: Masked Acoustic Model

Mockingjay

Transformer Encoders (Bidirectional, Self-Attention)

Masked Frames (Phoneme level, Spectrogram)

Real Frames

Mockingjay Representations
Pre-Training Task: Masked Acoustic Model

- Masking: L1 Loss on Predicted Frames
- Prediction Head: Feed forward
- Mockingjay Representations: (Bidirectional, Self-Attention)
- Transformer Encoders: (Phoneme level, Spectrogram)
- Masked Frames: A B 0 D E 0 0 H 0 J K

Mockingjay

(A) B C D E F G H I J K

Phoneme level, Spectrogram
Pre-Training Task: Masked Acoustic Model

- Considers the whole utterance
- Reconstructs from corrupted input

Transformer Encoders

Mockingjay

Prediction Head

L1 Loss on Predicted Frames

Mockingjay Representations

Transformer Encoders

Masked Frames

Bidirectional, Self-Attention

(Phoneme level, Spectrogram)

(feed forward)
Probabilistic Policy for Masking Frames

1) Select 15% of the frames for prediction (highlighted in green).
Probabilistic Policy for Masking Frames

1) Select 15% of the frames for prediction (highlighted in green).

2) For all selected frames:
   - mask to zero 80% of the time
   - replace randomly 10% of the time
   - leave untouch 10% of the time

Mask all 15%
1) Select 15% of the frames for prediction (highlighted in green).

2) For all selected frames:
   - mask to zero 80% of the time
   - replace randomly 10% of the time
   - leave untouch 10% of the time
Probabilistic Policy for Masking Frames

1) Select 15% of the frames for prediction (highlighted in green).

2) For all selected frames:
   - mask to zero 80% of the time
   - replace randomly 10% of the time
   - leave untouch 10% of the time

Mask all 15%
Replace all 15%
Do nothing, frames remain the same
Input Feature: Masked Spectrogram
Input Feature: Masked Spectrogram

- Masked to Zero
- 80-dim mel-spectrogram
- First derivative
Visualizations

Masked Frames

Mockingjay

L1 Loss on Predicted Frames

Prediction Head

Mockingjay Representations

Transformer Encoders

Masked Frames
The model was able to reconstruct spectrogram form hidden representations.
Migrating from text to speech

**Acoustic Features**: long and locally smooth in nature,

need to 1) **shorten the sequence** and 2) **mask over a longer span**
Migrating from text to speech

**Acoustic Features**: long and locally smooth in nature, need to 1) **shorten the sequence** and 2) **mask over a longer span**

Address the long and smooth problem with: *Downsampling*, and *consecutive masking*
Migrating from text to speech

**Acoustic Features:** long and locally smooth in nature,

need to **1) shorten the sequence** and **2) mask over a longer span**

Address the long and smooth problem with:

*Downsampling*, and *consecutive masking*

\[ R=3 \]
Migrating from text to speech

**Acoustic Features**: long and locally smooth in nature, need to 1) shorten the sequence and 2) mask over a longer span

Address the long and smooth problem with: *Downsampling*, and *consecutive masking*
Model Architecture

Mockingjay

L1 Loss on Predicted Frames
Prediction Head
Mockingjay Representations
Transformer Encoders
Masked Frames
Model Architecture

Transformer Encoders

- Add & Norm
- Feed Forward
  \[ F_{\text{dim}} \]
- Multi-Head Attention
  \[ A_{\text{num}} \]
- H_{\text{dim}}

Mockingjay

- Prediction Head
- L1 Loss on Predicted Frames
- Masked Frames

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>H</td>
<td></td>
<td>P</td>
<td>H</td>
<td></td>
<td>P</td>
<td>H</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Transformer Encoders

Mockingjay Representations
Model Architecture

- **H_dim = 768**
- **F_dim = 3072**
- **A_num = 12**

- Train on LibriSpeech **360 hrs**
- Pre-train steps = **500k**
- Fine-tune steps = **50k** (2-epochs)
Incorporating with Downstream Tasks

1) Feature Extraction
Incorporating with Downstream Tasks

1) Feature Extraction
Incorporating with Downstream Tasks

1) Feature Extraction

Mockingjay

Frozen

Used for feature extraction

Pre-train

L1 Loss on Predicted Frames

Prediction Head

Mockingjay Representations

Transformer Encoders

Masked Frames
Incorporating with Downstream Tasks

1) Feature Extraction

Which class?

Trained with little paired data

Frozen

Used for feature extraction

Mockingjay

Pre-train

L1 Loss on Predicted Frames
Prediction Head
Mockingjay Representations
Transformer Encoders
Masked Frames
Incorporating with Downstream Tasks

2) Weighted Sum from All Layers (WS)

Mockingjay

Trained with little paired data

Classifier

Frozen
Incorporating with Downstream Tasks

2) Weighted Sum from All Layers (WS)

Mockingjay

Classifier

Which class?

Trained with little paired data

Frozen

Repr

Layer1

Layer2

Layer3

0.5

0.2

0.3

WS: Learn weighted sum on all layers

Similar to ELMo in NLP
Incorporating with Downstream Tasks

3) Fine-tune (FT2)
Incorporating with Downstream Tasks

3) Fine-tune (FT2)
Incorporating with Downstream Tasks

3) Fine-tune (FT2)

- Fine-tune with little paired data
- Not Frozen

Classifier

Mockingjay

Initialize for FT

Pre-train

L1 Loss on Predicted Frames
Prediction Head
Mockingjay Representations
Transformer Encoders
Masked Frames
Experiments

We report results on 3 different downstream tasks:

- Phoneme Classification

- Speaker Recognition

- Sentiment Classification on spoken content
Experiments

We report results on 3 different downstream tasks:

- **Phoneme Classification** (72 classes):
  Train: LibriSpeech 360 / Test: LibriSpeech test-clean

- **Speaker Recognition**

- **Sentiment Classification on spoken content**
Experiments

We report results on 3 different downstream tasks:

- **Phoneme Classification (72 classes):**
  Train: LibriSpeech 360 / Test: LibriSpeech test-clean

- **Speaker Recognition (63 classes):**
  Train: 90% of LibriSpeech 100 / Test: 10% of LibriSpeech 100

- **Sentiment Classification on spoken content (2 classes):**
  To demonstrate domain invariant transferability, we use another dataset: MOSEI [3]
Experiments - 1/3

<table>
<thead>
<tr>
<th>Acoustic Features</th>
<th>Phoneme Classification</th>
<th>Speaker Recognition</th>
<th>Sentiment Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mel Features</td>
<td>49.1</td>
<td>70.1</td>
<td>64.6</td>
</tr>
<tr>
<td>BASE</td>
<td>60.9</td>
<td>94.5</td>
<td>67.4</td>
</tr>
<tr>
<td>LARGE</td>
<td>64.3</td>
<td>96.3</td>
<td>70.1</td>
</tr>
</tbody>
</table>

Consistent results over all three tasks: Mel < BASE < LARGE
### Experiments - 2/3

<table>
<thead>
<tr>
<th>Acoustic Features</th>
<th>Phoneme Classification</th>
<th>Speaker Recognition</th>
<th>Sentiment Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mel Features</td>
<td>49.1</td>
<td>70.1</td>
<td>64.6</td>
</tr>
<tr>
<td>BASE</td>
<td>60.9</td>
<td>94.5</td>
<td>67.4</td>
</tr>
<tr>
<td>LARGE</td>
<td>64.3</td>
<td>96.3</td>
<td>70.1</td>
</tr>
<tr>
<td>LARGE-WS</td>
<td><strong>69.9</strong></td>
<td><strong>96.4</strong></td>
<td><strong>71.1</strong></td>
</tr>
</tbody>
</table>

Consistent results over all three tasks:
LARGE < LARGE-WS
<table>
<thead>
<tr>
<th>Acoustic Features</th>
<th>Phoneme Classification</th>
<th>Speaker Recognition</th>
<th>Sentiment Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mel Features</td>
<td>49.1</td>
<td>70.1</td>
<td>64.6</td>
</tr>
<tr>
<td>BASE</td>
<td><strong>60.9</strong></td>
<td><strong>94.5</strong></td>
<td><strong>67.4</strong></td>
</tr>
<tr>
<td>LARGE</td>
<td>64.3</td>
<td>96.3</td>
<td>70.1</td>
</tr>
<tr>
<td>LARGE-WS</td>
<td>69.9</td>
<td>96.4</td>
<td><strong>71.1</strong></td>
</tr>
<tr>
<td>BASE-FT2</td>
<td><strong>84.3</strong></td>
<td><strong>98.1</strong></td>
<td>68.5</td>
</tr>
<tr>
<td>APC [2]</td>
<td>74.1</td>
<td>85.9</td>
<td>66.0</td>
</tr>
</tbody>
</table>

We demonstrate how pre-training on speech can improve supervised training in low resource scenarios, we train with reduced amount of labels.
Low-Resource Experiments - 2/6

Phone accuracy (test-clean, %)

Amount of labeled data (train-clean-360, hr)

Mel < BASE
Low-Resource Experiments - 3/6

Mel < BASE < LARGE
Low-Resource Experiments - 4/6

LARGE < LARGE-WS with an avg 5.75% improvement
With 0.1% of labels, LARGE-WS (52.8%) outperformed Mel (49.1%) that uses all 100% hours of labeled data.
Low-Resource Experiments - 5/6

Phone accuracy (test-clean, %)

Amount of labeled data (train-clean-360, hr)

All < BASE-FT2
Low-Resource Experiments - 5/6

Phone accuracy (test-clean, %)

Amount of labeled data (train-clean-360, hr)

All < BASE-FT2
Low-Resource Experiments - 5/6

With 0.1% of labels, BASE-FT2 (57.9%) outperformed Mel (49.1%) that uses all 100% hours of labeled data.
APC works well on full resource but fails to generalize for limited labeled data.
Conclusion

We conclude that unsupervised Mockingjay improves supervised training!
Links

This slide (with speaker notes) can be found here:

Our code and implementation can be found here:
https://github.com/andi611/Mockingjay-Speech-Representation