Self-supervised for speech processing

Facebook AI Research

Alexei Baevski, Alexis Conneau, Steffen Schneider, Henry Zhou, Abdelrahman Mohamed, Anuroop Sriram, Naman Goyal, Wei-Ning Hsu, Michael Auli, Kritika Singh, Yatharth Saraf, Geoffrey Zweig, Qiantong Xu, Tatiana Likhomanenko, Paden Tomasello, Ronan Collobert, Gabriel Synnaeve
Speech technology

Video captioning

Mobile devices

Home devices
Speech applications

- **Speech to text (Speech recognition)**
- Text to speech
- Keyword spotting (“Hey Alexa/Portal”)
- Speaker identification
- Language identification
- Speech translation
Overview

- Speech recognition
- Speech processing with less supervision / self-supervised learning
- Cross-lingual self-supervised learning for speech
I like black tea with milk
Traditional automatic speech recognition (ASR)

\[ W^* = \arg \max_W p(W|X) \]
\[ W^* = \arg \max_W p(F|W)p(W) \]
Traditional automatic speech recognition (ASR)

- Represent words as sequences of phonemes
- hello  =  h   eh   l   ow
- Distinct units of sound to distinguish words
Traditional automatic speech recognition (ASR)

\[
W^* = \arg \max_W p(W|X)
\]

\[
W^* = \arg \max_W \sum_Q p(F|Q)p(Q|W)p(W)
\]
Traditional automatic speech recognition (ASR)

\[ W^* = \arg \max_W p(W|X) \]

\[ W^* = \arg \max_W \sum_Q p(F|Q)p(Q|W)p(W) \]

Focus of this talk
Feature representation

• Typical sample rates for speech: 8KHz, 16KHz.
• Traditionally: build spectrogram
Spectrogram

- Small window, e.g., 20ms of waveform
- Compute FFT and take magnitude
- Describes frequency content in local window
Spectrogram

- Concatenate frames from adjacent windows to form a spectrogram
Self-supervised speech representation learning
Training speech recognition models

I like black tea with milk

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

( , cat )

potential train/test mismatch

Need to annotate lots of data!
Meanwhile in other fields

Pre-training in NLP

- Glue score:
  - 50
  - 58
  - 66
  - 74
  - 82
  - 90

- Models:
  - biLSTM
  - ELMo
  - GPT
  - BERT

Pre-training in Computer Vision

Thanks to Priya Goyal for sharing the vision graph.
Unsupervised / Self-supervised Pre-training

- Learn good representations **without labels**

- NLP: Predict occluded parts of sentence

- Vision: make representations invariant to augmentations
Learning good representations of audio data from unlabeled audio
I like tea

Speech recognition

Pre-trained model

0.1
0.5
...
-0.9
Ich mag Tee

Speech translation

Pre-trained model
Pre-trained model

Audio event detection

"music"

0.1
0.5
...
-0.9
wav2vec: Latent speech audio representations

- CNN encodes waveform as latent representations $z_t$ spanning 25ms each
- Another CNN builds context representations $c_t$ of ~300ms
- Training: predict future latents $p(z_{t+1}|c_t), p(z_{t+2}|c_t), ...$
- Inference: feed $c_t$ into traditional ASR system - instead of logmel etc.
vq-wav2vec: Learning **discrete** latent speech representations

- Human language has a relatively fixed number of possible sounds.
- Mimic this by constraining the latents to a fixed number
- **Vector quantize** the latents = assign each $z_t$ to an entry in a fixed size codebook $q$ by, e.g., online k-means
- Learn an inventory of acoustic units, basic sounds
wav2vec 2.0

- Bi-directional contextualized representations
- Vector quantized targets for training
Objective

Codebook diversity penalty to encourage more codes to be used

\[
\mathcal{L}_m = - \log \frac{\exp(sim(c_t, q_t) / \kappa)}{\sum_{\tilde{q} \sim Q_t} \exp(sim(c_t, \tilde{q}) / \kappa)}
\]

- Cosine similarity
- Context representation
- Discrete latent speech representation
- Negative samples
- Temperature
Masking

- Sample starting points for masks without replacement, then expand to 10 time-steps (1 time-step is 25ms but 10ms stride)

- Spans can overlap

- For a 15s sample, ~49% of the time-steps masked with an average span length of ~300ms
Fine-tuning

- Add a single linear projection on top into target vocab and train with CTC loss with a low learning rate (CNN encoder is not trained).

- Use modified SpecAugment in latent space to prevent early overfitting

- Uses wav2letter decoder with the official 4gram LM and Transformer LM
Results

Librispeech 960h setup + Neural LM

Word error rate

- ContextNet (supervised only)
- wav2vec (supervised only)
- Noisy Student (60k-h unlabeled)
- wav2vec (60k-h unlabeled)
Results

Librispeech 960h setup + Neural LM

<table>
<thead>
<tr>
<th>test other</th>
<th>Word error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ContextNet (supervised only)</td>
<td>4.1</td>
</tr>
<tr>
<td>wav2vec (supervised only)</td>
<td>4.6</td>
</tr>
<tr>
<td>Noisy Student (60k-h unlabeled)</td>
<td>3.4</td>
</tr>
<tr>
<td>wav2vec (60k-h unlabeled)</td>
<td>3.3</td>
</tr>
</tbody>
</table>
Results

Librispeech 960h setup + Neural LM

Word error rate

Low resource setup

Word error rate

<table>
<thead>
<tr>
<th>Test Other</th>
<th>Word Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Librispeech 960h setup + Neural LM</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ContextNet (supervised only)</td>
</tr>
<tr>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>2.3</td>
<td>1.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Other</th>
<th>Word Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low resource setup</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noisy Student 100h labeled (+860h unlabeled)</td>
</tr>
<tr>
<td>0</td>
<td>8.6</td>
</tr>
<tr>
<td>2.75</td>
<td>5</td>
</tr>
<tr>
<td>5.5</td>
<td>5</td>
</tr>
<tr>
<td>8.25</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
</tr>
</tbody>
</table>
Results

Effects of model size and amount of unlabeled data

Word error rate on test-other

<table>
<thead>
<tr>
<th>Labeled data</th>
<th>Base (100m)</th>
<th>Large (300m)</th>
<th>+ 60k-h</th>
</tr>
</thead>
<tbody>
<tr>
<td>10m</td>
<td>12.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1h</td>
<td>9.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10h</td>
<td>7.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100h</td>
<td>6.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>960h</td>
<td>4.8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Effects of model size and amount of unlabeled data

Word error rate on test-other

Labeled data

Base (100m)  Large (300m)  + 60k-h

<table>
<thead>
<tr>
<th>Labeled data</th>
<th>Base (100m)</th>
<th>Large (300m)</th>
<th>+ 60k-h</th>
</tr>
</thead>
<tbody>
<tr>
<td>10m</td>
<td>12.9</td>
<td>10.8</td>
<td></td>
</tr>
<tr>
<td>1h</td>
<td>9.3</td>
<td>7.6</td>
<td></td>
</tr>
<tr>
<td>10h</td>
<td>7.8</td>
<td>6.1</td>
<td></td>
</tr>
<tr>
<td>100h</td>
<td>6.3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>960h</td>
<td>4.8</td>
<td>4.1</td>
<td></td>
</tr>
</tbody>
</table>
Results

Effects of model size and amount of unlabeled data

Word error rate on test-other

Labeled data

- Base (100m)
- Large (300m)
- + 60k-h

<table>
<thead>
<tr>
<th>Labeled data</th>
<th>Base (100m)</th>
<th>Large (300m)</th>
<th>+ 60k-h</th>
</tr>
</thead>
<tbody>
<tr>
<td>10m</td>
<td>12.9</td>
<td>10.8</td>
<td>8.2</td>
</tr>
<tr>
<td>1h</td>
<td>9.3</td>
<td>7.6</td>
<td>5.8</td>
</tr>
<tr>
<td>10h</td>
<td>7.8</td>
<td>6.1</td>
<td>4.9</td>
</tr>
<tr>
<td>100h</td>
<td>6.3</td>
<td>5</td>
<td>4.0</td>
</tr>
<tr>
<td>960h</td>
<td>4.1</td>
<td>4.8</td>
<td>3.3</td>
</tr>
</tbody>
</table>
Examples (10 min labeled data)

HYP (no LM): she SESED and LUCHMAN GAVE 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
wav2vec on HuggingFace

- HuggingFace is a popular NLP model zoo
- HuggingFace community fine-tuned our models to do speech recognition in 73 languages.
Pre-training and self-training
Pre-training and self-training

- Self-training very successful in speech recognition: generate pseudo-labels
Pre-training and self-training

- Self-training very successful in speech recognition: generate pseudo-labels

Supervised model
Pre-training and self-training

- Self-training very successful in speech recognition: generate pseudo-labels
Pre-training and self-training

- Self-training very successful in speech recognition: generate pseudo-labels
Pre-training and self-training

- Self-training very successful in speech recognition: generate pseudo-labels
- Do both have the same effect?
- Recipe: pre-train on the unlabeled data, pseudo-label, fine-tune pre-trained model
Amount of labeled data used

Librispeech benchmark, WER on test-other

960h labeled

Data based on Papers with Code (25 Oct 2020)
Amount of labeled data used

960h labeled

Data based on Papers with Code (25 Oct 2020)

Librispeech benchmark, WER on test-other

Results based on wav2vec 2.0
XLSR: cross lingual speech representation learning with wav2vec
Why *cross-lingual* self-supervised learning

- Little labeled data -> little unlabeled data
- Leverage unlabeled data from high-resource languages
- To improve performance on low-resource languages
- One model for each of the 6500 languages, for each domain? No.
- Instead: one pertained model for all languages
Meanwhile in multilingual research

Cross-lingual understanding (XLU)

<table>
<thead>
<tr>
<th>Model</th>
<th>XNLI Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>biLSTM</td>
<td>50</td>
</tr>
<tr>
<td>mBERT</td>
<td>58</td>
</tr>
<tr>
<td>XLM</td>
<td>66</td>
</tr>
<tr>
<td>XLM-R</td>
<td>82</td>
</tr>
</tbody>
</table>

Multilingual machine translation
XLSR: cross lingual speech representation learning with wav2vec
XLSR: Results - cross-lingual transfer

XLSR significantly outperforms previously published approaches on CommonVoice/BABEL

CommonVoice results:

- **Phoneme Error Rate (PER)**
- **High-resource languages**
- **Low-resource languages**
XLSR: Results - cross-lingual transfer

XLSR significantly outperforms previously published approaches on CommonVoice/BABEL

**CommonVoice results:**

- XLSR-Mono
- XLSR-10 (Base)
- XLSR-10 (Large)

**BABEL (average) results:**

- Multi-BLSTM^+VGG
- XLSR-Mono
- XLSR-10 (Base)
- XLSR-10 (Large)
XLSR: Results - multilingual fine-tuning

Multilingual finetuning leads to *one model for all languages* with little loss in performance.
XLSR: Results - multilingual fine-tuning

Multilingual finetuning leads to one model for all languages with little loss in performance.
XLSR: Results - impact of language similarity

Language similarity plays an important role in cross-lingual transfer

Similar higher-resource language data helps the most for low-resource language

- 5h Italian + 1h labeled + 50h unlabeled in another language
XLSR: Results - impact of language similarity

Language similarity plays an important role in cross-lingual transfer

Similar higher-resource language data helps the most for low-resource language

<table>
<thead>
<tr>
<th>Language</th>
<th>Phoneme Error Rate (PER)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>50</td>
</tr>
<tr>
<td>Italian</td>
<td>12.5</td>
</tr>
<tr>
<td>Spanish</td>
<td>25</td>
</tr>
<tr>
<td>German</td>
<td>25</td>
</tr>
<tr>
<td>English</td>
<td>25</td>
</tr>
<tr>
<td>Russian</td>
<td>25</td>
</tr>
<tr>
<td>Kazakh</td>
<td>25</td>
</tr>
<tr>
<td>Chinese</td>
<td>25</td>
</tr>
</tbody>
</table>

5h Italian + 1h labeled + 50h unlabeled in another language
XLSR: Analysis of discrete latent speech representations

PCA visualization of latent discrete representations from the multilingual codebook

Similar languages tend to share discrete tokens and thus cluster together.
Conclusion

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

- Cross-lingual training improves low-resource languages.

- Pre-training and self-training are complementary.

- 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.
Thank you