Building Intelligent Machines that Learn from Human Speech

Michael Auli

FAIR
California

Arun Babu
Alexis Conneau
Steffen Schneider
Henry Zhou
Abdelrahman Mohamed
Jiatao Gu
Naman Goyal
Wei-Ning Hsu
Alexei Baevski
Michael Auli

Kushal Lakhotia
Andros Tjandra
Kritika Singh
Yatharth Saraf
Geoffrey Zweig
Qiantong Xu
Tatiana Likhomanenko
Paden Tomasello
Ronan Collobert
Gabriel Synnaeve
Speech is Rich in Information

- Voice carries a lot of information: what you say/how you say it.
- Stress in our voice, intonation, and other paralinguistic features.
- Children learn a lot by listening to others.
Speech is Natural & Interactive

- Intelligent machine communicating using speech.
- High-latency (text) vs. low-latency (speech).
- We speak faster than we can type (2-3 words/sec).
Speech is Ubiquitous

- Large language models are trained on a lot of data
- YouTube adds 500 hours of video data/minute. Up to 2.6T tokens/year.
Access to Technology and Information
World’s Languages Dying Off Rapidly

By John Noble Wilford
Sept. 18, 2007

Of the estimated 7,000 languages spoken in the world today, linguists say, nearly half are in danger of extinction and are likely to disappear in this century. In fact, they are now falling out of use at a rate of about one every two weeks.
Speech Applications

- Speech to text/speech recognition - dictation etc.
- Text to speech - reading out aloud
- Keyword spotting - “Hey Alexa/Portal”
- Speaker identification - is it your voice?
- Language identification
- Speech translation
This Talk

- **wav2vec**: a self-supervised algorithm for speech representations.
- **wav2vec-U**: self-supervised learning enables unsupervised speech recognition.
- **data2vec**: unified objective for self-supervised learning in multiple modalities.
Self-supervised Speech Representation Learning
Supervised Machine Learning

Need to annotate lots of data!

potential train/test mismatch

Not how humans learn!
Supervised Machine Learning
Self-supervised Learning

- Learn good data representations (structure, features etc.) \textbf{without labels}
- \(|\text{Unlabeled data}| \gg |\text{Labeled data}|\)
- Use representations to solve the task
**Supervised learning** simultaneously performs representation learning of the data and associating these features with labels.

**Limitation:** relies on labeled data to learn feature encoding.
Self-supervised learning:
1/ representation learning of the data
2/ learn to associate labels with the representations

Reduces reliance on labeled data!
Training Speech Recognition Models

I like black tea with milk

- Train on 1,000s of hours of data for good systems.
- Many languages, dialects, domains etc.
I like tea

Speech recognition

Pre-trained model

0.1
0.5
...
-0.9
Speech translation

Ich mag Tee

wav2vec
Audio event detection

"music"

wav2vec

0.1
0.5
...
-0.9
wav2vec 2.0

- Masked prediction with transformer, bidirectional contextualized representations (similar to BERT).

- But predict what? Learn an inventory of speech units with vector quantization via Gumbel softmax.

- Learning task: Joint VQ & context representation learning.

- Contrast true quantized latent with distractor latents.
• Masked prediction with transformer, bi-directional contextualized representations (similar to BERT).

• But predict what? Learn an inventory of speech units with vector quantization via Gumbel softmax.

• Learning task: Joint VQ & context representation learning.

• Contrast true quantized latent with distractor latents.
wav2vec 2.0

- Masked prediction with transformer, bidirectional contextualized representations (similar to BERT).

- But predict what? Learn an inventory of speech units with vector quantization via Gumbel softmax.

- Learning task: Joint VQ & context representation learning.

- Contrast true quantized latent with distractor latents.
Objective

Codebook diversity penalty to encourage more codes to be used
Masking

- Sample starting points for masks without replacement, then expand to 10 time-steps
- Spans can overlap
- For a 15s sample, ~49% of the time-steps masked with an average span length of ~300ms
Fine-tuning

- Fine-tune model on labeled data for ASR with CTC (or other speech tasks)
- SpecAugment-style regularization & remove quantization
## Results

### High resource setup
(Librispeech 960h labeled)

<table>
<thead>
<tr>
<th>Method</th>
<th>Word error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ContextNet (supervised)</td>
<td>3.3</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>

### Low resource setup
(Librispeech 10min - 100h labeled)

<table>
<thead>
<tr>
<th>Method</th>
<th>Word error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noisy Student 100h</td>
<td>8.6</td>
</tr>
<tr>
<td>wav2vec 100h</td>
<td>7.6</td>
</tr>
<tr>
<td>wav2vec 1h</td>
<td>7.6</td>
</tr>
<tr>
<td>wav2vec 10m</td>
<td>8.3</td>
</tr>
<tr>
<td>wav2vec 10m + (60k-h unlabeled)</td>
<td>11.0</td>
</tr>
</tbody>
</table>

---

**Effective both for high and low-resource settings!**
Results

Effects of model size and amount of unlabeled data

Word error rate on test-other

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

Amount of labeled data
Examples (10 min Labeled Data)

HYP (no LM): she SESED and LUCHMAN GAIVE 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
## Librispeech benchmark, WER on test-other

<table>
<thead>
<tr>
<th>Method</th>
<th>Results (test-other)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSpeech 2 (Baidu '15)</td>
<td>5.2</td>
</tr>
<tr>
<td>Fully Conv ASR (FB '18)</td>
<td>5.2</td>
</tr>
<tr>
<td>tdnn / Kaldi ('18)</td>
<td>3.3</td>
</tr>
<tr>
<td>SpecAugment (Google '19)</td>
<td>3.4</td>
</tr>
<tr>
<td>RW/TH Hybrid ('19)</td>
<td>3.9</td>
</tr>
<tr>
<td>Pseudo-labeling (FB '20)</td>
<td>3.3</td>
</tr>
<tr>
<td>Conformer (Google '20)</td>
<td>2.6</td>
</tr>
<tr>
<td>Noisy Student (Google '20)</td>
<td>5.2</td>
</tr>
<tr>
<td>wav2vec 2.0 (FB, '20)</td>
<td>8.6</td>
</tr>
<tr>
<td>wav2vec 2.0 + Conf. + NST (Google, '20)</td>
<td>8.6</td>
</tr>
<tr>
<td>wav2vec 2.0 (FB, '20)</td>
<td>5.2</td>
</tr>
</tbody>
</table>

Data based on Papers with Code (25 Oct 2020)
Summary

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

- 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.
Unsupervised Speech Recognition
Unsupervised Speech Recognition

- Important step towards agents that can learn without supervision.
- Unsupervised machine translation exists, what about speech?
- Key problem: what are the units in the speech audio?
wav2vec Unsupervised: Key Ideas

- Learn good representations of speech audio
- Unsupervised segmentation of the speech audio into phonemic units
- Learn mapping between speech segments and phonemes using adversarial learning
wav2vec Unsupervised

Step 1: Learn speech representations

Step 2: k-means cluster representations

Step 3: Segment into phonemic units

Step 4: Build segment representations

Step 5: Generate phoneme sequence

Step 6: Phonemize

Step 7: GAN training
Text Data Pre-processing

Phonemize

he spoke soothingly

Unlabeled text

hh iy s ow k s uw dh ih ng l iy
GAN inputs

Unlabeled phonemized text

Unlabeled speech audio

Generator

Discriminator

Segment representations

Phoneme probability distributions

Combine identical phoneme predictions

real or fake

phoneme representations (1-hot vectors)
Generator / Discriminator

- Generator: 1 layer CNN with 90k parameters
  w2v features frozen

- Discriminator: 3 layer CNN

- Train time: 12-15h on a single V100
Training Details

- Unsupervised metric for early stopping, hyper-parameter selection
- Self-training after GAN training (HMM and fine-tuning w2v)
Comparison to Prior Unsupervised Work

Phoneme error rate

<table>
<thead>
<tr>
<th></th>
<th>Phoneme Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chen et al. Matched 2018</td>
<td>26.1</td>
</tr>
<tr>
<td>WAV2VEC-U</td>
<td>11.3</td>
</tr>
</tbody>
</table>
Comparison to Best Supervised Systems

Amount of labeled data used

- 960 hrs.+
- 1 hr.
- 0 hrs.

Word error rate

- DEEP SPEECH 2 2015: 13.25
- Fully Conv ASR 2016: 10.47
- TDNN/KALDI 2018: 7.53
- SPEC-AUGMENT 2019: 5.8
- RWTH HYBRID 2019: 5
- PSEUDO-LABELING 2020: 4
- NOISY STUDENT 2020: 3.4
- WAV2VEC 2.0 2020: 3.3
- WAV2VEC 2.0 + NOISY ST. 2020: 2.6
- WAV2VEC -U: 5.9

Librispeech benchmark
Low-resource Languages

- **Tatar**
  - Fer et al. '17
  - Riviere et al. '20
  - Conneau et al. '21
  - wav2vec-U

- **Kyrgyz**
  - Supervised (Besacier et al. '15)
  - Riviere et al. '20
  - wav2vec-U

- **Swahili**
  - wav2vec-U + ST

*wav2vec-U uses much less speech audio than prior work: 1.8h vs. 17h for Kyrgyz, 4.6h vs. 17h for Tatar*
Discussion

- Very lightweight approach (except for wav2vec 2.0)
- Why does it work? Good audio features are main driver of performance
- Phonemizer still required
- Segment construction
data2vec: A Unified Objective for Self-supervised Learning
Natural Language Processing

BERT

GPT-3
Computer Vision

SimCLR, BYOL, Masked AutoEncoders (MAE), …
Speech: Unsupervised Speech Recognition

CPC, wav2vec 2.0, wav2vec Unsupervised, WavLM, w2v-BERT, HuBERT, ...

Amount of labeled data used

Word error rate

Librispeech benchmark (test-other) compared with the best systems over time. Source: paperswithcode.com
Two Challenges
Modality-specific Learning Algorithms

- Most algorithms developed for one modality - specific designs and learning biases.
- General idea of SSL. Biology of learning (Friston, ’10).
- **This talk:** single objective for vision, speech and text.
Little Focus on Efficiency

- Great progress but model sizes and compute requirements are ever growing.

- Are we using the best algorithms to push the boundaries?

- Scaling an efficient learner may ultimately get you further than an inefficient one.

- **This talk:** compute efficient SSL
A Single Learning Objective
data2vec

- General algorithm that works very well across modalities.
- Same learning objective for each modality.
<table>
<thead>
<tr>
<th>Images</th>
<th>Speech</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td><img src="image1.png" alt="Image of tea" /></td>
<td>I like tea with milk</td>
</tr>
<tr>
<td>Masked</td>
<td><img src="image2.png" alt="Image of tea with milk" /></td>
<td>I like tea milk</td>
</tr>
</tbody>
</table>
masked prediction
data2vec

Original

Masked

Images

Speech

Language

Model in teacher-mode

Model in student-mode

I like tea with milk

I like tea milk
data2vec

Images

Speech

Language

Model in teacher-mode

I like tea with milk

Model in student-mode

I like tea ☕️ milk

contextualized targets

Predict model representation of original input
data2vec

Original

Masked

Images

Speech

Language

Model in teacher-mode

Model in student-mode

Predict model representation of original input

Teacher tracks student parameters

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

- Momentum teacher (Grill et al., ’20, Caron et al., ’21)

- Contextualized targets (Hsu et al., ’21)
Vision Results

ViT-L on ImageNet-1K

Multiple models
- BEiT: 85.2
- PeCO: 86.5

Single models
- MoCo-v3: 84.1
- MAE: 85.9
- MaskFeat: 85.7
- data2vec: 86.6
Speech & NLP Results

Librispeech test-other, Large models

<table>
<thead>
<tr>
<th>Amount of labeled data</th>
<th>Word error rate on test-other</th>
</tr>
</thead>
<tbody>
<tr>
<td>10m</td>
<td>11</td>
</tr>
<tr>
<td>1h</td>
<td>9.4</td>
</tr>
<tr>
<td>10h</td>
<td>8.2</td>
</tr>
<tr>
<td>100h</td>
<td>7.8</td>
</tr>
<tr>
<td>960h</td>
<td>6.2</td>
</tr>
</tbody>
</table>

- wav2vec 2.0
- HuBERT
- WavLM
- data2vec

GLUE score

RoBERTa baseline: 77
Ours: 85
Teacher Representation Construction

Averaging multiple layers helps!

(a) Speech

(b) NLP

(c) Vision

Word error rate

GLUE score

Top-1 valid accuracy

$K$

$K$

$K$
Target Context Size

Contextualized targets improve accuracy!

(a) Speech

(b) Vision
Limitations

- Modality specific feature encoder -> Perceiver work!
- Requires two forward-passes -> data2vec 2.0
Efficient Self-supervised Learning
data2vec 2.0

- MAE: Do not encode masked time-steps.
- Multi-masking: Learn from different views & share target representation.
  - Amortizes the cost of the teacher.
- Result: train with less compute, fewer epochs & smaller batch size.
Example

Images

Teacher encoder

Masked

Example

Masked
Example

Images

Teacher tracks student weights

Masked

Teacher encoder

Student encoder

Predict model representation of original input

Teacher tracks student weights

Example

Masked
Example

Masked

Images

Teacher encoder

Student encoder

Student CNN decoder

Predict model representation of original input

Teacher tracks student weights

≈

≈

Example

Masked

Teacher
c

Predict model
representation of
original input

≈

≈
Teacher tracks student weights

Targets are contextualized

Example

Masked #1

Masked #n

Speech Language Images

I drink black tea

Predict model representation of original input

Example

Masked #1

Masked #n

Targets are contextualized

Predict model representation of original input

Example

Masked #1

Masked #n

Targets are contextualized

Predict model representation of original input

Example

Masked #1

Masked #n

Targets are contextualized

Predict model representation of original input
Compute Efficiency in Vision

ViT-B, pre-train and fine-tune on ImageNet-1K, eval on dev
All training times are for 32 A100 GPUs
Compute Efficiency in Speech

Transformer Base, pre-train on Librispeech, fine-tune on Libri-light 10h, eval on dev-other, no language model
Compute Efficiency in NLP
Vision Results - ViT-L accuracy vs. epochs

Multiple models/external data

Top-1 Accuracy vs. Epochs

- data2vec 2.0
- PeCo
- BEiT
- BEiT-2
- TEC
- MAE
- MaskFeat
- MoCo v3

Multiple models/external data
Conclusion

• A single learning objective can perform very well compared to the best modality-specific algorithms for vision/speech/NLP.

• Contextualized targets lead to a rich learning task which enables efficient training.

• Think about multiple modalities from the outset.
Thank you