Bigger is Not Always Better: The Effect of Context Size on Speech Pre-Training

Sean Robertson, Ewan Dunbar

University of Toronto

sdrobert@cs.toronto.edu,ewan.dunbar@utoronto.ca

VECTOR INSTITUTE

UNIVERSITY OF TORONTO

DATA SCIENCES INSTITUTE

Motivations

- Some speech processing tasks can be performed on less than a second of input; others need much more.
- In the supervised setting (*e.g.* ASR), excess is OK: the network can learn to filter out irrelevant context.
- Pre-trained speech models uncritically adopt the "more is better" mantra of their supervised cousins, but how do they know what's irrelevant?

Hypothesis: Providing too much input as context to a pre-trained speech model is detrimental to phoneme discriminability.

Phoneme Discriminability

- Phonemes are defined as the minimal acoustic unit of speech such that changing the identity of one changes the identity of the containing word.
- Though additional context helps humans distinguish between phonemes, we can do so with high accuracy with only a fraction of a second's exposure.
- Pre-trained speech representations ought to be able to do the same!
- ABX error rates estimate the frequency with which pairs of spans of speech features are more dissimilar *within* phoneme classes than *between*.
- Formally, for sets of all spans of speech features A, B aligned to unique phonemes (of N possible phonemes), the ABX error rate is approximated as:

$$ABX \approx 1 - \frac{1}{N(N-1)} \sum_{A,B} \Delta(A, B), \text{ where}$$

$$\Delta(A,B) = \frac{\sum_{\mathbf{a} \in A} \sum_{\mathbf{b} \in B} \sum_{\mathbf{x} \in A \setminus \{\mathbf{a}\}} I[d(\mathbf{a},\mathbf{x}) < d(\mathbf{b},\mathbf{x})] + \frac{1}{2} I[d(\mathbf{a},\mathbf{x}) = d(\mathbf{b},\mathbf{x})]}{|A|(|A|-1)|B|}$$

- $d(\mathbf{a}, \mathbf{x})$ is computed with Dynamic Time warping over spans $\mathbf{a}$ and $\mathbf{x}$.

Limiting Context

- To limit the context available to speech representations without changing the size of our models, we implement *causal, chunked self-attention*.
- Define query vector $q \in \mathbb{R}^{d_Q}$, key vectors $\mathbf{k}$, $k_t \in \mathbb{R}^{d_K}$, and the same number of value vectors $\mathbf{v}$, $v_t \in \mathbb{R}^{d_V}$. Parametrized by a score function $score : \mathbb{R}^{d_Q} \times \mathbb{R}^{d_K} \to \mathbb{R}$, *attention* produces a vector in $\mathbb{R}^{d_V}$ defined as:

$$attend(q, \mathbf{k}, \mathbf{v}) = \sum_{t=1}^{T} \alpha(q, \mathbf{k}) v_t, \text{ where}$$

$$\alpha(q, \mathbf{k}) = \frac{\exp\left(score(q, k_t)\right)}{\sum_{t'=1}^{T} exp\left(score(q, k_{t'})\right)}.$$

- For fixed context width $W \in \mathbb{N}$ and input sequence $\mathbf{x}^{(in)}$, $x_t^{(in)} \in \mathbb{R}^d$, causal, chunked attention is defined as:

$$x_t^{(out)} = attend\left(x_t^{(in)}, \mathbf{x}^{(in)}_{\max(t-W,1):t}, \mathbf{x}^{(in)}_{\max(t-W,1):t}\right).$$

Pre-training Objective

- Pre-training is a round of unsupervised training designed to transform audio input $\mathbf{x}$ into speech representation vectors $\mathbf{c}$, $c_t \in \mathbb{R}^d$.
- *Contrastive predictive coding* (CPC) learns useful representations $\mathbf{c}$ by using them to *predict* latent vectors $\mathbf{z}$, $\mathbf{x} \mapsto \mathbf{z} \mapsto \mathbf{c}$ and *contrasts* those latent vectors against one another.
- Formally, using $\mathbf{c}$ to generate S prediction sequences, $\mathbf{c} \mapsto \mathbf{v}^{(s)}$, and drawing some number of "distractor" latent vectors $\tilde{z}$ uniformly from $\mathbf{z}$, the CPC loss is defined as:

$$\mathcal{L}_{\text{CPC}} = \frac{1}{S} \sum_{s=1}^{S} \mathcal{L}_{\text{CPC}}^{(s)}, \text{ where}$$

$$\mathcal{L}_{\text{CPC}}^{(s)} = -\frac{1}{T-S} \sum_{t=1}^{T-S} \log \frac{\exp\left(z_{t+s}^{\intercal} v_t^{(s)}\right)}{\sum_{\tilde{z}} \exp\left(\tilde{z}^{\intercal} v_t^{(s)}\right)}.$$

Experiments

Open source URL: `https://github.com/sdrobert/scpc`

- Our experiments primarily measure the impact of context on ABX error rates.
- We modify the baseline CPC architecture used in the Zero Resource Speech Challenges:
  - $\mathbf{x} \mapsto \mathbf{z}$ is a 5-layer stack of dilated convolutions;
  - $\mathbf{z} \mapsto \mathbf{c}$ is a single Transformer layer with causal, chunked self-attention; and
  - $\mathbf{c} \mapsto \mathbf{v}^{(s)}$ is a single Transformer layer with causal (non-chunked) self-attention, which is thrown away after pre-training.
- We control the context width W in the causal, chunked self-attention layer.
  - W ranges between 2 frames (40ms) and 128 frames (1300sm).
  - The phoneme duration in the training data (LibriSpeech) is 90 ± 50 ms.
- We first measure the significance of context width on ABX error rates *via* a repeated-measures ($N = 5$) one-way ANOVA.
- We probe the source of significance via *post-hoc* Wilcoxon signed-rank tests.
- In addition, we run a series of auxiliary experiments to determine whether context effects are robust:
  - we increase the number of chunked Transformer layers – *2-layer* or *4-layer*;
  - we replace Transformer layers with convolutions, either with a fixed output size – *conv (fixed H_2)* – or a fixed number of layer parameters – *conv (fixed param)*;
  - we extend the duration of training – *long train*;
  - we increase the amount of training data – *960h*;
  - we set $\mathcal{L}_{\text{CPC}} = \mathcal{L}_{\text{CPC}}^{(6)}$ – *last at $S = 6$*; and
  - we replace the CPC loss with a masked prediction loss – *BEST-RQ*.
- Finally, we replicate the SUPERB ASR task on the best-performing models from the main setup.

Results

Main result: context has a significant effect on phoneme discriminability, with better performance on shorter windows.

ABX error rates (%). Lower is better.

	context width W						
condition	2	4	8	16	32	64	128
main	15.6	**14.2**	15.1	15.3	14.8	16.9	17.7
main (best)	15.3	**12.6**	13.3	13.8	13.6	16.2	15.9
2-layer	13.7	15.5	15.8	**12.6**	13.6	16.8	13.0
4-layer	13.8	**13.0**	14.1	15.8	15.8	16.8	17.1
conv (fixed param)	17.0	**14.6**	16.8	17.0	19.0	19.4	20.7
conv (fixed H_2)	15.0	14.6	16.9	**14.2**	50.0	22.3	43.7
long train	13.4	**13.0**	14.0	14.4	14.3	14.7	15.0
960h	–	–	**17.2**	–	–	–	18.5
last at $S = 6$	14.0	13.9	**13.1**	15.4	13.3	14.9	16.0
BEST-RQ	27.6	**24.1**	24.5	26.6	26.1	28.0	25.2

ASR error rates (%) with LMs. Lower is better.

	partition			
context width W	dev-clean	dev-other	test-clean	test-other
2	19.1	39.4	18.0	43.9
4	**16.5**	**36.6**	**15.4**	41.0
8	17.0	37.3	16.1	**40.9**
16	19.4	39.9	17.9	43.9
32	18.1	38.4	17.1	42.6
64	22.7	42.9	21.3	48.5
128	19.9	39.6	18.9	44.2

- $F(6, 28) = 6.026, p < 0.001$, with Wilcoxon tests significant when compared widths are between a $W \leq 32$ model and a $W > 32$ model.
- Best performer is always one of $W \in \{4, 8, 16\}$, even for ASR.
- $W = 128$ occasionally nears best performer, though not reliably.
- ABX performance does not map nicely to validation loss.

Discussion and Conclusions

- Experiments lend strong support to our hypothesis that one can have "too much" context when pre-training for downstream tasks focusing on short-term phenomena.
- The preference for short windows in ASR may be due to a large, flexible dowstream model (~ 44 million parameters).

Take-home: rather than look for a "universal context," we recommend learning heterogenous representations of multiple contexts.

LaTeX TikZposter