INVESTIGATION ON INSTANCE MIXUP REGULARIZATION STRATEGIES FOR SELF-SUPERVISED SPEAKER REPRESENTATION LEARNING

WOO HYUN KANG,
JAHANGIR ALAM,
ABDERRAHIM FATHAN
SELF-SUPERVISED SPEAKER REPRESENTATION LEARNING

• Why do we need self-supervised speaker representation learning?
  • Nowadays, vast speech data can be obtained for training a speaker verification system
    • E.g., YouTube, Soundcloud, TikTok, etc.
  • However, most of these speech samples do not have any speaker labels
    • Also, collecting speaker labeled speech samples can be very expensive in terms of resources
Although the conventional deep embedding schemes showed impressive performance, **they require speaker labels to be trained**:

- End-to-end systems: require speaker labels to define positive and negative pairs for contrastive objective functions

→ Therefore we should utilize pseudo-labels to apply these frameworks to self-supervised scenarios
For **contrastive objectives**, we need to define positive pairs and negative pairs

- In a self-supervised scenario, we can **consider the utterance identity as pseudo-labels**
- For each training utterance, we apply **two different types of augmentations**, resulting in two samples
  - Samples created from the same utterance are considered as positive pair
  - Samples created from different utterances are considered as negative pair
- This way, we can apply contrastive loss functions, such as angular prototypical objective

\[
L_{AP} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\cos(\omega_i^1, \omega_i^2))}{\sum_{j=1}^{N} \exp(\cos(\omega_i^1, \omega_j^2))}
\]
PROBLEM WITH TRAINING WITH PSEUDO-LABELS

- Using pseudo-labels can allow us to train speaker embedding systems with unlabeled dataset, but the **performance is limited as these are not actual speaker labels**
  - Therefore, overfitting on the pseudo-labels can cause critical performance degradation
    - As the system is optimized more to the pseudo-labels, it is likely for the **system to learn non-speaker attributes**
INSTANCE MIXUP (I-MIX)

• I-mix is a data-driven augmentation strategy for improving the generalization of the self-supervised representation

• For arbitrary objective function $L_{\text{pair}}(x, y)$, where $x$ is the input sample and $y$ is the corresponding pseudo-label, giving two data instances $(x_i, y_i)$ and $(x_j, y_j)$,

\[
L^{i-\text{mix}}_{\text{pair}}((x_i, y_i), (x_j, y_j)) = L_{\text{pair}}(\lambda x_i + (1 - \lambda)x_j, \lambda y_i + (1 - \lambda)y_j)
\]

• For cross-entropy-based loss (e.g., prototypical loss), this equation can be rewritten as

\[
L^{i-\text{mix}}_{\text{pair}}((x_i, y_i), (x_j, y_j)) = \lambda L_{\text{pair}}(x_i, y_i) + (1 - \lambda)L_{\text{pair}}(x_j, y_j).
\]

→ Essentially, this creates synthetic training samples with new pseudo-identities
• Here, the mixup coefficient $\lambda \sim \text{Beta}(\alpha, \alpha)$
  • This distribution yields $\lambda$ with value between 0 and 1
  • Depending on the $\alpha$, the distribution shape varies (symmetric)
    • $\alpha < 1.0$: U-shaped distribution, where the sampled $\lambda$ is likely to have value close to either 1.0 or 0.0
    • $\alpha = 1.0$: a uniform distribution across 0 to 1
    • $\alpha > 1.0$: a bell-shaped distribution, where the sampled $\lambda$ is likely to have value close to 0.5
I-MIX ANGULAR PROTOTYPICAL OBJECTIVE (I-AP)

- To enhance the generalization of the self-supervised speaker embedding system, we applied the i-mix strategy to the angular prototypical objective.
  - We apply interpolation on the input acoustic features and utterance identity pseudo-labels.

\[
L_{i-AP} = -\lambda \frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(cos(\omega_{\text{mix}(i,r\neq i)}, \omega_i^2))}{\sum_{j=1}^{N} \exp(cos(\omega_{\text{mix}(i,r\neq i)}, \omega_j^2))} \\
- (1 - \lambda) \frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(cos(\omega_{\text{mix}(i,r\neq i)}, \omega_{\text{mix}(i,r\neq i)}^2))}{\sum_{j=1}^{N} \exp(cos(\omega_{\text{mix}(i,r\neq i)}, \omega_{\text{mix}(i,r\neq i)}^2))}
\]
POSSIBLE LIMITATION OF THE I-AP

• Although applying mixup augmentation to the raw data have proven its strength in many tasks (e.g., speech recognition, image classification), there is room for improvement
  • Due to the linear interpolation, i-mix strategy can only generate new samples between the original samples on the feature space
  • This restricts the diversity of the synthetic training samples, thus limiting the generalization of the system
LATENT SPACE INSTANCE MIXUP (L-MIX)

• In order to overcome this limitation, we propose an i-mix strategy applied to the latent space of speech (l-mix)
  • The latent variable of speech will include essential, disentangled information of various speech attributes
  • We use a variational autoencoder for extracting the latent variable from the given acoustic features (i.e., MFCC)
    • Prior to training the embedding system, we train a VAE for reconstructing the acoustic features

\[
L_{VAE} = D_{KL}(q_{\phi}(z|x)||p(z)) - E_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)],
\]

Encoder parameters

Decoder parameters
Once the VAE is trained, we can use this for extracting the latent variable and reconstructing the acoustic feature

- The VAE encoder generates the Gaussian posterior latent distribution $z \sim N(\mu, \sigma^2)$
- The latent distributions are linearly interpolated, which yields a new Gaussian distribution (weighted sum of independent normal distributions)

$$z_{mix} = \lambda z_1 + (1 - \lambda) z_2$$
$$\sim N(\lambda \mu_1 + (1 - \lambda) \mu_2, \lambda^2 \sigma_1^2 + (1 - \lambda)^2 \sigma_2^2),$$

- The mixed up latent variable is fed into the decoder network to generate a synthetic acoustic feature $x_{l-mix}$
L-MIX ANGULAR PROTOTYPICAL OBJECTIVE (L-AP)

- Analogous to i-AP, we can apply the l-mix strategy to the angular prototypical objective

\[
L_{l-AP} = -\lambda \frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\cos(\omega_{l-mix}^{1}(i,r\neq i), \omega_{i}^{2})\right)}{\sum_{j=1}^{N} \exp\left(\cos(\omega_{l-mix}^{1}(i,r\neq i), \omega_{j}^{2})\right)}
- (1 - \lambda) \frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\cos(\omega_{l-mix}^{1}(i,r\neq i), \omega_{r\neq i}^{2})\right)}{\sum_{j=1}^{N} \exp\left(\cos(\omega_{l-mix}^{1}(i,r\neq i), \omega_{j}^{2})\right)}.
\]
• **VoxCeleb dataset**
  • Training set
    • VoxCeleb2 development set
      • 5994 speakers included (no labels were used for our experiments)
  • Evaluation set
    • VoxCeleb1 trial

• **Acoustic features**
  • 40 dim. MFCC (mel filterbank cepstral coefficients) features
  • Augmentations:
    • Wave-level augmentation: MUSAN noise or RIR simulation
    • Cepstrum-level augmentation: random cepstrum/frame masking (similar to SpecAugment)
• **Embedding system**
  - ECAPA-TDNN architecture: state-of-the-art system for supervised text-independent speaker recognition
  - Attentive channel- and context-dependent statistics pooling
  - Multi-layer aggregation
  - Embedding dimension: 512

• **Variational autoencoder (VAE)**
  - 10 layered convolutional VAE

<table>
<thead>
<tr>
<th>Layer #</th>
<th>Encoder</th>
<th>Decoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3×3 2D-Conv, 32 ReLU, stride 3</td>
<td>64×32 FC</td>
</tr>
<tr>
<td>2</td>
<td>3×3 2D-Conv, 64 ReLU, stride 3</td>
<td>3×3 2D-TransposedConv, 32 ReLU, stride 3</td>
</tr>
<tr>
<td>3</td>
<td>3×3 2D-Conv, 32 ReLU, stride 3</td>
<td>3×3 2D-TransposedConv, 64 ReLU, stride 3</td>
</tr>
<tr>
<td>4</td>
<td>3×3 2D-Conv, 32 ReLU, stride 3</td>
<td>3×3 2D-TransposedConv, 32 ReLU, stride 3</td>
</tr>
<tr>
<td>5</td>
<td>32×64 FC for each $\mu$ and $\log \sigma^2$</td>
<td>3×3 2D-TransposedConv, 1 ReLU, stride 3</td>
</tr>
</tbody>
</table>
**EXPERIMENT**

- **Analysis on synthetic samples**
  - Since i-mix and l-mix applies mixup on different space, *they can create very different samples even when using the same mixup coefficient*.
EXPERIMENT

• **Analysis on synthetic samples**
  - Since the i-mix strategy applies linear interpolation on the feature space, the generated samples are placed on the line between the two original samples.
  - On the other hand, the samples created via l-mix are not necessarily placed on the line.
    - This indicates that the l-mix can create samples with more diversity on the feature space.
EXPERIMENT

• Speaker verification performance
  • Here, we compare performance of the systems trained with various objective functions and augmentations on the VoxCeleb1 evaluation set

→ i-mix and l-mix can both improve the performance when the right coefficient is used

→ The best performance was observed when using l-mix along with wave-level augmentation and ceapsaugin

<table>
<thead>
<tr>
<th>Augmentation</th>
<th>Objective</th>
<th>EER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Benchmark</td>
<td>(Huh et al. 2020)</td>
<td>15.7700</td>
</tr>
<tr>
<td>None</td>
<td>i-vector (Huh et al. 2020)</td>
<td>15.2800</td>
</tr>
<tr>
<td></td>
<td>AP (FastResNet34)</td>
<td>25.3700</td>
</tr>
<tr>
<td>waveaug</td>
<td>GCL (ResNet18)</td>
<td>15.2600</td>
</tr>
<tr>
<td></td>
<td>AP (FastResNet34)</td>
<td>11.6000</td>
</tr>
<tr>
<td>waveaug</td>
<td>i-AP (α = 0.5)</td>
<td>11.6384</td>
</tr>
<tr>
<td></td>
<td>i-AP (α = 1.0)</td>
<td>11.9618</td>
</tr>
<tr>
<td>waveaug</td>
<td>l-AP (α = 0.5)</td>
<td>11.2407</td>
</tr>
<tr>
<td></td>
<td>l-AP (α = 1.0)</td>
<td>11.8240</td>
</tr>
<tr>
<td></td>
<td>l-AP (α = 1.0)</td>
<td>10.7741</td>
</tr>
<tr>
<td>waveaug</td>
<td>l-AP (α = 32.0)</td>
<td>11.8876</td>
</tr>
<tr>
<td></td>
<td>l-AP (α = 0.5)</td>
<td>11.7179</td>
</tr>
<tr>
<td></td>
<td>r-AP (α = 1.0)</td>
<td>10.6257</td>
</tr>
<tr>
<td></td>
<td>l-AP (α = 0.5)</td>
<td>10.9279</td>
</tr>
<tr>
<td></td>
<td>l-AP (α = 1.0)</td>
<td>12.1633</td>
</tr>
<tr>
<td></td>
<td>l-AP (α = 32.0)</td>
<td>10.4931</td>
</tr>
<tr>
<td></td>
<td>l-AP (α = 1.0)</td>
<td>10.5408</td>
</tr>
<tr>
<td></td>
<td>l-AP (α = 32.0)</td>
<td>11.8399</td>
</tr>
</tbody>
</table>
CONCLUSION

• We incorporate the i-mix strategy to the self-supervised speaker embedding learning framework for robust speaker verification

• We also propose a latent space i-mix strategy (l-mix), which performs i-mix on the latent space of the speech

• Our experimental results show that the self-supervised speaker embedding learning can benefit greatly from the i-mix regularization strategy

• Moreover, the proposed l-mix strategy can further improve the performance, by yielding much diverse synthetic training samples
Q & A