The Hong Kong Polytechnic University
香港理工大學

Robust Speaker Verification Using Population-based Data Augmentation

Weiwei LIN, Man-Wai MAK
Department of Electronic and Information Engineering, The Hong Kong Polytechnic University

Introduction

- Data Augmentation is an important procedure in the training of speaker embedding networks.
- However, we often use a set of pre-defined DA parameters whose values were intuitively set instead of optimally determined.
- In this paper, we propose to use population-based learning to automatically learn DA parameters.

Methodology

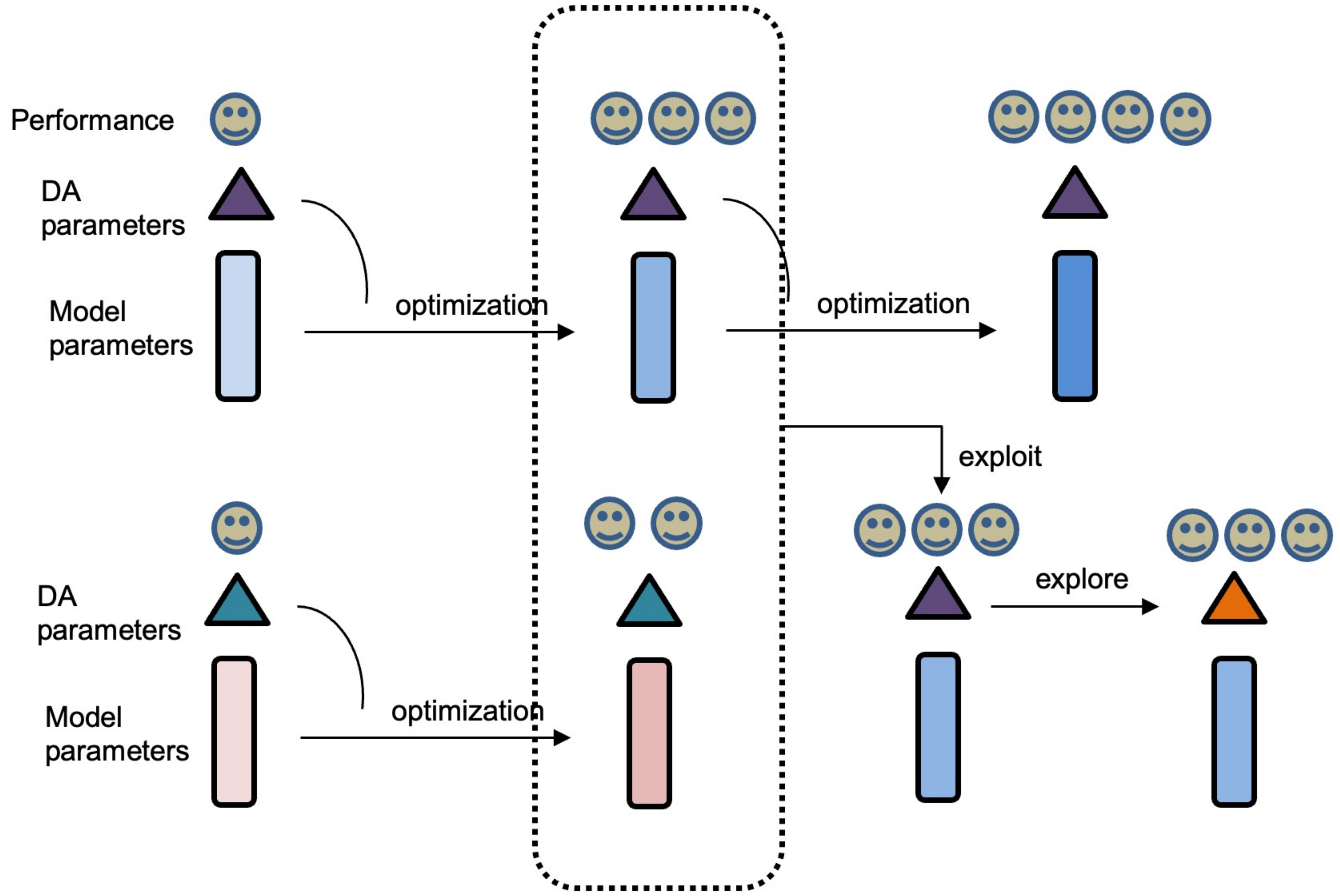
PBA learns a schedule for changing the DA parameters. It involves the following steps:

- Initialization: The augmentation parameters for each model in a population are randomly initialized with some predefined ranges.
- Optimization: The network parameters of each model are optimized independently (using stochastic gradient descent (SGD) on the augmented data.)
- Evaluation: Each of the models in the population is evaluated on a validation set.
- Exploitation: The network parameters of the models in the bottom 25% of the ranked list are replaced by those in the top 25%.
- Exploration: Apply the “explore” function to the augmentation parameters.

References

1. Ho, Daniel, et al. "Population based augmentation: Efficient learning of augmentation policy schedules." *International Conference on Machine Learning*. PMLR, 2019.
2. Snyder, David, et al. "X-vectors: Robust dnn embeddings for speaker recognition." *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018.

PBA Explore and Exploit Functions



- “Exploit” selects the top-ranked models and DA params.
- “Explore” randomly perturbs the DA params.

Training procedures

- Voxceleb1 dev and Voxceleb2 dev set.
- VOiCES-19 dev was used as the validation set.
- Adding noise, babble, and music from MUSAN and reverberation from the RIR dataset to speech. Time and frequency masking is also applied

Input acoustic features

- 40-dimensional filter bank features with a frame length of 25ms at 10ms shift;
- using Kaldi energy-based voice activity detection (VAD) to remove silence frames;
- using small chunks of acoustic sequences with a chunk length of 400 frames for training.

Algorithm 1 Applying data augmentation to a mini-batch

```
Input: mini-batch 𝒳, parameters ℋ   ▷
ℋ is a list of augmentation hyperparameters comprising (trans, prob, mag)
ℒ = [ ]                                  ▷ Empty List
for x in 𝒳 do
    x = sample_segment(x, T)             ▷ Sample a random segment with duration T
    for (trans, prob, mag) in ℋ do
        if random(0, 1) < prob then
            z = trans(x, mag)
            ℒ = append(ℒ, z)
        else
            ℒ =append(ℒ, x)
        end if
    end for
end for
Return ℒ
```

Algorithm 2 PBA “explore” function for magnitude parameters. Magnitude parameters can be any from 0 to 9 inclusive.

```
Input: MagParams ℳ            ▷ ℳ is a list of magnitude parameters
ℳ_new = [ ]     ▷ Initialize an empty list for new parameters
for m in ℳ do
    if random(0, 1) < 0.2 then
        m_new = random_int(0, 9)          ▷ resample a new parameter
    else
        inc = random_int(0, 3)            ▷ Randomly choose an increment value
        if random(0, 1) < 0.5 then
            m_new = m + inc   ▷ Increase the aug. parameter
        else
            m_new = m − inc   ▷ Decrease the aug. parameter
        end if
    end if
    m_new = max{0, m_new}        ▷ Clip m_new within [0, 9]
    m_new = min{m_new, 9}        ▷ Clip m_new within [0, 9]
    ℳ_new = append(ℳ_new, m_new)
end for
Return ℳ_new
```

Comparison with Kaldi Aug.

Network	Aug	EER
X-vector	Kaldi+SpecAug	6.87%
X-vector	PBA	4.82%
DenseNet121	Kaldi+SpecAug	5.53%
DenseNet121	PBA	3.98%

Ablation Study

Without	EER
None (using all aug.)	3.98%
Additive noise	4.42%
Reverb	4.63%
Time masking	4.03%
Freq masking	3.88%

Discussions

- Deeper networks, such as DenseNet121, achieve much better performance than the X-vector networks.
- For both X-vector and DenseNet121, PBA obtains better performance than Kaldi augmentation.
- Ablation study shows that reverberation is the most important augmentation and frequency mask is the least important augmentation. Removing frequency masking actually improves the performance. This could be that it does not work well with other augmentation operations.