# Unsupervised learning - Project

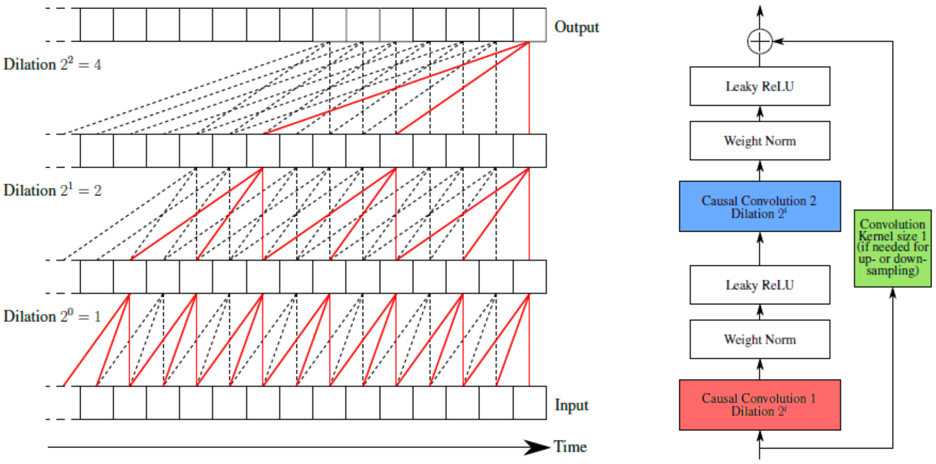
presenter: Mordehay Moradi

This project is relied on the “Unsupervised Scalable Representation Learning for Multivariate Time Series” paper.

This paper investigates the topic of unsupervised general-purpose representation learning for time series. A few articles explicitly deal with general-purpose representation learning for time series without structural assumption on non-temporal data.
This problem is interesting because real-life time series are rarely or sparsely labeled and hence unsupervised representation learning would be strongly preferred.
The paper’s authors propose an unsupervised method to learn general-purpose representations for multivariate time series that comply with the issues of varying and potentially high lengths of the studied time series.

There are some works done to tackle this problem.Lei et al. (2017)[1] expose an unsupervised method designed so that the distances between learned representations mimic a standard distance (Dynamic Time Warping, DTW) between time series, Malhotra et al. (2017)[2] design an encoder as a recurrent neural network, jointly trained with a decoder as a sequence-to-sequence model to reconstruct the input time series from its learned representation. However, these methods either are not scalable nor suited to long time series, due to the sequential nature of a recurrent network, or to the use of DTW with a quadratic complexity with respect to the input length, are tested on no or very few standard datasets and with no publicly available code, or do not provide sufficient comparison to assess the quality of the learned representations.
Scalable models and extensive analysis are suggested in the paper to overcome these problems while also outperforming these approaches. Another advantage of this paper is that the authors demonstrate the quality, transferability and practicability of the learned representations with thorough experiments and comparisons

The paper suggests encoder-only architecture for avoiding the need to jointly train with a decoder as in autoencoder-based standard representation learning methods as done by Malhotra et al. (2017)[2], those would induce a larger computational cost.
In more details, the proposed architecture is number of stacked dilated causal convolution



(a)

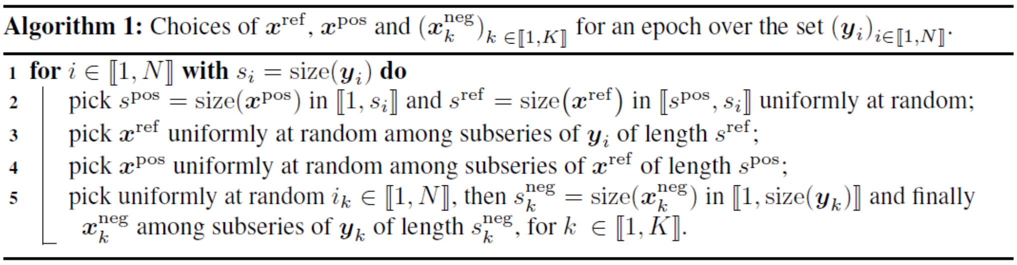
(b)

In the figure above we can see: (a) Illustration of three stacked dilated causal convolutions. Lines between each sequence represent their computational graph. Red solid lines highlight the dependency graph for the computation of the last value of the output sequence, showing that no future value of the input time series is used to compute it. (b) Composition of the $i$-th layer of the chosen architecture.The output of this causal network(which has a size of (B, C, L), where B, C and L is the batch size, number of channels and the number of frames correspondly ) is then given to a global max pooling layer squeezing the temporal dimension and aggregating all temporal information in a fixed-size vector (as proposed by Wang et al. (2017) in a supervised setting with full convolutions)(which has a size of (B, $\tilde{C}$)). A linear transformation of this vector is then the output of the encoder, with a fixed, independent from the input length, size(which has size of (B, fixed_size), in this work the fixed size is equal to 2 for visualization purposes).
It is worth mentioning the big advantage of causal convolution layers since we can parallelize processes, the model can be trained in a faster way than conventional convolutions layers.

The triplet loss through which the model is trained, was inspired by the classic word representation learning method known as word2vec. This work is the first in the time series literature to rely on a triplet loss in **a fully unsupervised setting**. The triplet loss is formulated as:

$$L(\theta) = -\log(\sigma(\mathbf{f}(\mathbf{x}^{ref},\theta)^T\mathbf{f}(\mathbf{x}^{pos},\theta)) - \sum_{k=1}^{K}\log(\sigma(\mathbf{f}(\mathbf{x}^{ref},\theta)^T\mathbf{f}(\mathbf{x}_k^{neg},\theta))$$

where $\sigma$ is the sigmoid function, $f(\cdot,\theta)$ is a deep neural network.
This loss pushes the computed representations to distinguish between $\mathbf{x}^{ref}$ and $\mathbf{x}^{neg}$ , and to assimilate $\mathbf{x}^{ref}$ and $\mathbf{x}^{pos}$. Overall, the training procedure consists in traveling through the training dataset for several epochs, picking tuples $(\mathbf{x}_{ref}, \mathbf{x}_{pos}, (\mathbf{x}_k^{neg})_k)$ at random as detailed in Algorithm 1(The length of the negative examples is chosen at random in Algorithm 1), and performing a minimization step on the corresponding loss for each pair, until training ends.



**Algorithm 1:** Choices of $\boldsymbol{x}^{\text{ref}}$, $\boldsymbol{x}^{\text{pos}}$ and $(\boldsymbol{x}_k^{\text{neg}})_{k\,\in[\![1,K]\!]}$ for an epoch over the set $(\boldsymbol{y}_i)_{i\in[\![1,N]\!]}$.

```
for $i \in [\![1, N]\!]$ with $s_i = \text{size}(\boldsymbol{y}_i)$ do
    pick $s^{\text{pos}} = \text{size}(\boldsymbol{x}^{\text{pos}})$ in $[\![1, s_i]\!]$ and $s^{\text{ref}} = \text{size}(\boldsymbol{x}^{\text{ref}})$ in $[\![s^{\text{pos}}, s_i]\!]$ uniformly at random;
    pick $\boldsymbol{x}^{\text{ref}}$ uniformly at random among subseries of $\boldsymbol{y}_i$ of length $s^{\text{ref}}$;
    pick $\boldsymbol{x}^{\text{pos}}$ uniformly at random among subseries of $\boldsymbol{x}^{\text{ref}}$ of length $s^{\text{pos}}$;
    pick uniformly at random $i_k \in [\![1, N]\!]$, then $s_k^{\text{neg}} = \text{size}(\boldsymbol{x}_k^{\text{neg}})$ in $[\![1, \text{size}(\boldsymbol{y}_k)]\!]$ and finally
      $\boldsymbol{x}_k^{\text{neg}}$ among subseries of $\boldsymbol{y}_k$ of length $s_k^{\text{neg}}$, for $k \in [\![1, K]\!]$.
```

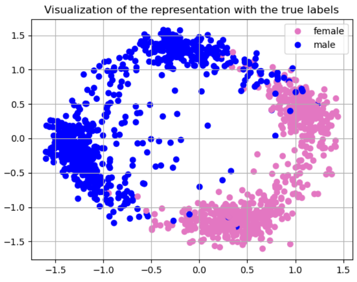
## Experiment

In this project I was examining the performance of the representation vectors yielding from the model that has been described in the previous section, for auditory data. The purpose is to divide speakers by their gender, male and female in a **fully** unsupervised way .
Dataset that contains male and female recordings is for example the LibriSpeech dataset[3]. The LibriSpeech corpus is a collection of approximately 1,000 hours of audiobooks such that each recording in this dataset is spoken by one man or woman. Since there are one-speaker recordings for each gender, we can build negative and positive samples. After choosing the recordings(negative and positive), I computed the corresponding log mel-spectrogram for each recording and build tuple $(\mathbf{x}_{ref}, \mathbf{x}_{pos}, (\mathbf{x}_k^{neg})_k)$ that will be inputted to the model. Mel spectrogram is used here because is good for classification task.The building of the tuple from the log mel - spectrograms will be as follows: taking from the positive mel - spectrogram samples, $k_{pos}$ frames labeled as $\mathbf{x}_{pos}$ and $k_{ref}$ frames labeled as $\mathbf{x}_{ref}$. On the same way, for $K$ different recordings of negative samples. In practice, we only know the speaker's id of the positive recording such that the negative recordings are sampled from the rest of the data that does not belong to this positive speaker.

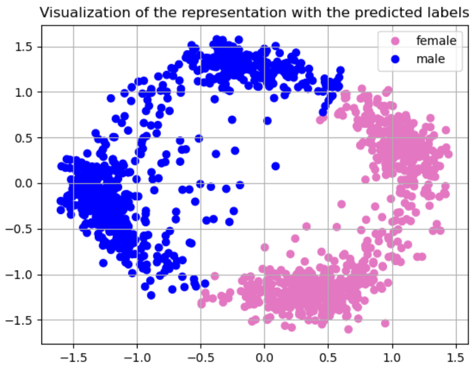
Finally the input to the model has the following shape: $\mathbf{x}_{ref} \in \mathbb{R}^{F\times k_{ref}}$, $\mathbf{x}_{pos} \in \mathbb{R}^{F\times k_{pos}}$, $(\mathbf{x}_{neg} \in \mathbb{R}^{F\times k_{negk}})_k$ where $F$ is the number of frequencies. As can be seen, there is difference between these three variables along the temporal axis, but this doesn't matter since the layer before the last is global max pooling layer which squeeze this dimension
For visualization purposes, the size of the latent space will be 2d or 3d. I will expect to see samples belonging to the male recordings separated from the samples belonging to the female recordings.
I run the model with the following hyperparameters:
batch size = 64, number of the middle channel = 40, input channel = number of mel band = 128, kerne size = 3, Adam optimizer with lr = 0.001, number of negative samples = 10, negative penalty = 1, number of output channels = 2.
I trained and tested the model on recordings of 5 seconds sampled at 16 kHz.
For more comprehensive analysis, I tested the model also on the Wall street Journal(WSJ) corpus[4]. For our purpose this dataset has the same properties as the LibrSpeech dataset[3].

## Results

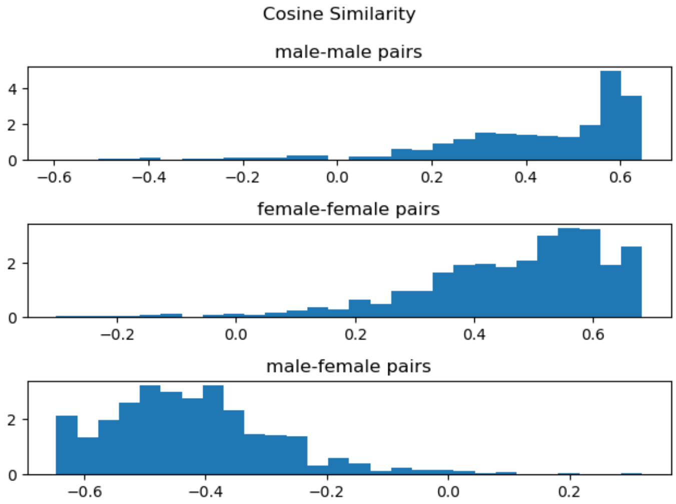
After training the model on the train set I was examining the performance for unseen samples from the LibriSpeech test set. The test set includes 40 speakers with balance in the gender of the speakers(20 male and 20 female).
Based on the following figure, it can be seen that the vectors represent differences between speakers based on their gender



In addition to the representation in 2d I fitted SVM classifier and got  almost the same results:



For evaluating the performance I computed some metrics of clustering denoting the similarity of two different clustering(the true one and the predicted classes from the classifier).
The first metric is the cosine similarity, which was used also to train the model. between all within-class pairs ((male, male), (female, female)) and all not within-class pairs (male, female).



As can be seen, speakers with the same gender have more similarity than speakers with different gender. This result denotes the success of separating speakers by their gender.
One can also compute the mean of these paris and get:
cosine similarity within class -  higher is better
cosine similarity not within class  - lower is better
In addition to the cosine similarity I compute more other metrics and get the following results:

- Completeness: all members of a given class are assigned to the same cluster - higher is better
- Homogeneity: each cluster contains only members of a single class - higher is better
- Mutual Information:  a function that measures the agreement of the two assignments, ignoring permutations. Values close to zero indicate two label assignments that are largely independent, while values close to one indicate significant agreement.
- Accuracy

| Metric Name | Value - LibriSpeech | Value - WSJ |
|---|---|---|
| Accuracy | 0.97 | 0.96 |
| Homogeneity | 0.82 | 0.77 |
| Completeness | 0.82 | 0.77 |
| Normalized Mutual Information | 0.82 | 0.77 |
| Cosine similarity within class | 0.44 | 0.58 |
| Cosine similarity not within class | -0.42 | -0.41 |

Moreover, as I mentioned above I test the quality of the representation vectors on the wsj corpus. I built a test set of 5545 female recordings and 4888 male recordings. The results can be seen in the table above.

In conclusion, I trained the proposed model with multivariate auditory data(the frequency dimension in the mel spectrogram) in order to extract a 2d vector that will be used to svm classifier to classify the speakers by their gender. Overall, I got impressive results considering the fact that the training was done in an unsupervised way. In my opinion, the model misses when some gender distorts his voice and makes him sound like the other gender. In this project I was impressed by the power of discovering underlying insights about the data without any labels and external information.
During this project, I worked through the issue of low accuracy (0.82) that resulted from using a lot of unnecessary parameters, which led me to reduce the number of channels and layers.

References

[1] Lei, Q., Yi, J., Vaculin, R., Wu, L., and Dhillon, I. S. Similarity preserving representation learning for time series analysis. arXiv preprint arXiv:1702.03584, 2017.
[2] Malhotra, P., TV, V., Vig, L., Agarwal, P., and Shroff, G. TimeNet: Pre-trained deep recurrent neural network for time series classification. arXiv preprint arXiv:1706.08838, 2017.
[3] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.
[4] J. Garofolo, D. Graff, D. Paul, and D. Pallett, “Csr-i (wsj0) complete ldc93s6a,” Web Download. Philadelphia: Linguistic Data Consortium, vol. 83, 1993.