Assignment 4. Manipulating data to 'fool' Artificial Neural Network

Submission deadline: Monday, 24 April 2023, 12:00

Submission format:

- link to a cloud storage with full project and files you added (original dataset, generated dataset, training file, testing file, spectrograms generation files, audio manipulation file)

Grading criteria:

- Quality of exploratory data analysis: +2 points,
- Justification of your 'attack' strategy: +2 points,
- The target metric has dropped by at least 0.05: +2 points,
- The target metric has dropped by at least 0.10: +2 points,
- Signal generation applied (Assignment 1): + 2 points,
- Noises applied (Assignment 2): +2 points,
- Live grading: +8 points;

Task

Imagine, if you get access to some smart speaker. Your task here is to 'confuse' or 'fool' Artificial Neural Network, which classifies audio data from spectrograms. In order not to expose yourself prematurely, you can modify only 15% of packages (in our case timesteps) in each recording. Try to manipulate test data and get lowest evaluation metrics for the Artificial Neural Network I prepared for you.

Directories 'test_audio' and 'train_audio' contain .wav files, spectrograms were created from them. ANN were trained on these spectrograms (directories 'train_spec' and 'test_spec'). Python notebook 'to_spectrograms.ipynb' is to be used for spectrograms generation. Python notebook test.ipynb' is to be used for evaluation of your attack. Directory 'my_model' contains Tensorflow model files.

What I expect you to do:

- Download project from cloud storage (link),
- Use all techniques and tricks you learned during previous assignments and run exploratory data analysis of training and test sets (no limits, no constraints),
- Explain your findings. Develop a strategy to 'fool' ANN. Justify it.
- Use .wav files from 'test_audio' and manipulate them to generate new spectrograms for the test set. You can modify up to 15% of timesteps for each file. Use new spectrograms as a test set and evaluate classifiers. Try to get the lowest score possible. Evaluation metric is 'accuracy' (0.6906).

Constraints:

- Your test set shape has to match the shape of the original test set.
- You CAN'T change more than 15% of timesteps in each .wav file (if you change more, it means zero).
- Generate spectrograms using 'to_spectrograms.ipynb' only.
- Additional lab points for 10 best-performing solutions.

References:

- Adversarial Attacks on Convolutional Neural Networks in Facial Recognition Domain: https://arxiv.org/abs/2001.11137
- Data Augmentation techniques in time series domain: A survey and taxonomy: https://arxiv.org/abs/2206.13508
- Convolutional Neural Network: Feature Map and Filter Visualization: https://towardsdatascience.com/convolutional-neural-network-feature-map-and-filter-visualization-f75012a5a49c
- Inspiration for this assignment: https://ods.ai/competitions/data-fusion2023-attack
- Original dataset: https://www.tensorflow.org/datasets/catalog/speech_commands