Synopsis for 02456 Project "Various Deep Learning Architectures for Urban Sound Classification"

s161041	Sébastien Demortain	s161027	Péter Semság
s161174	Lorenzo Belgrano	s161463	Benjamin Jüttner

Background and Motivation:

Sound classification is a task commonly solved by RNNs rather than CNNs, which in turn are rather suitable for image data. However, since a spectrogram of an audio sequence can be interpreted as an image, CNNs too can be used for sound data, as was done e.g. in [1]. Therefore, sound data is a good opportunity to work with two of the architectures we learned in 02456, namely CNNs and RNNs. The dataset chosen for the project was the *UrbanSound8K*[2], which is a collection of over 8000, up to 4 seconds long, audios from urban environments with labels such as dog barking or jackhammer.

Milestones:

1. (also safe plan B) Reproduce the CNN architecture proposed in [1], with each audioclip processed into several 60×41 pixel spectrograms.
2. Same architecture and same data as in Milestone 1, but now train easier observations first and more difficult observations afterwards (see *curriculum learning* [3]). See if the performance improves.
3. Implement an architecture combined of CNN and RNN (maybe realizing some of the ideas in [4]).
4. Experiment with mixed data: (i) Artificially overlap two audios of classes a and b and make it one observation. The network should give a softmax output where the highest values are in a and b. (ii) Concatenate audios, e.g. dog – jackhammer – silence – jackhammer. Use CTC[5, 6] to automatically segment the new audio and give a label for each segment.

References

[1] K. J. Piczak: "ENVIRONMENTAL SOUND CLASSIFICATION WITH CONVOLUTIONAL NEURAL NETWORKS", in *2015 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING*, Sept. 17–20, 2015, Boston, USA

[2] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research", in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[3] Y. Bengio, J. Louradour, R. Collobert, J. Weston: "Curriculum Learning"

[4] Baidu Research – Silicon Valley AI Lab: "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin"

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber: "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks." In *ICML*, pages 369–376. ACM, 2006.

[6] TensorFlow implementation of [5]: `https://www.tensorflow.org/versions/r0.12/api_docs/python/nn/connectionist_temporal_classification__ctc_`