Plain PCA layers on patches of CQT

Carmine-Emanuele Cella

November 16, 2016

1 Dataset

The following experiments has been done on the *ESC-50* dataset, made of 50 classes with 40 samples per class (totalling 2000 samples) of environmental sounds. Each sample is 5 seconds of sound sampled at 44.1 Khz (220500 samples).

The state of the art results for this dataset are (accuracy): 73% by myself (convnet on top of CQT), 73.5% by Joakim Anden (joint-scattering) and 74.2% by the paper [SoundNet] (semisupervised network).

2 Algorithm

The algorighm applied in this context is inspired by the paper [PCANet] and is as follows:

$$x \to |CQT_{84\times 215}| \to patches_{9\times 9} \to |PCA_{50}| \to patches_{9x9} \to |PCA_{25}| \to log.$$

At each PCA layer, the number of components is reduced to almost half and to almost a quarter respectively. The experiments have been done on the whole dataset and on subsets made of 5, 10, 20 and 25 classes respectively. I computed the plain CQT (modulus) on each set and then the version with 2 PCAs described above; the data have been transformed by *log* but not standardized. The classifier used is a linear SVM.

3 Results

The table 3 shows the obtained results. For each set, the increment given by the PCAs is substantial and increases when the size of the dataset increases. Figure 3 shows the filters learned at layer of the PCA; they look substantially DCTs bases.

Figure 1: PCA components for the first layer.

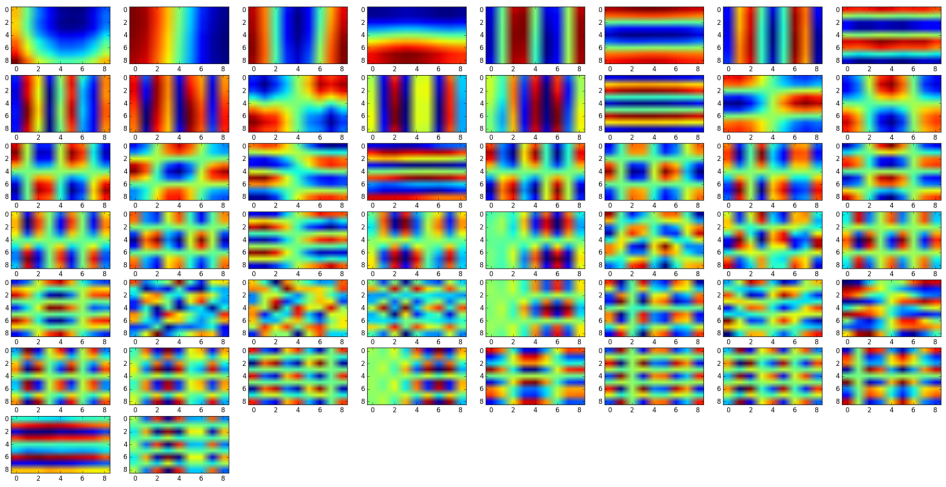

Table 1: Results on different subsets of ESC-50.

no. classes	**method**	**accuracy**	**improvement over CQT**
5	CQT	69%	-
5	CQT + 2 PCA	80%	11%
10	CQT	56%	-
10	CQT + 2 PCA	71%	15%
20	CQT	50%	-
20	CQT + 2 PCA	65%	15%
25	CQT	38%	-
25	CQT + 2 PCA	62%	24%
50	CQT	33%	-
50	CQT + 2 PCA	53%	20%

References

[SoundNet] Y. Aytar et al., SoundNet: Learning Sound Representations from Unlabeled Video, NIPS 2016, Barcellona.

[PCANet] T. Chan et al., PCANet: A Simple Deep Learning Baseline for Image Classification?, 2014, arXiv.