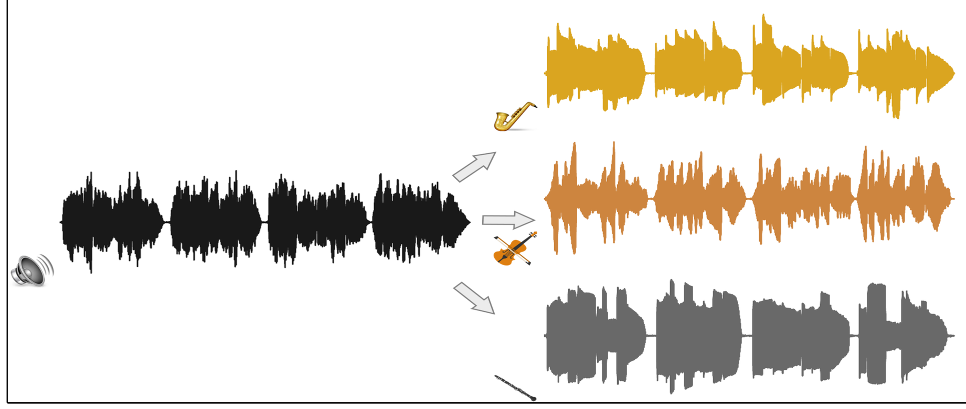
Single Channel Source Separation


[1]James Webb, [2]Joshua Chu, & [2]Shayok Dutta

[1]Department of Neuroscience, Baylor College of Medicine, [2]Department of Electrical and Computer Engineering, Rice University, Houston, Texas

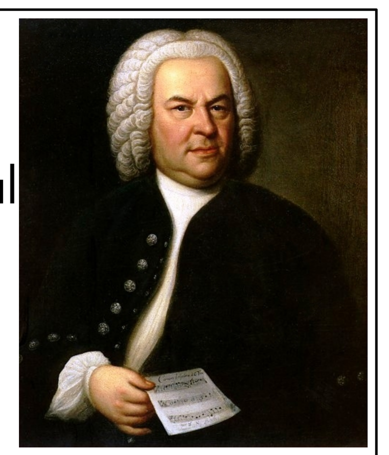

Introduction

Single channel source separation (SCSS) has important applications (e.g. speech recognition). Approaches taken have been NNMF, denoising autœncoders, and CNNs. Ability of CNNs to interpret local connectivity is appealing. Recently, [1] proposed a variant of an autœncoder that uses convolutional layers to process spatiotemporal information that achieved comparable results to previous architectures. We further investigated the performance of the convolutional autœncoder on the Bach10 dataset.

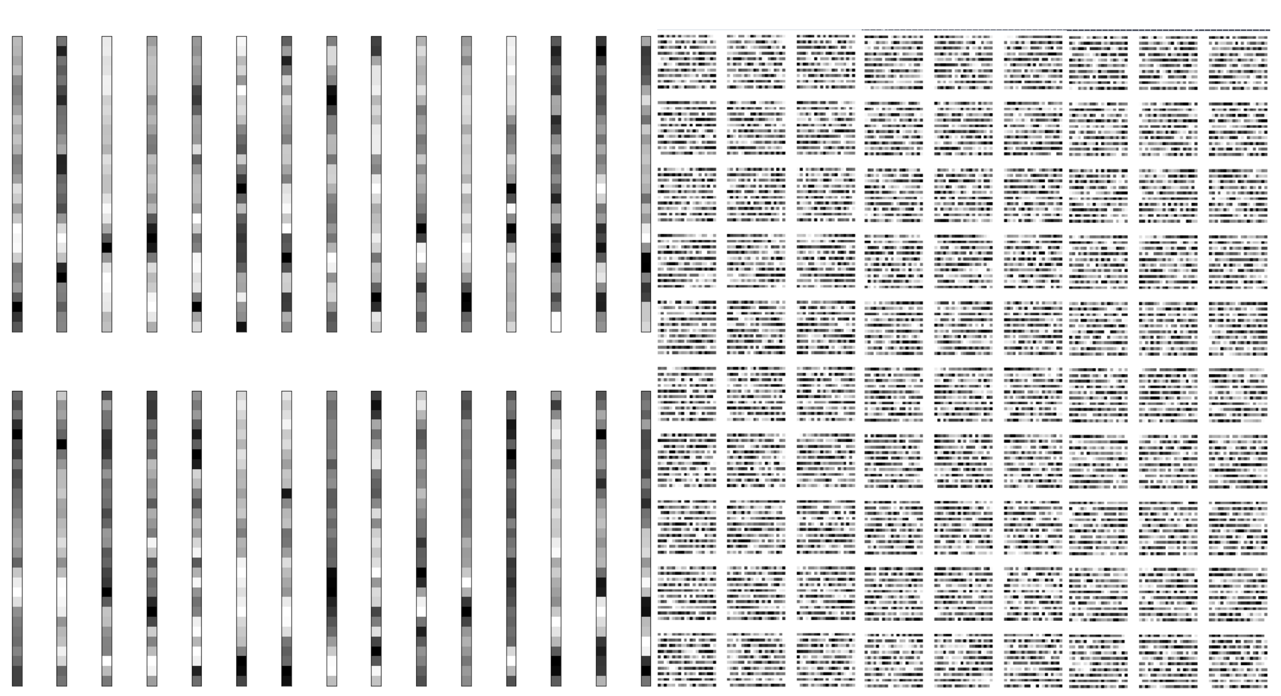


Dataset

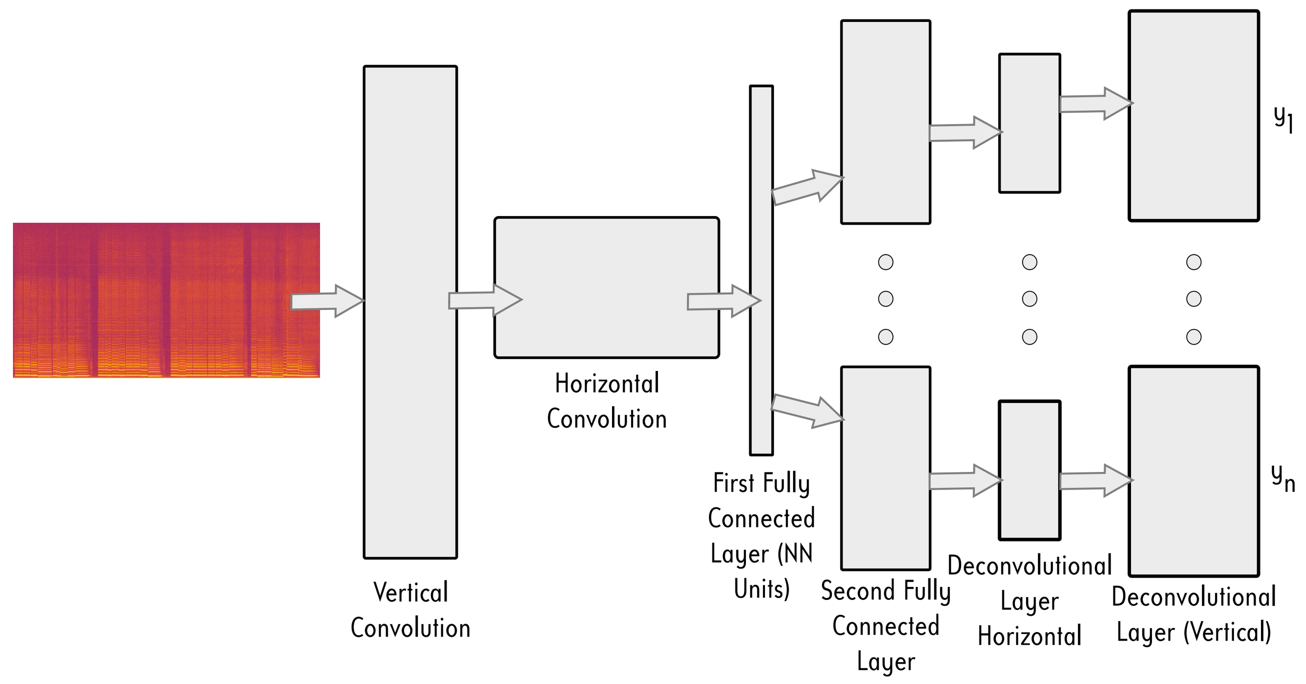
Bach10 Dataset
- 10 songs
- Contains full song and individual sources
- Individual sources include:
  (1) Bassoon
  (2) Clarinet
  (3) Saxphone
  (4) Violin



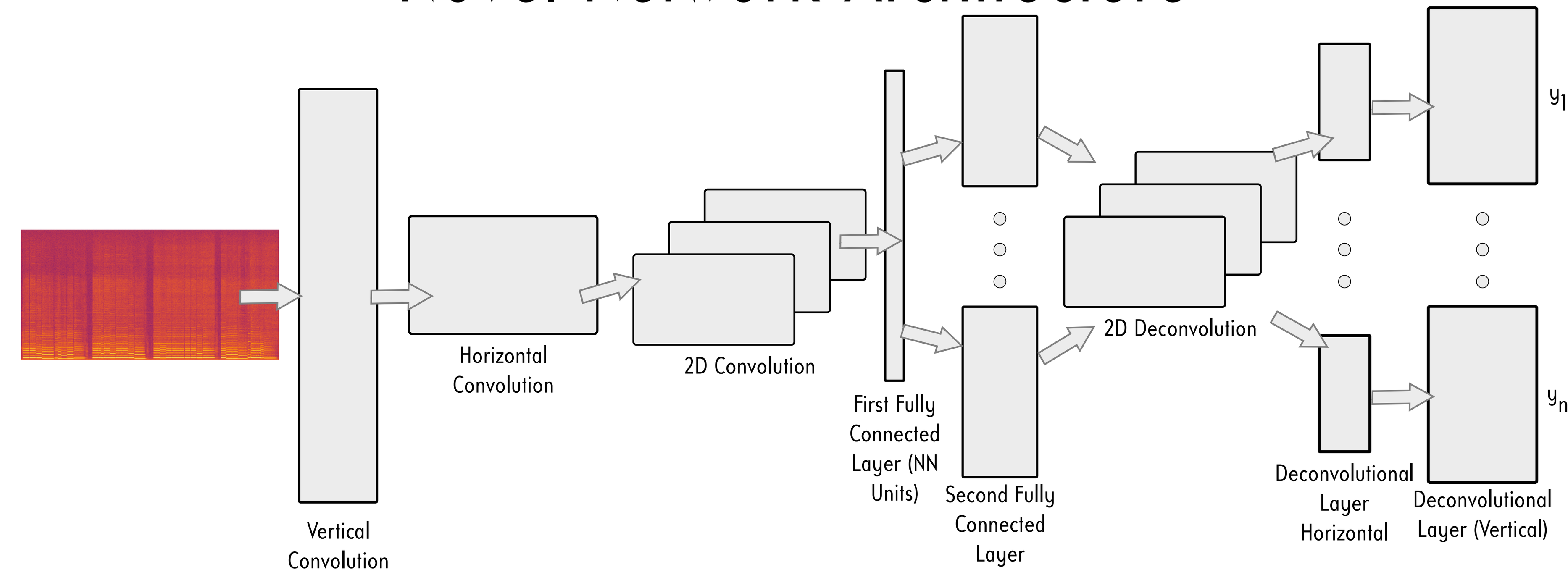
Filters



General Network Archtecture



Novel Network Architecture

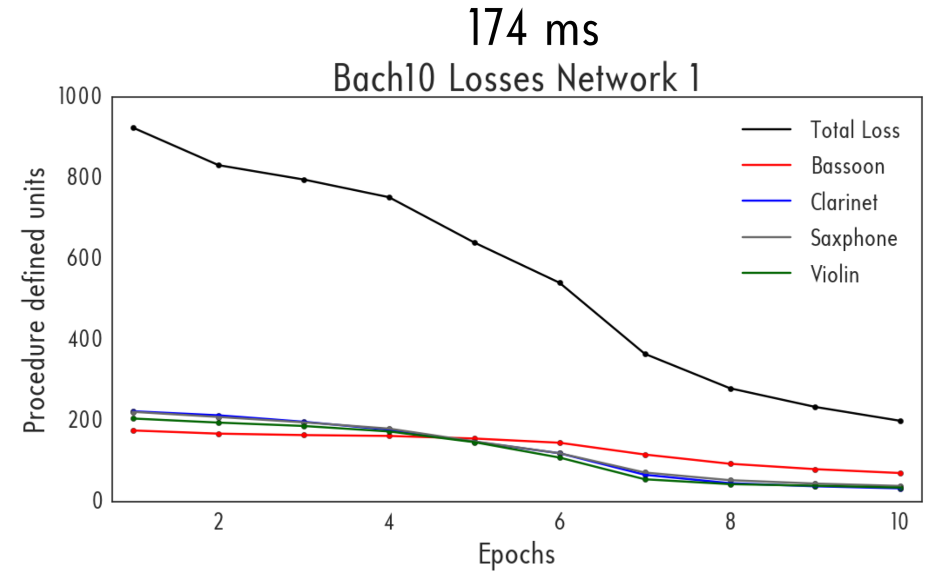


Network Parameters (unless noted otherwise):
- Max # epochs: 10-20
- Training songs: First 5
- Initial learning rate: 0.001
- Input feature size: 1025

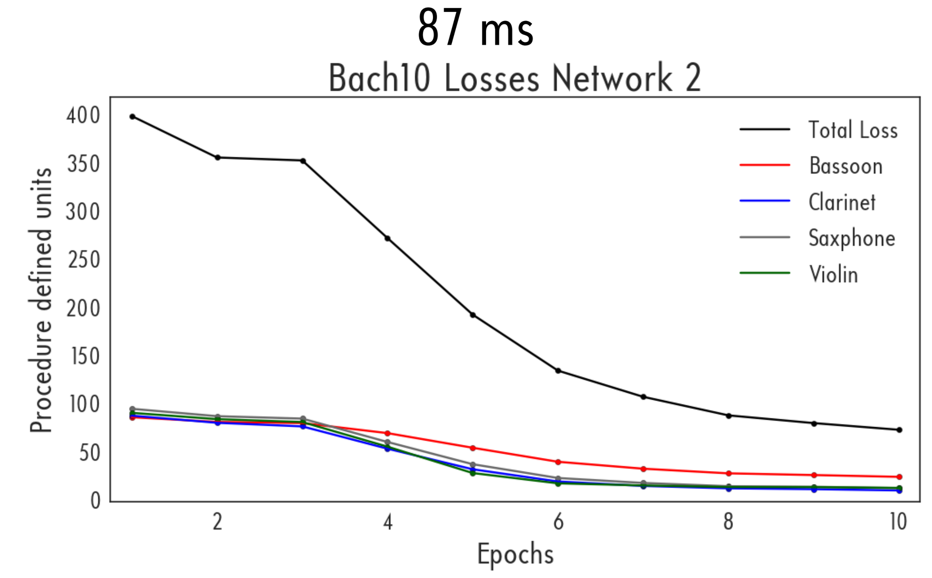
$$m_n(f) = \frac{|\hat{y}_n(f)|}{\sum_{n=1}^{N} |\hat{y}_n(f)|} \qquad \tilde{y}_n(f) = m_n(f)x(f) \qquad L_{sq} = \sum_{i=1}^{N} \|\tilde{y}_n - y_n\|^2$$

Varying Spectogram Time Contexts

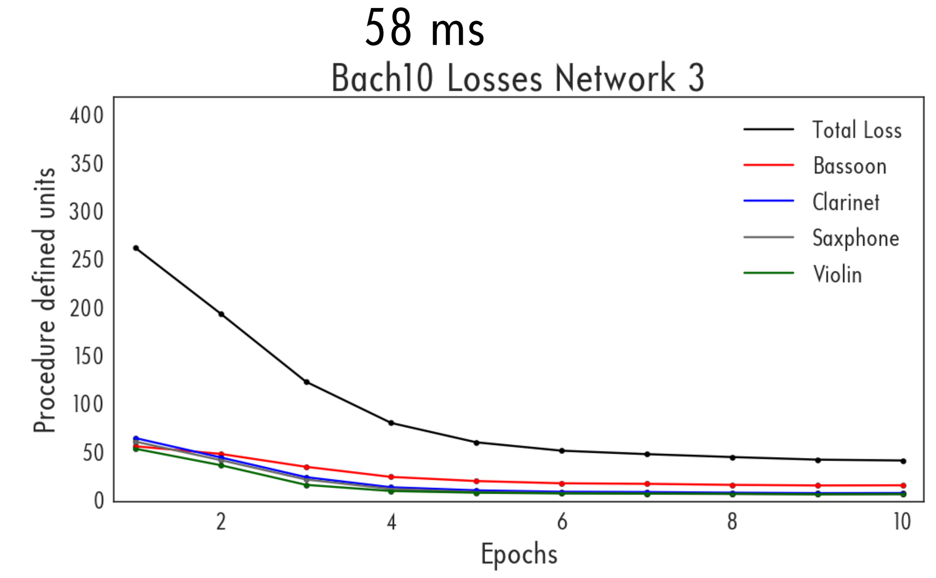
Network 1
174 ms



Network 2
87 ms

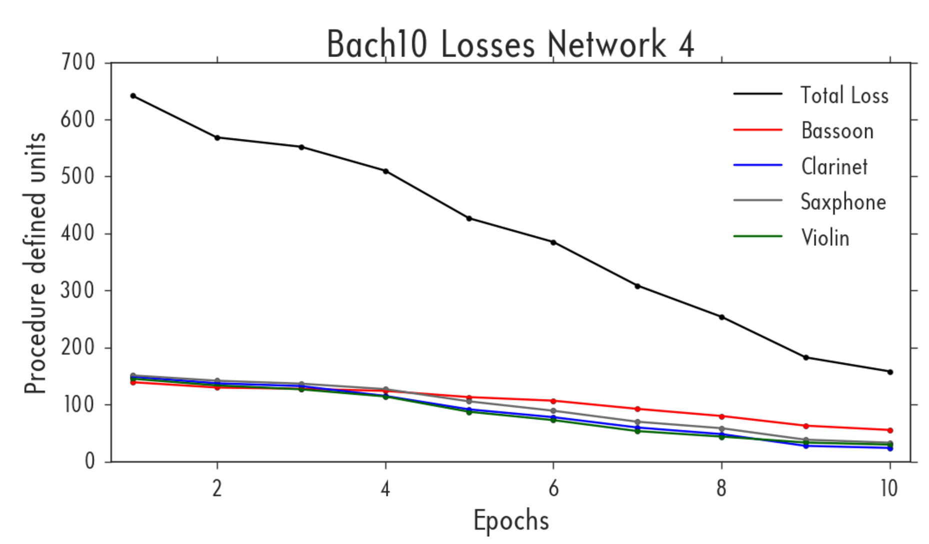


Network 3
58 ms

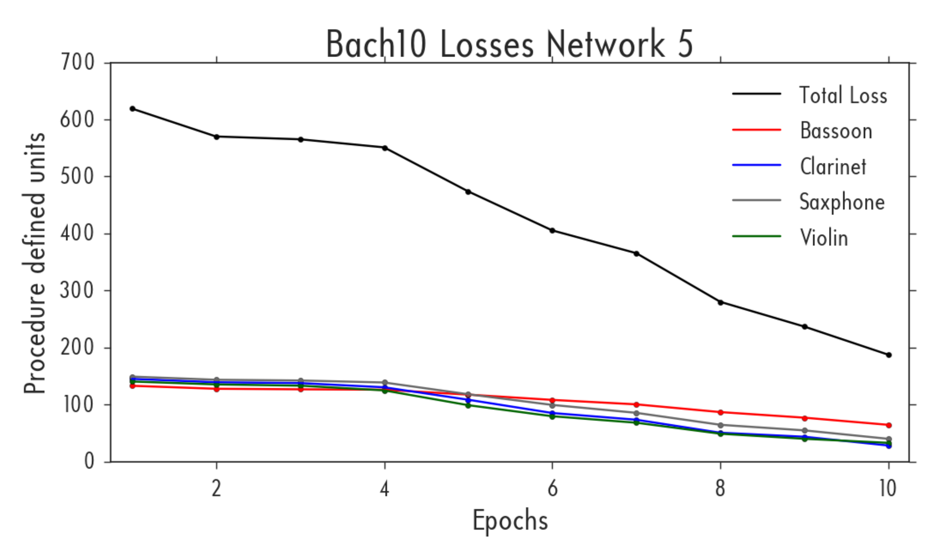


Varying Fully Connected Layer Units

Network 4 Modification
512 units fully connected layer

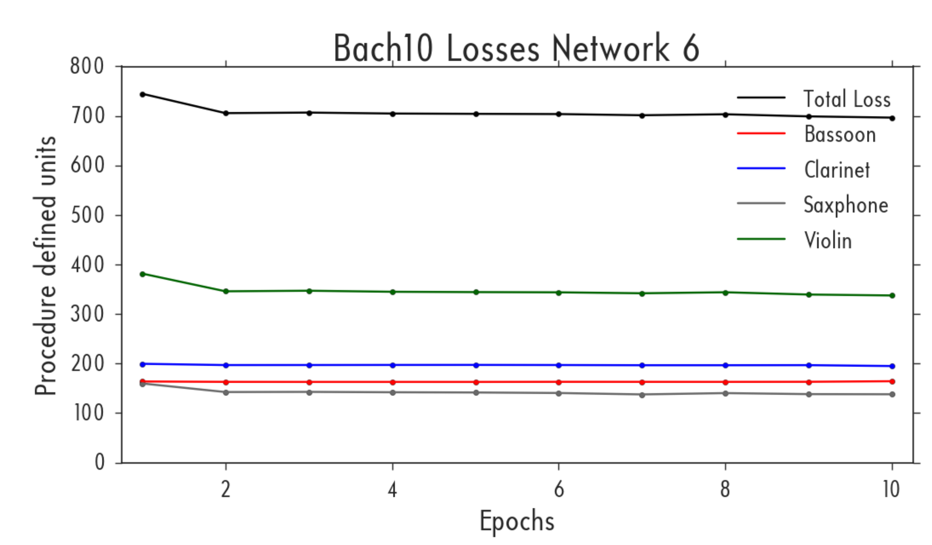


Network 5 Modification
128 units fully connected layer



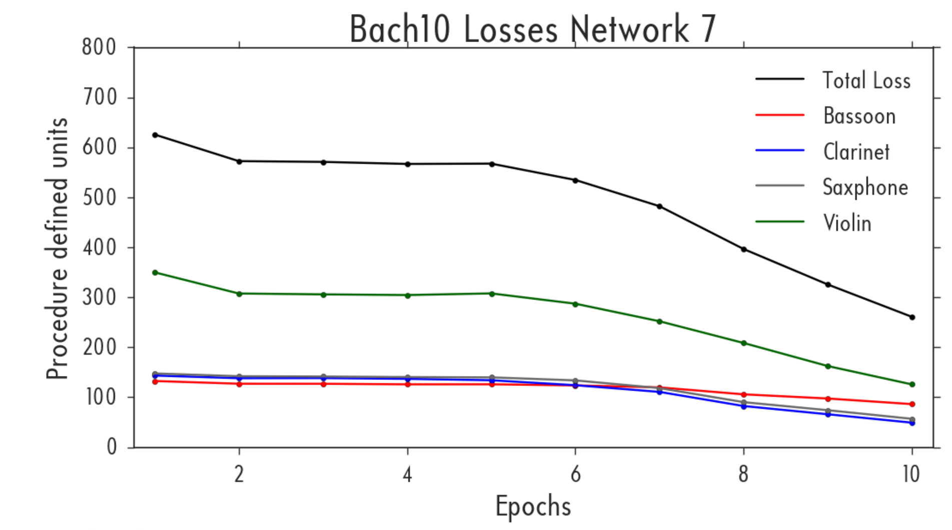
Novel Architecture

Network 6 Modification
Third convolutional/deconvolutional layer



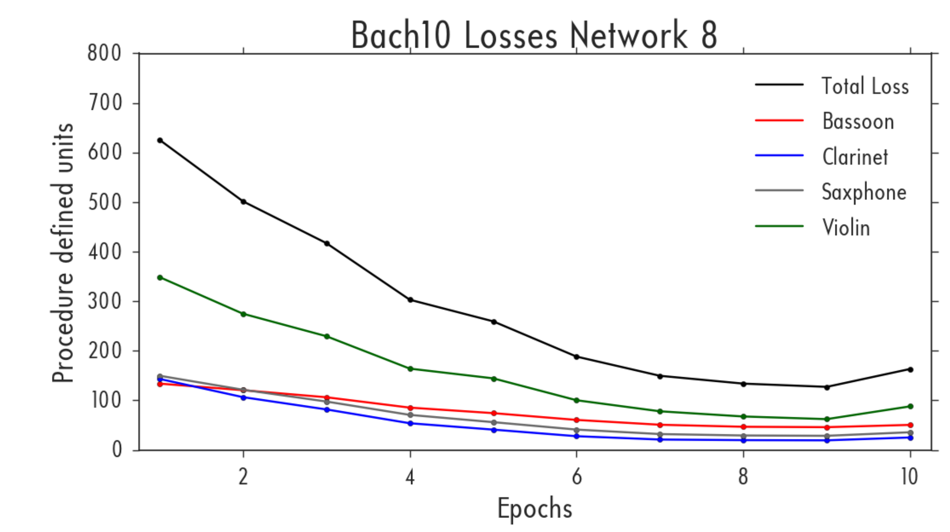
Modifying Conv Layer Filters

Network 7 Modification
64 convolution filters per layer



Adding Activation Function

Network 8 Modification
ReLU transfer function



Conclusions & Future Works

All network variants learned the training set, as evidenced by decreasing, stabilizing learning curves. Surprisingly, network performance varied inversely with the time context of the spectograms, suggesting time-dependent features only exist on a small scale in the dataset. Additionally, performance with respect to individual sources depended on the network architecture. Generally, the network learned the source signals of the violin and clarinet best, and of the bassoon the worst. However, when the convolutional processing was increased, the performance with respect to the bassooon and violin flipped.

Further investigation should extend these results to larger datasets, such as the DSD100, to test the scalability of our findings. The features of each source could be better understood by investigating gradients with respect to input.

References

[1] Chandna, Pritish. "Monoaural Audio Source Separation Using Deep Convolutional Neural Networks," 2017.
[2] Grais, Emad M., and Mark D. Plumbley. "Single Channel Audio Source Separation Using Convolutional Denoising Autœncoders." arXiv:1703.08019 [Cs], March 23, 2017.