Comparison of neural networks for segmentation of vocalizations

Vocalizations such as speech can be segmented into units like words or phonemes. Neural networks for speech to text typically avoid segmenting into units of time, instead mapping directly from the input features to a sequence of text. Neglecting the segmentation problem allows these networks to achieve high accuracy. However there are many cases where it is desirable to find segments, such as diagnosis of speech disorders.

To understand how the brain learns and produces speech and similar motor skills like playing the piano, many neuroscientists study songbirds. Songbirds learn their songs as juveniles from adult tutors, much like a baby learns to talk from its parents. I have previously shown with collaborators that birdsong provides a good testbed for networks that segment vocalizations, and that a hybrid convolutional-recurrent neural network can outperform previously proposed architectures that segment birdsong into its constituent elements, called syllables.

https://youtu.be/1XEDFpUGmqs

https://nickledave.github.io/neural-network-segment-birdsong.html

However it remains unclear whether networks for segmentation require recurrent connections, or whether alternatively a fully convolutional network can recover segments. It is also unclear how robust the different architectures are to noise. Initial results suggest that a dilated convolutional network yields the best segmentation and is most robust to noise. Experiments in progress test performance on a subset of the TIMIT speech dataset as well. I will discuss whether recurrent connections are always advantageous in neural networks or whether convolutional architectures can always be competitive, provided the input can be represented as an image.