Neural Networks and Natural Language

Vladimir Feinberg

July 19, 2017

We review techniques for natural language processing (NLP) with DNNs. Content is mostly from Dr. Goodfellow's Deep Learning Book, but also taken from Dr. Hinton's Coursera Class, lecture week 4.

1 Word Prediction

With many word outputs, softmax penalties make extremely sparse gradients if each word is a class. Resolve this per Mikolov et al 2013 by moving the output class into the input, and output a single scalar probability when parameterized. This is the serial architecture, used for predicting the next word in a sequence (Fig. 2). The serial architecture takes a long time to find candidates which are assessed by the model as likely, and it

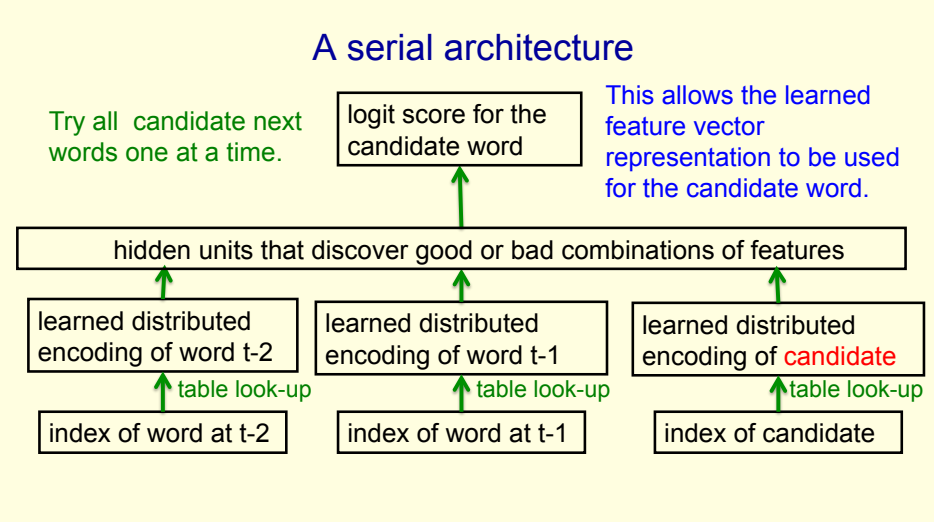

Figure 1: The serial architecture, which folds output complexity into input complexity, from Hinton's Coursera course lecture slides, week 4, slide 25.

can be improved to consider fewer candidates (Mnih and Hinton 2009).

2 Text Classification

Text classification deals with taking usually variable-length sequences of text, and extracting a label (such as spam/not-spam or a topic).

Though variable-length text parsing does require an RNN, a component of this RNN may be convolutional. A grid topology may be induced by concatenating embedding vectors for a sentence into a matrix, and treating that as an image. See this blog for details. Though convolutions may not capture relationships outside of their fixed width, they still may be useful as lower-level feature detectors.

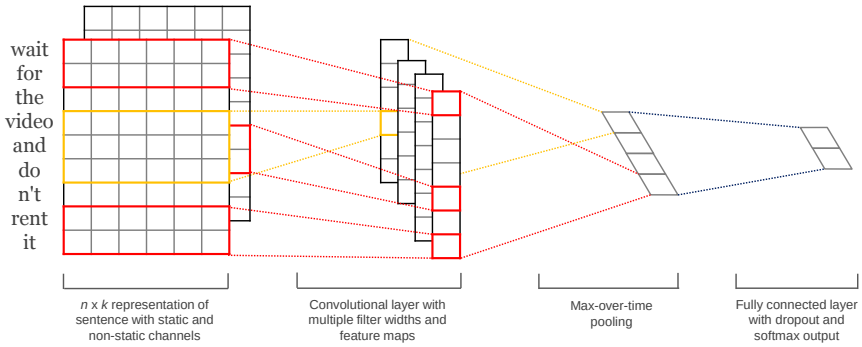

Figure 2: Figure 1 from Kim 2014 demonstrates how matrices of word embeddings can be used for sentence classification