CS 224N: Assignment 3

RYAN MCMAHON SATURDAY 15TH APRIL, 2017

Contents

Problem 1: A Window Into NER (30 pts)

See "~/03-HW3/assignment3.pdf" for the full introduction to the question.

... With these, each input and output is of a uniform length (w and 1 respectively) and we can use a simple feedforward neural net to predict $\boldsymbol{y}^{(t)}$ from $\tilde{\boldsymbol{x}}^{(t)}$: As a simple but effective model to predict labels from each window, we will use a single hidden layer with a ReLU activation, combined with a softmax output layer and the cross-entropy loss:

$$
\begin{aligned}
\boldsymbol{e}^{(t)} &= [\boldsymbol{x}^{(t-w)}L, \ldots, \boldsymbol{x}^{(t)}L, \ldots, \boldsymbol{x}^{(t+w)}L] \\
\boldsymbol{h}^{(t)} &= \text{ReLU}(\boldsymbol{e}^{(t)}W + \boldsymbol{b}_1) \\
\hat{\boldsymbol{y}}^{(t)} &= \text{softmax}(\boldsymbol{h}^{(t)}U + \boldsymbol{b}_2) \\
J &= \text{CE}(\boldsymbol{y}^{(t)}, \hat{\boldsymbol{y}}^{(t)}) \\
\text{CE}(\boldsymbol{y}^{(t)}, \hat{\boldsymbol{y}}^{(t)}) &= -\sum_i y_i^{(t)} \log(\hat{y}_i^{(t)}) \ ,
\end{aligned}
$$

where $L \in \mathbb{R}^{V \times D}$ are word embeddings, $\boldsymbol{h}^{(t)}$ is dimension H and $\hat{\boldsymbol{y}}^{(t)}$ is of dimension C, where V is the size of the vocabulary, D is the size of the word embedding, H is the size of the hidden layer and C is the number of classes being predicted (here 5).

1.1 (a) Conceptual (5 pts)

1.1.1 i) (2 pts)

Provide 2 examples of sentences containing a named entity with an ambiguous type (e.g. the entity could either be a person or an organization, or it could either be an organization or not an entity).

Answer:

1. What have you heard about Louis Vuitton?
2. We had dinner at that new restaurant, Frank's, last night.

1.1.2 ii) (1 pt)

Why might it be important to use features apart from the word itself to predict named entity labels?

Answer:
The word feature matrix is an extremely sparse representation format, wherein it is going to be difficult to recognize entities that don't appear very often.

1.1.3 iii) (2 pts)

Describe at least two features (apart from the word) that would help in predicting whether a word is part of a named entity or not.

Answer:
The most obvious additional predictor would be capitalization. Another would be part-of-speech tags (e.g., if the previous word, $w^{(t-1)}$, is a determiner, the current word is more likely to be an entity).

1.2 (b) Network Components (5 pts)

1.2.1 i) (2 pts)

What are the dimensions of $\boldsymbol{e}^{(t)}$, W and U if we use a window of size w?

Answer:

1. $\boldsymbol{e}^{(t)}$ is going to be a row vector of length $(2w + 1) \times D$
2. W is going to be a matrix of dimensionality $|\boldsymbol{e}^{(t)}| \times H$
3. U is going to be a matrix of dimensionality $H \times C$

1.2.2 ii) (3 pts)

What is the computational complexity of predicting labels for a sentence of length T?

Answer: