COMP9444 Neural Networks and Deep Learning

Quiz 6 (Language Processing)

This is an optional quiz to test your understanding of the material from Week 6.

---

1. What are the potential benefits of continuous word representations compared to synonyms or taxonomies?

   Synonyms, antonyms and taxonomy require human effort, may be incomplete, and force discrete choices. Continuous representations have the potential to capture gradations of meaning and more fine-grained relationships between words, as well as being extracted automatically without human involvement.

2. What is meant by the Singular Value Decomposition of a matrix X? What are the special properties of the component matrices? What is the time complexity for computing it?

   The Singular Value Decomposition of X is $X = U\ S\ V^T$ where U, V are unitary (all columns of unit length) and S is diagonal with all entries ≥ 0.
   The time to compute it is proportional to $L \times M^2$ if X is L-by-M and $L \geq M$.

3. What cost function is used to train the word2vec skip-gram model? (remember to define any symbols you use)

   If the text is w_1 ... w_T then the cost function is

   $-(1/T) \sum_{t=1}^{T} \sum_{-c \leq r \leq c,\ r \neq 0} \log \text{prob}(w_{t+r} \mid w_t)$

4. Explain why full softmax may not be computationally feasible for word-based language processing tasks.

   The number of outputs is equal to the total number of words in the lexicon (approximately 60,000) and all of them would need to be evaluated at every step.

5. Write the formula for Hierarchical Softmax and explain the meaning of all the symbols.

   $\text{prob}(w = w_t) = \prod_{j=1}^{L(w)-1} \sigma([n(w, j+1) = \text{child}(n(w, j))]\ \mathbf{v'}_{n(w,j)}^{\mathrm{T}}\ \mathbf{h})$

   $n(w,1)$, ..., $n(w,L(w))$ are the nodes along the path in a Binary Search Tree from the root to w_t
   $\mathbf{h}$ = hidden unit activations, $\sigma(u) = 1/(1 + \exp(-u))$,
   $[n' = \text{child}(n)] = +1$, if n' is left child of node n; -1, otherwise.

6. Write the formula for Negative Sampling and explain the meaning of all the symbols.

   $E = -\log \sigma(\mathbf{v'}_{j*}^{\mathrm{T}}\ \mathbf{h}) - \sum_{j \in W\text{neg}} \log \sigma(-\mathbf{v'}_{j}^{\mathrm{T}}\ \mathbf{h})$
   j* = target word, Wneg = set of negative examples drawn from some distribution

7. From what probability distribution are the negative examples normally drawn?

   $P(w) = U(w)^{3/4}/Z$,
   $U(w)$ = Unigram distribution determined by previous word,
   Z = normalizing constant.

8. Explain how an *attention mechanism* can improve the performance of a neural machine translation system.

An additional network layer assigns a weighting to each timestep in the RNN or LSTM, in order to focus attention on those timesteps which are most relevant for choosing the next word in the translated sentence.
