UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11061 NATURAL LANGUAGE UNDERSTANDING (LEVEL 11)

Monday 8th May 2017

14:30 to 16:30

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY. If both QUESTION 2 and QUESTION 3 are answered, only QUESTION 2 will be marked.

All questions carry equal weight.

CALCULATORS MAY NOT BE USED IN THIS EXAMINATION

MSc Courses

Convener: F. Keller
External Examiners: A. Burns, P. Healey, M. Niranjan

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. THIS QUESTION IS COMPULSORY

The following questions are short answer questions. Answer as concisely as possible. Often a single sentence is sufficient; never write more than a paragraph. Longer answers will not receive more credit.

(a) Latent Dirichlet Allocation (LDA) and skip-gram are two models for learning word vector meaning representations.

i. Describe how the two models are trained and list two major differences between them. [*3 marks*]

ii. Using LDA and skip-gram, how would you express the similarity of two words w_1 and w_2? Provide formulas with your answer. [*2 marks*]

iii. Both LDA and skip-gram yield embeddings. What do these embeddings represent in each model? [*2 marks*]

(b) Give the formula for the reconstruction error in Socher's recursive autoencoder. [*2 marks*]

(c) Could you use a recurrent neural network to obtain compositional sentence representations? Justify your answer. [*2 marks*]

(d) Neural networks have been successfully applied in a variety of NLP tasks such as sentiment classification and paraphrase detection. Can you think of a modeling scenario where it would not be a good idea to use neural networks? Justify your answer. [*2 marks*]

(e) What are the two key features that set Bayesian estimation apart from maximum likelihood estimation? Use formulae as appropriate. [*2 marks*]

(f) Why is a Bayesian HMM for unsupervised part-of-speech tagging much more effective if it uses a lexicon? [*2 marks*]

(g) The long short-term memory (LSTM) is a standard architecture used for many NLP tasks.

i. The LSTM uses a set of gates to control information flow: output gate, input gate, forget gate. Describe how these gates work, and which problem with other architectures they solve. [*3 marks*]

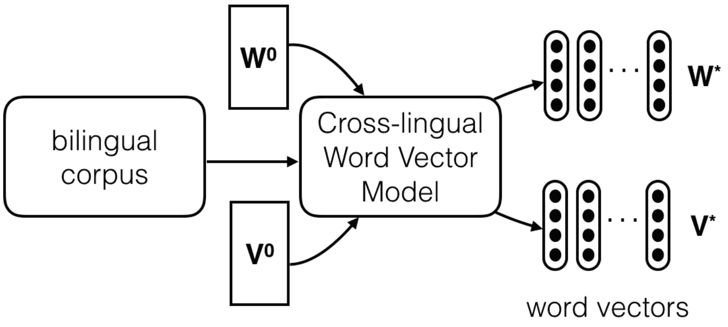
ii. How could you use an LSTM to do unsupervised part of speech tagging? Sketch the input and out representations you would use. [*3 marks*]

iii. When doing unsupervised part of speech tagging with an LSTM, you want to test if using a lexicon improves performance. How would you do this? [*2 marks*]

2. ANSWER EITHER THIS QUESTION OR QUESTION 3

The skip-gram model learns embeddings for individual words in a single language (e.g., English). We have also seen techniques for learning embeddings for sentences and documents using *monolingual* distributional information. In this question we will adapt some of these models to work on *parallel* or *multilingual* data. A parallel corpus is a collection of texts, each of which is translated into one or more other languages. We will look at the simplest case where only two languages are involved: one of the corpora is an exact translation of the other.

Within a monolingual context, the distributional hypothesis forms the basis of most approaches for learning word representations. We will extend this hypothesis to bilingual data and learn bilingual embeddings using a parallel corpus.



Algorithm 1 General Algorithm

```
Initialize $\mathbf{W} \leftarrow \mathbf{W}^0$,$\mathbf{V} \leftarrow \mathbf{V}^0$
$(\mathbf{W}^*, \mathbf{V}^*) \leftarrow \arg\min \alpha A(\mathbf{W}) + \beta B(\mathbf{V}) + C(\mathbf{W}, \mathbf{V})$
```

Figure 1: Schema for induction of crosslingual word vector representations.

(a) A general schema for inducing bilingual embeddings is shown in Figure 1. A word vector model generates embeddings which incorporate distributional information *crosslingually*. At the bottom of Figure 1 a general algorithm for bilingual word embeddings is shown, where α, β, $\mathbf{W}^0$, $\mathbf{V}^0$ are parameters and A, B, C are suitable objectives. A and B are monolingual objectives, whereas C is a bilingual objective.

Assume that you will be training your model on a bilingual corpus (e.g., English–French) containing sentence and word alignment information. In other words, you can assume that any two parallel sentences are translations of each other and that you know which words are translations of each other. Example word alignments are shown in Figure 2a.

How would you extend the skip-gram model to learn bilingual embeddings following the schema in Figure 1? Define the objectives A, B, and C.

QUESTION CONTINUES ON NEXT PAGE

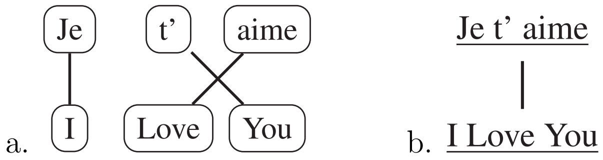


Figure 2: French and English parallel sentences with word alignments (a) and sentence alignments (b), indicated by arcs.

Notation: Let $W = \{w_1, w_2, \dots w_{|W|}\}$ be the vocabulary of a language l_1 with $|W|$ words and $V = \{v_1, v_2, \dots, w_{|V|}\}$ be the vocabulary of another language l_2 with $|V|$ words. Let $\mathbf{w}$ denote the vector for word w. [*8 marks*]

(b) What is the role of the hyper-parameters α and β in the algorithm in Figure 1? [*4 marks*]

(c) Now assume that you don't have access to any word alignment information. All you know is that parallel sentences are translations of each other, and you want to leverage the fact that aligned sentences have equivalent meaning, and thus their sentence representations should be similar. Example sentence alignments are shown in Figure 3b. First show how you will represent the aligned sentences in two languages. Assume two functions f and g which map sentences $\vec{v} = \langle \mathbf{x_1}, \dots, \rangle$ and $\vec{w} = \langle \mathbf{y_1}, \dots, \rangle$ into semantic representations (where $\mathbf{x_i} \in \mathbf{V}$, $\mathbf{y_i} \in \mathbf{W}$ are vectors corresponding to the words in the sentences). Discuss at least two ways of representing the sentences compositionally. Then, define objective C over *entire* sentences. **Hint:** use the distance between the two sentence representations as objective. A sketch of the model is shown in Figure 3. [*8 marks*]

(d) Now imagine that you don't have alignments at the sentence level but at the document level. How would you change the model in Figure 3 to extend to document-level learning? Draw your modified model. [*5 marks*]

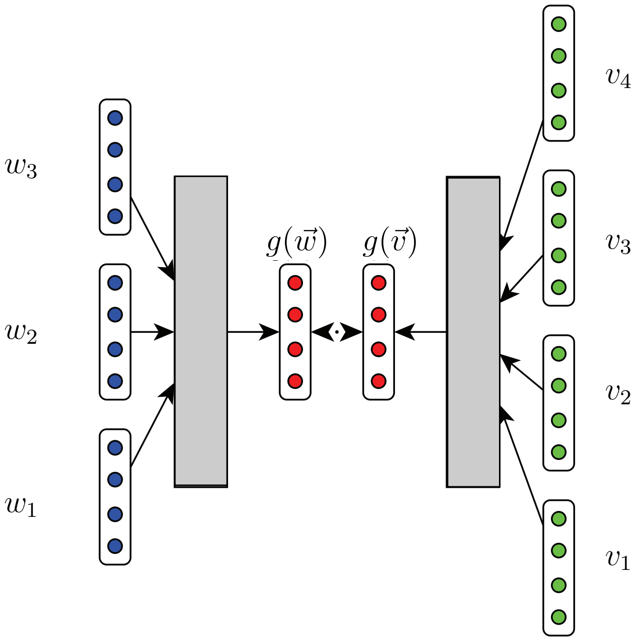


Figure 3: Model for induction of word vector representations over bilingual sentences. Gray boxes indicate composition functions.

3. ANSWER EITHER THIS QUESTION OR QUESTION 2

Image description is the task of taking an image I and generating a sentence S that describes the visual content of the image. An example of an image with a description is given in Figure 4.

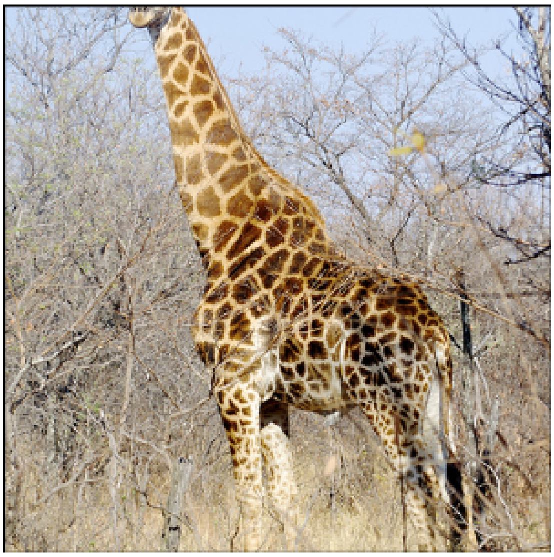
In this question, you will develop a neural network architecture that is able to describe images. Your training data will be a set of images with human-generated descriptions. At test time, you only have the images, and your model needs to generate the descriptions. We will assume that images are represented as sets of pixels, while descriptions are represented as strings of words.

To estimate the parameters θ of your model, you need to use the following formulation:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{(I,S)} \log p(S|I;\theta) \tag{1}$$

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1}) \tag{2}$$

where $S_0, \ldots, S_N$ are the words in a description S with length N.



"a giraffe standing in a grassy area with trees in the background"

Figure 4: Example for an image with a textual description.

QUESTION CONTINUES ON NEXT PAGE

(a) Design a neural network that computes $p(S|I)$ as defined in equation (2). For this, assume that you have a pre-existing convolutional neural network (CNN) that turns an image (a set of pixels) into a suitable feature representation. You can also assume pre-trained word embeddings. You will need to integrate the CNN with another type of network in order to generate image descriptions. Provide a diagram of your architecture and explain using formulae how your network computes the probabilities in equation (2). [*10 marks*]

(b) How would you train your network, i.e., which loss function would you minimize, and what training algorithm would you use? [*4 marks*]

(c) Once you have trained your network, how would you use it to generate sentences (image descriptions) at test time? Describe a procedure for generating descriptions word by word. What potential problem would you encounter with this word-by-word approach, and how can you address this problem? [*6 marks*]

(d) How would you evaluate your model, assuming that you have a test set that contains human-generated image descriptions? How would you evaluate your model if you don't have such a test set? Discuss the advantages and disadvantages of both approaches. [*5 marks*]