Solutions to Practice Midterm 2

Part I

1.
   a. True. SVD is often used when it comes to dimension reduction.
   b. True, the two important keywords is that it is an optimal **linear** auto-encoder under **L2** loss.
   c. False. It is not optimal under cross-entropy loss, as it does not consider softmax or cross-entropy. It is optimal for L2 loss.
   d. False, negative sampling is a concept from word2vec. Here we just have a matrix we try to decompose and recompose optimally.
   e. False, there's no notion of center and context words, this is a word2vec concept.

2.
   a. False, we are considering pairs of (word, context) that are in close proximity in some document in our dataset. Furthermore, one motivation for developing word2vec is precisely to avoid storing all our words in a giant data matrix, which has numerous downsides.
   b. True. See above.
   c. False, the output loss is a cross-entropy: we are classifying whether a word2 was in the context of word1 (binary classification).
   d. Correct, see above.
   e. True, there are two matrices, one used for the center words and one for the context words. This means each word is actually represented by two vectors. While this sounds parameter inefficient, it empirically works better and it is still far more efficient than storing global information in a giant matrix.
   f. True, word analogies are observed to happen in word vectors trained using word2vec.

3.
   a. True.
   b. False, the loss in Glove is more complicated and does not directly minimize L2 of the original matrix. See our lectures for details.
   c. True. The model and loss were designed to improve performance on word analogy tasks.
   d. False, there is again a center-word embedding and a context word embedding.

4.
   a. False, skip-thought models generating sentence embeddings, not word embeddings.
   b. True.
   c. True. This and the above are the 2 tasks it is trained on.

   d. False.
   e. True.

5. GANs question.
   a. False. While it's true that the discriminator has to distinguish between two images, and one of them is a real image, the second image is some other image that was generated by the Generator. It is not the synthetic image that "most closely matches the real image" but one that is sampled (and generated) by the Generator.
   b. False, it seems better to keep parameterizations separate as they are "competing" against each other in the minimax game.
   c. True, see the paper and (Monday's) slides for details.
   d. False. To be effective the discriminator must be repeatedly retrained, otherwise it will be easy for the generator to create non-realistic fooling images for it.
   e. False, although the images are generated using an adversarial network, they are not adversarial. Humans will normally classify the images to the correct category: i.e. the one used to generate them.

6. [Not relevant for this midterm]

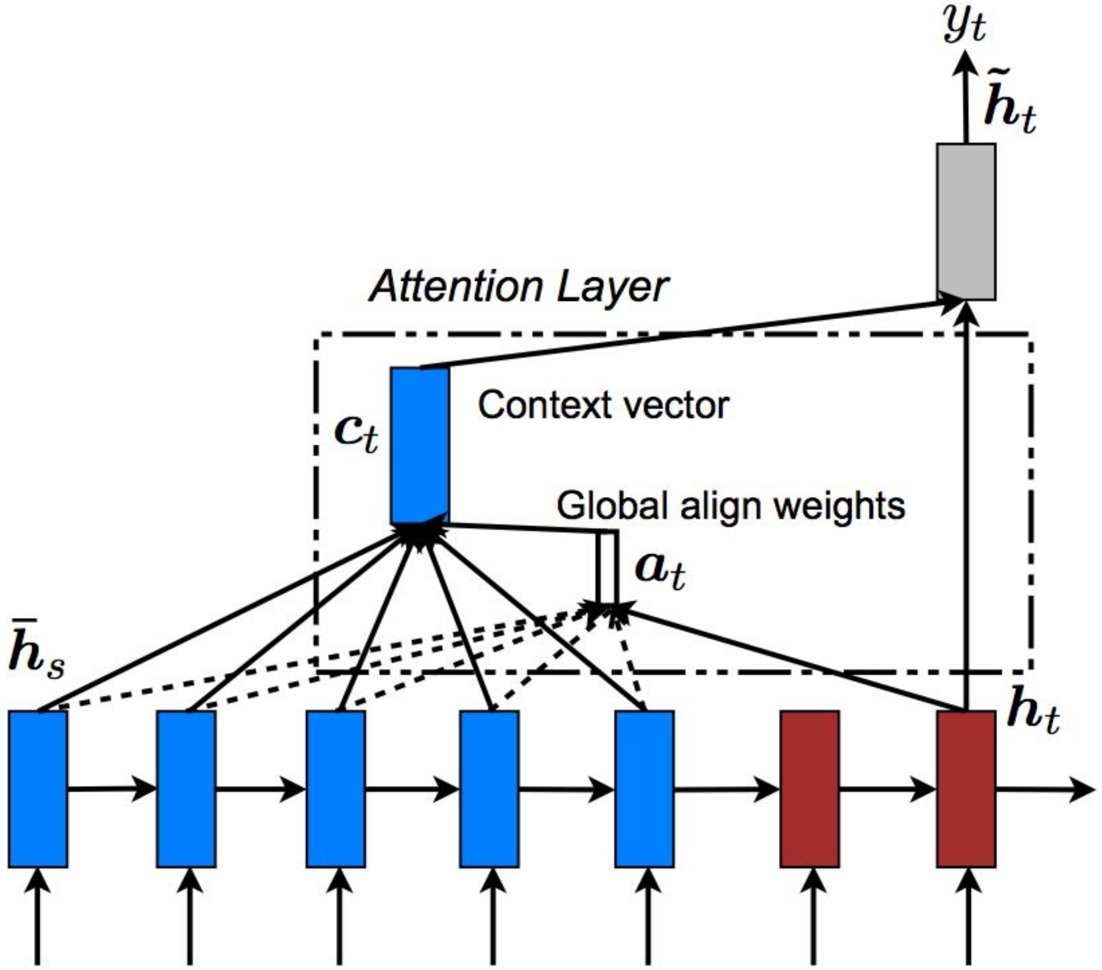
7. [Not relevant for this midterm]

8.
   a. False, targeted examples are harder to generate. Many adversarial examples misclassify to a class different from the one used to train them, that is, they are untargeted. To create an adversarial example that is targeted (i.e. reliably misclassifies to a given class) one must work harder, e.g. by finding a common perturbation that misclassifies to the same class on many different classifiers.
   b. Yes. The sign of the gradient also works.
   c. Yes.
   d. Yes (however this is not impervious to attack, see transferability)

9. [Not relevant for this midterm]

<u>Part II</u>

10. That's basically what this figure does:



11.

a. Position encoding is used to ensure that word position is known. Because attention is applied symmetrically to all input vectors from the layer below, there is no way for the network to know which positions were filtered through to the output of the attention block. Position encoding also allows the network to compare words (nearby position encodings have high inner product) and find nearby words.

b. Multi-Head attention allows for a single attention module to attend to multiple parts of an input sequence. This is useful when the output is dependent on multiple inputs (such as in the case of the tense of a verb in translation). Attention heads find features like start of sentence and paragraph, subject/object relations, pronouns etc.

c. When k is the hidden dimension of the network:
   i. O(M^2k)
   ii. O(MNk)

iii. O(N^2k)

d. No. The encoder activations do not depend on the decoder activations. Thus, you only need O(MN + N^2) additional computation to decode into a new sequence.

12.

a. Yes. These models are trained end-to-end with SGD (Hence the title of the paper “End to End Memory Networks”). This is because of the inclusion of softmax over the hard-max in traditional memory nets.

b.

1 Mary moved to the bathroom.
2 John went to the hallway.
3 Where is Mary? bathroom
4 Daniel went back to the hallway.
5 Sandra moved to the garden.
6 Where is Daniel? hallway
7 John moved to the office.
8 Sandra journeyed to the bathroom.
9 Where is Daniel? hallway
10 Mary moved to the hallway.
11 Daniel travelled to the office.
12 Where is Daniel? office

c. Large datasets can be bad for attention-based architectures, as it can be difficult to learn the embeddings as the attention is relatively spread out - and the gradients to useful parts of the attention can be small. For larger datasets, full-text search can be used to find passages of the input text that are likely to be relevant and the memory network applied only to those segments.

d. The major difference is that in traditional memory networks, the keys and values are computed from the context directly. Key-Value networks compute a unique set of keys and values for each input context element (and can thus leverage additional information such as the context of a word).