

## Practice Midterm 2

NAME \_\_\_\_\_ SID \_\_\_\_\_

**This is a closed-book exam with 12 questions. You can use one sheet of notes. Please write all your answers in this book. There is a total of 75 points for all questions, and you have 75 minutes to complete the exam. Please budget your time accordingly. Good luck!**

**Part I (45 points)** The following short answer-questions are *multiple-answer*. Be sure to circle *\*all\** the options that apply.

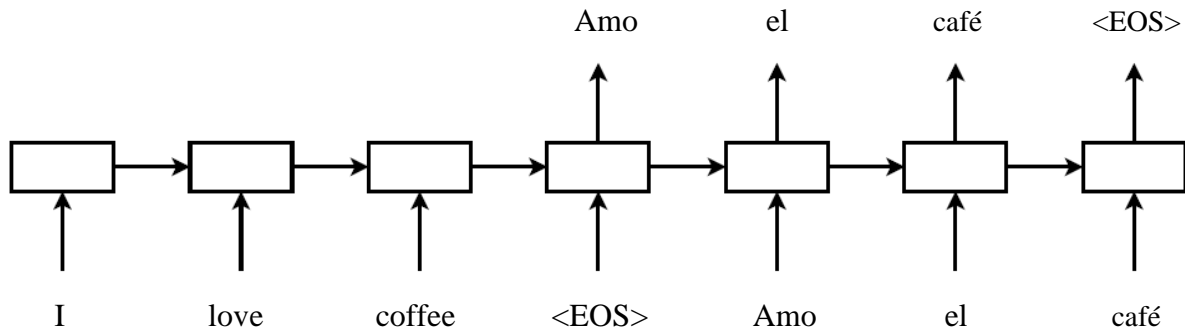
1. (5 points) Latent Semantic Analysis (LSA) is used to embed words in a vector space, and is computed from a document/word matrix  $T$ , each of whose rows is a BoW representation of a document. Circle *\*all\** the correct answers that apply:
  - a) LSA can be computed using an SVD (singular value decomposition) of the matrix  $T$ .
  - b) LSA is an optimal linear auto-encoder for L2 loss on reconstruction of  $T$ .
  - c) LSA is an optimal linear auto-encoder for cross-entropy loss on reconstruction of  $T$ .
  - d) LSA sometimes uses negative sampling to improve the performance of its output softmax.
  - e) LSA computes two word embedding matrices  $U$  and  $V$  for center and context words respectively.
  
2. (6 points) For the word2vec algorithm, circle all correct answers below:
  - a) Word2vec takes as input a document-word matrix  $T$  similar to LSA.
  - b) Word2vec models local structure between a center word and neighboring words up to a fixed distance in the input text sequence.
  - c) Word2vec models can be computed using a deep network with L2 output loss.
  - d) Word2vec models can be computed using a deep network with cross-entropy output loss.
  - e) Word2vec computes two word embedding matrices  $U$  and  $V$  for center and context words respectively.
  - f) Implements word analogies using the formula:  $\text{vec}(\text{"woman"}) - \text{vec}(\text{"man"}) + \text{vec}(\text{"aunt"}) - \text{vec}(\text{"uncle"}) \approx 0$ .
  
3. (4 points) For the GloVe algorithm, circle all correct answers below:
  - a) GloVe takes as input a co-occurrence matrix  $C$  where  $C_{ij}$  = number of occurrences of word  $i$  near word  $j$ .
  - b) GloVe minimizes the L2 loss between original and predicted co-occurrence matrix  $C$ .
  - c) GloVe is optimized to improve the accuracy of analogies.
  - d) GloVe generates a single embedding matrix  $V$  for both context and center words.

4. (5 points) For the skip-thought embedding method, circle all correct answers:
- a) Skip-thought models include separate embedding matrices  $U$  and  $V$  for center and context words.
  - b) Skip-thought models include a loss for prediction of the next sentence from the current sentence.
  - c) Skip-thought models include a loss for prediction of the previous sentence from the current sentence.
  - d) Skip-thought loss is explicitly optimized for word analogies.
  - e) Skip-thought models embed entire sentences as a single context vector.
5. (5 points) For Generative Adversarial Networks (GANs) circle all that apply:
- a) A discriminator is trained to distinguish a real image from a synthetic image that most closely matches the real image.
  - b) The discriminator and generator should share parameters for best performance.
  - c) A GAN (locally) minimizes the Jensen-Shannon divergence between the input image distribution and the distribution of generated images.
  - d) The discriminator should always be fully trained first, and then the generator can be trained using negative discriminator loss.
  - e) The images generated by a GAN are adversarial (fooling) images.
6. (6 points) For imitation learning of a control policy  $\pi$ , circle all correct answers:
- a) Imitation learning relies on on-policy data.
  - b) DAgger trains with on-policy trajectories, with action selection by a human teacher.
  - c) In pure imitation learning, the target policy is executed by a human and the behavior policy is learned by the agent.
  - d) Training on off-policy data can lead the trained policy to diverge from an expert trajectory.
  - e) The more off-policy data is used for training (assuming only off-policy data is available), the less accurate and stable is the learned policy.
  - f) The gap between off-policy data and on-policy training can sometimes be bridged by domain adaptation.

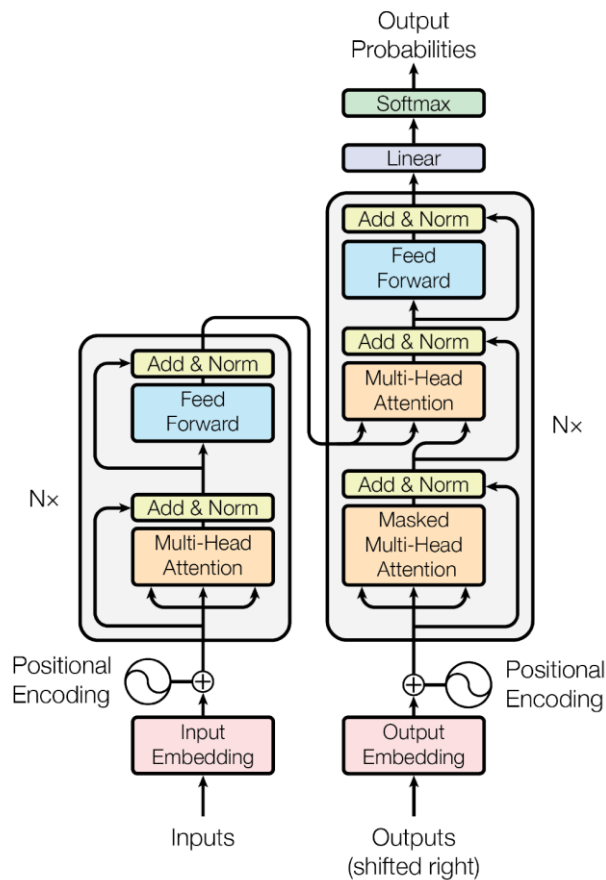
7. (4 points) For GAIL (Generative Adversarial Imitation Learning), circle all that apply:
- a) The policy  $\pi$  is trained to minimize a cost function learned by Inverse Reinforcement Learning (IRL).
  - b) An adversary is trained to discriminate expert (real) and policy-generated (synthetic) trajectories based on the sequence of actions performed.
  - c) An adversary is trained to discriminate expert (real) and learned (synthetic) trajectories based on the sequence of (state, action) pairs along the trajectory.
  - d) The policy parameters are optimized to maximize the expected reward only.
8. (5 points) Adversarial Examples. Circle all the answers that are correct:
- a) It is harder to generate non-targeted adversarial examples than targeted examples.
  - b) Adversarial examples can be generated in a single step by perturbing images in the direction of gradient of the loss.
  - c) Adversarial examples can be generated iteratively by maximizing the label loss with an optimizer such as SGD.
  - d) Ensembles of models are used to resist against adversarial image attacks.
9. (5 points) Negotiation dialog agents have the following characteristics. Circle all that apply:
- a) Agents always reach a deal.
  - b) Agents trained solely to mimic human dialog generate human-understandable dialog and optimal rewards.
  - c) Agent expected reward is a differentiable function of model parameters, allowing end-to-end training of reward-optimal agents with methods like SGD.
  - d) Agent dialog is consistent, i.e. agents only accept the exact offer made by the other agent.
  - e) Rollouts – i.e. multi-step simulation of dialog and rewards, improve expected agent rewards.

**Part II: General questions (30 points)**

10. (10 points) The diagram below shows a design for a language translation system without attention. Modify the design to include soft attention. Show only one copy of the unit that computes attention weight and use word-position indices as appropriate to explain its inputs and output.



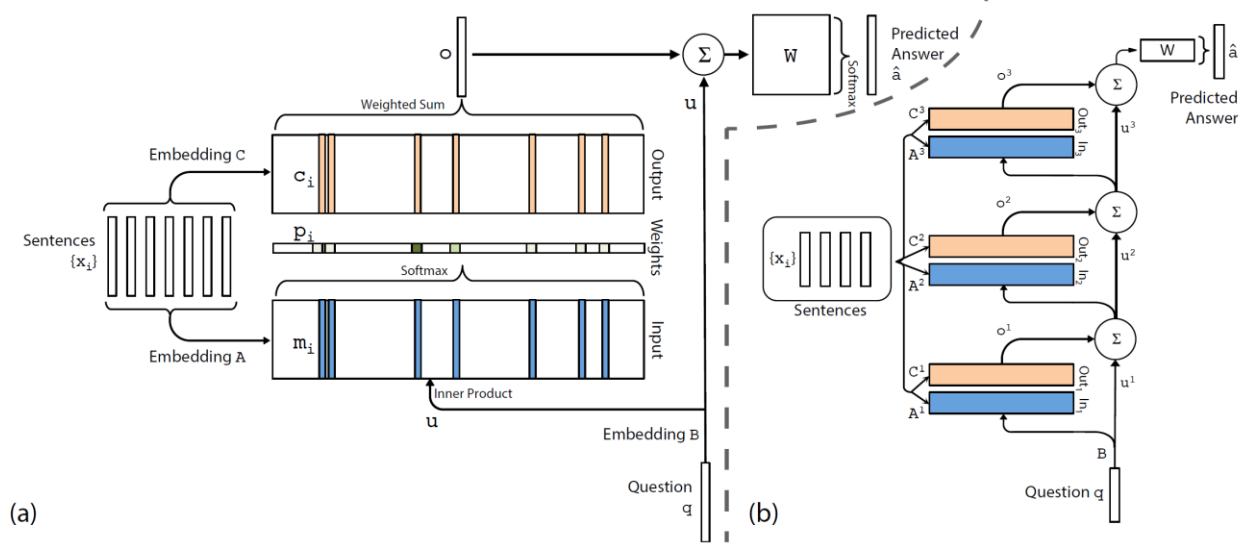
11. (10 points) Machine Translation. A diagram of a transformer network stage is shown below:



Answer the following questions about this network:

- (3 points) What is the reason for positional encoding, and how is it typically implemented? Use diagrams as appropriate.
- (2 points) What is the advantage of multi-head attention (vs. a single head). Give examples of what structure can be found by multi-head attention.
- (3 points) For input sequences of length  $M$ , output sequences of length  $N$ , what are the complexities of
  - Encoder self-attention
  - Decoder-encoder attention
  - Decoder self-attention
- (2 points) Do activations of the encoder depend on decoder activations? How much additional computation is needed to translate a source sequence into a different target language, in terms of  $M$  and  $N$ ?

12. (10 points) Memory Networks. A basic memory network is shown below:



- (2 points) Is this network typically trained end-to-end with gradient methods (e.g. SGD) or does it require reinforcement learning? Explain.
- (4 points) Give an example of a query that requires multiple hops to answer. You may wish to list a sequence of input facts that the networks refers to. In terms of number of hops, how would you decide if your query is answerable with the network above?
- (2 points) What difficulties does this design have when used for queries over large databases? Give one solution that addresses these difficulties.
- (2 points) What is the difference between the network above and a key-value memory network? Give one example of why keys and values are different and how this is useful.