MT Marathon 2016 – Written Exam

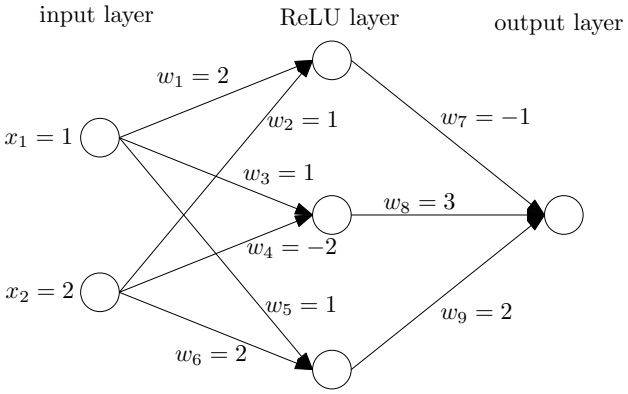
Please answer the questions directly on this paper. Note that the answers will be checked by some other participant.

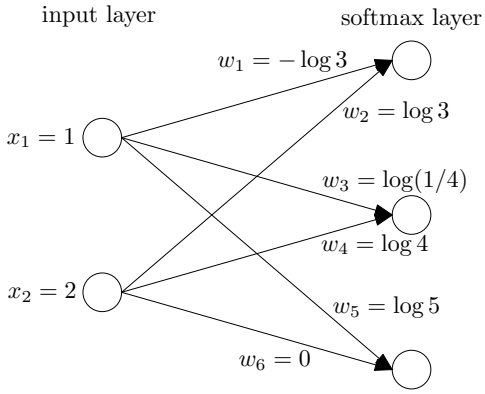
The final exam score is a complex number – some questions from the exam were not discussed on the slides, but are important NN topics. These questions are written in ***italics*** and are awarded with imaginary points.

You can get a hint for any question (for example if you are unsure how to compute softmax or how to derivate it) – just raise your hand and I will come to you. You will be awarded only 50% of points for such questions.

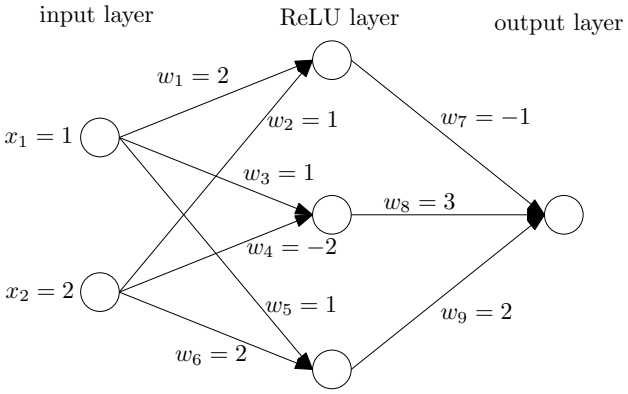
1) Compute BLEU (up to bigrams) and TER of the following translations.
Human Translation: *the cat jumped on the table*
Machine Translation: *the table the the cat*

2) Compute the output of the following neural networks.

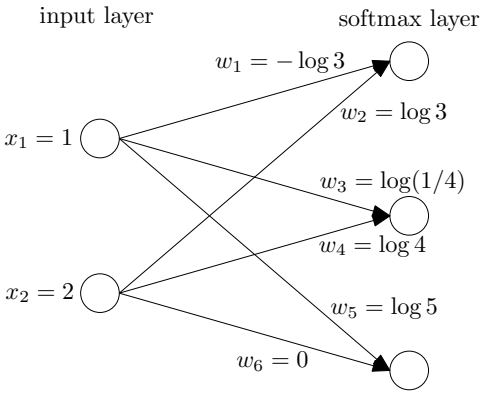




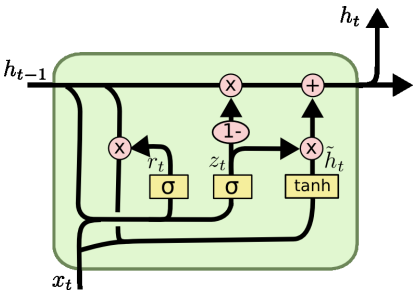
3) Compute derivatives of the networks with respect to all weights and inputs.
For the first network, the gold output is 3, and the loss function is MSE, i.e., $(output - gold)^2$.



For the second network, the gold output is the distribution $(0, 0, 1)$ and the loss function is cross-entropy (which in the *single correct class* case is equal to log probability of the correct class).



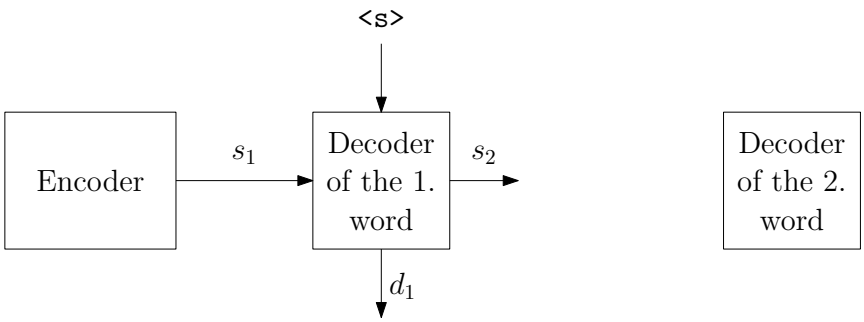
4) Write explicit GRU equations.
Compute h_t using h_{t-1} and x, using the GRU sketch from the slides. Include all weight matrices and bias terms.



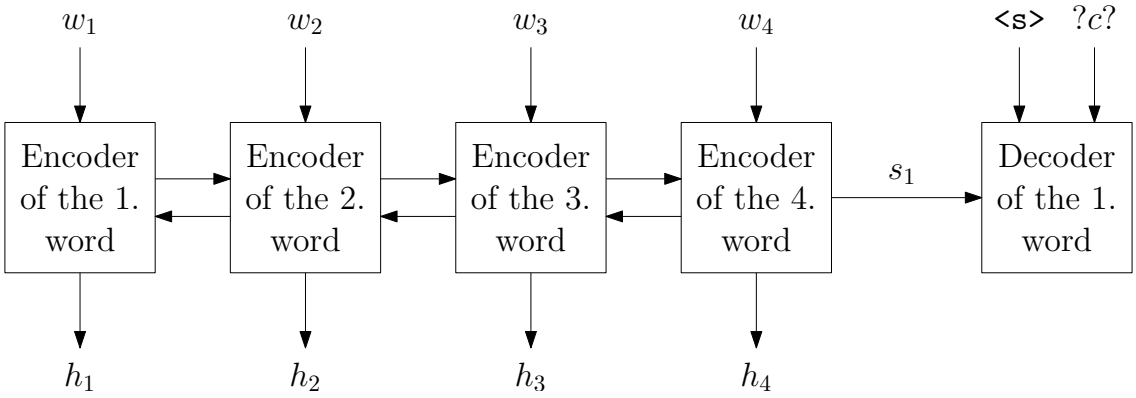
5) Explain (with formulas) how can we handle the exploding gradient problem.

6) What is vanishing gradient problem in plain RNNs? Try explaining why it occurs.

7) Assume basic encoder-decoder model without attention. Write/draw the inputs used for decoding the second word during inference time.
Include all operations like tanh, softmax, linear transformations, biases, embeddings, ...



8) Assume encoder-decoder model with attention. Write how exactly is computed the attention c used when decoding the first word.



9) Compute 3 steps and then further 4 steps of BPE on the following dictionary.

Word	*Frequency*
t a l l </w>	2
t a l l e r </w>	1
t a l l e s t </w>	1
b e s t </w>	1

10) Explain the skip-gram model of Mikolov et al.
Consider the basic skip-gram model (which uses full softmax output) with context size 1. Draw the network used in the model, specify exactly how is the output computed (notably it should be clear what all parameters of the model are), and try writing the loss function.

In practice, the full softmax output is too slow. If you know negative sampling, try sketching out how it works.

Recently, structures skip-gram (SSkip-gram) model is gaining popularity. If we tell you that it uses a separate output matrix for every context offset, try guessing how it is computed.

11) ***List as many NN training algorithms you know apart from standard SGD.***

12) ***Name as many NN regularization methods you know.***

13) ***Why are minibatches regularly used in NN training? Write at least two reasons.***