Assignment 3

Problem 1 a)

i) Carnegie Mellon awarded the students for their accomplishments

I am calling you from Stanford.

ii) Using features provides additional information, to help discern between ambigous named entities, like those above.

iii) Two additional feautures are part of speech and surrounding context of a word.

b)

i) $e^{(t)} \in \mathbb{R}^{1 \times (2w+1)D}$

$W \in \mathbb{R}^{(2w+1)D \times H}$

$U \in \mathbb{R}^{H \times C}$

ii)

Computing $e^{(t)}$ takes $O((2w+1)D)$ concatenation operations

Computing $h^{(t)}$ takes $O((2w+1)DH)$ time for matrix multiplication

Computing $y^{(t)}$ takes $O(HC)$ time for matrix multiplication

Overall the process takes $O(T((2w+1)DH + HC))$ time for T operations d)

Best F_1 = 83%

i)

gold\guess	PER	ORG	LOC	MISC	O
PER	2979	39	59	15	57
ORG	152	1639	113	57	131
LOC	59	104	1860	33	38
MISC	41	59	37	1024	107
O	39	51	14	30	42625

The confusion matrix indicates that the neural network misclassifies LOC as ORG and ORG as LOC quite frequently. The most frequent error is PER instead of ORG.

ii) The first limitation is that the model does not keep feature labels of the previous tokens. Something like a language model could be applied to NER.

EX: The fight, Duran's first on home soil for 10 years, is being billed here as the "Return of the Legend" and Duran (PER/ORG) still talks as if he were in his prime.

Here Duran could be infered to be a person from previous words

The second limitation is that the ~~model does not the~~ window based implementation is limited to the window size, and cannot scale to the size of the sentence.

EX: Duran, 45, takes on little-known Mexican Ariel (PER/ORG) Cruz (PER/ORG)

Here, knowing that Duran is a PER could inform us that Ariel is also a PER and not an ORG, but window is too small.

Problem 2

a)

i) The RNN contains

$\boxed{DH + H^2 - (2w+1)DH \text{ more parameters}}$

ii) Embedding lookup takes $O(D)$ time for $e^{(t)}$

Computing $h^{(t)}$ takes $O(DH + H^2)$ time for matrix multiplication

Computing $y^{(t)}$ takes $O(HC)$ time for matrix multiplication

for T sentence lengh T, computation takes

$O(T(HC + DH + H^2))$ time

b)

i) Any situation where precision remains constant but recall decreases, will decrease cross-entropy.

EX: John saw The KINGS SPEECH

PER O MISC MISC MISC

predicting PER O O MISC MISC

versus

PER O O O MISC

ii) F_1 is difficult to optimize because it is hard to differentiate a quotient efficiently.

d)

i) Including the padding labels would cause the RNN to attempt to predict the padding labels, affecting the downstream parameters via bachpropogation

g)

i) One limitation is that the RNN no longer has a context ahead of the label trying to be predicted.

EX: Panamanian (MISC/PER) Boxing Association President Ramon Manzanares said.

"Panamanian" was mislabelled becaase it did not have the future Context

Another limitation is that the RNN can saffer from vanishing gradients, so that long past tokens have no effect on current predictions

EX: Traders said the Fed's decision to adopt a tightening bias at the July FOMC (ORG/O) meeting...

Here the token "Fed" might have been hepful in determining that FOMC is also an ORG.

ii)

The first limitation can be addressed using either a bidirectional RNN or concatenating the current token with its context before input.

The second limitation can be addressed by implementing LSTM instead of RNN.

Problem 3

a)

i) One possible solution $\boxed{w_h = 1 \,,\; w_z = 1 \quad \text{and} \quad u_z = 0}$

ii) One possible solution $\boxed{w_z = 1 \,,\; u_z = 0 \,,\; w_h = 1 \,,\; u_h = 1}$

b)

i) we have 4 possible states

State 1	State 2	State 3	State 4
$h^{t-1} = 0$	$h^{t-1} = 0$	$h^{t-1} = 1$	$h^{t-1} = 1$
$x = 0$	$x = 1$	$x = 0$	$x = 1$
$h^t = 0$	$h^t = 1$	$h^t = 1$	$h^t = 0$

For state 1

$$h^t = \sigma(b_h) = 0 \implies b_h \leq 0$$

For state 2

$$h^t = \sigma(w_h + b_h) = 1$$

$$\implies w_h + b_h > 0$$

$$\implies w_h > -b_h \geq 0$$

$$\implies w_h > 0$$

For state 3

$$h^t = \sigma(u_h + b_h) = 1$$

$$\implies u_h > 0$$

For state 4

$$h^t = \sigma(w_h + u_h + b_h) = 0$$

$$\implies w_h + u_h + b_h \leq 0$$

However we know that

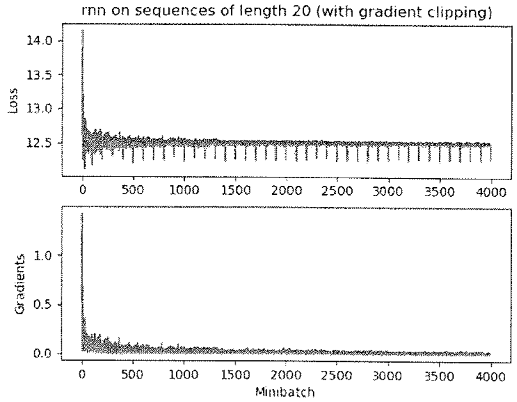
$$w_h + b_h > 0 \quad \text{and} \quad u_h > 0$$

$$\implies w_h + u_h + b_h > 0$$

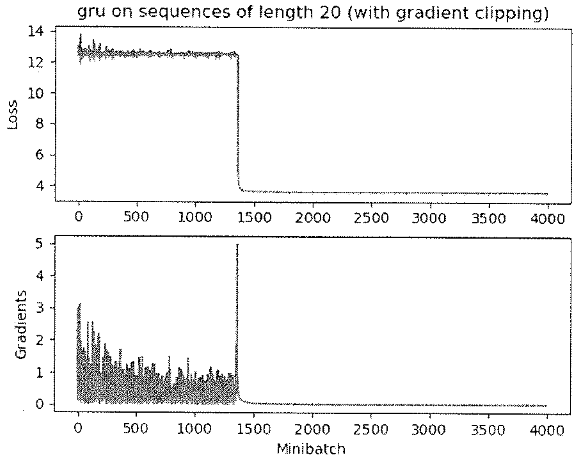
Therefore we have a contradiction,

Therefore an RNN is insufficient to replicate our desired behavior.

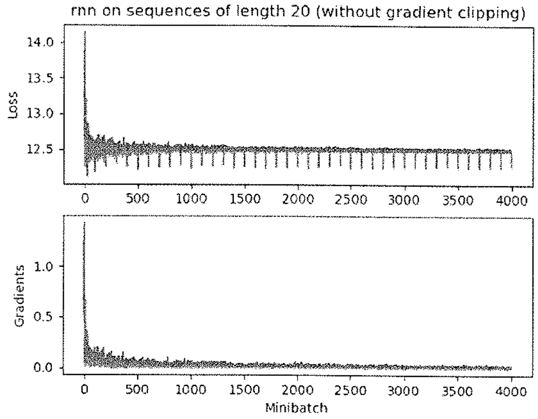
ii) One possible solution is $\boxed{w_z = -1 \,,\; w_h = 1 \,,\; u_z = 1 \,,\; b_r = 1 \,,\; u_r = -1}$



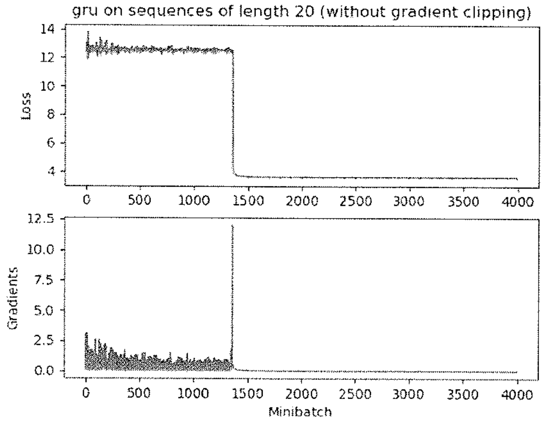
rnn on sequences of length 20 (with gradient clipping)



gru on sequences of length 20 (with gradient clipping)



rnn on sequences of length 20 (without gradient clipping)



gru on sequences of length 20 (without gradient clipping)

i) Oddly enough, GRU appears to experience minor exploding gradients which quickly subsides. Gradient clipping does not appear to make a huge qualitative difference in the graphs.

ii) GRU does better beccause it minimizes the loss overall much better than RNN.