Theory Part

[**Q1**] *The following is the AdaGrad algorithm for weight update.*

$$cache_i = cache_i + (\nabla_{w_i} L)^2$$

$$w_i = w_i - \frac{\eta}{\sqrt{cache_i} + \epsilon} \nabla_{w_i} L$$

where w_i is the weight to be updated, $\nabla_{w_i} L$ is the gradient of the loss w.r.t w_i , ϵ is a hyperparemeter between 10^{-8} and 10^{-4} and η is a hyperparameter similar to step size in SGD. List one difference between AdaGrad and SGD in terms of step size and **explain** what effects you expect from this difference.

ANSWER:

SGD use constant step size, which suffers from descending too slow in a flat, non-minumum areas and descending too fast in the steeper area.

$$w_i = w_i - \eta \nabla_{w_i} L$$

AdaGrad attempts to mitigate this by introducing a *cache*; in this case it is the sum of the previous gradients squared. By dividing the gradient descent by the square root of the cache, this means that the gradient will descend faster when the cache is small, and slower when the cache is big.

Small cache means that the gradient has been small for the past iterations, meaning that the model is currently at a "flat" area of the graph, and it should move faster as everything around the area is generally going to be flat, so that the model can move to an area that actually has a minimum. Big cache means that the gradient has been big for the past iterations, implying a "steep" area in which the model should descend more slowly on, in order to not miss the minimum point. To this end, AdaGrad will perform better than SGD in terms of both speed and performance.

[**Q2**] The following are the defining equations for an LSTM cell,

$$\begin{gathered} i_t = \sigma(W_{x_t}^i + U^i h_{t-1}) \\ f_t = \sigma(W_{x_t}^f + U^f h_{t-1}) \\ o_t = \sigma(W_{x_t}^o + U^o h_{t-1}) \\ \hat{c_t} = tanh(W_{x_t}^c + U^c h_{t-1}) \\ c = f_t \circ c_{t-1} + i_t \circ \hat{c_t} \\ h_t = o_t \circ tanh(c_t) \end{gathered}$$

The symbol $\circ$ denotes element-wise multiplication and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. Answer True/False to the following questions and give not more than 2 sentences explanation.

1. If $x_t = 0$ vector then $h_t = h_{t-1}$.
2. If f_t is very small or zero, then the error will not be back-propagated to earlier time steps.
3. The entries of f_t, i_t, o_t are non-negative.
4. f_t, i_t, o_t can be seen as probability distributions, which means that their entries are nonnegative and their entries sum to 1.

ANSWER:

1. False. The cell forward pass is still affected by $U^* \cdot h_{t-1}$ (where $*\epsilon\{i, f, o\}$).
2. False. Backpropagation will still occur according to i_t and o_t.
3. True. As i_t, f_t, o_t is the output of sigmoid functions, it will only lie in (0, 1) as output.
4. False. While the entries are non-negative, they graph does not sum up to 1. As $\sigma(\infty) = 1$, it is impossible for the sum of the graph to be equal to 1.

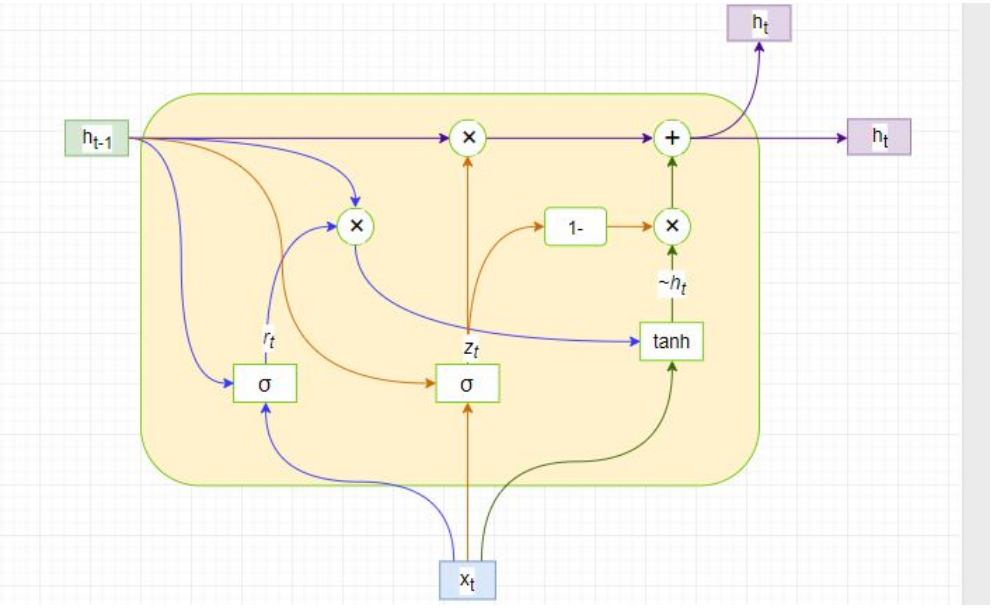
[**Q3**] The defining equations for a GRU cell are,

$$
\begin{aligned}
z_t &= \sigma(W^z x_t + u^z h_{t-1}) \\
r_t &= \sigma(W^r x_t + u^r h_{t-1}) \\
\hat{h}_t &= tanh(W x_t + r_t \circ U h_{t-1}) \\
h_t &= z_t \circ h_{t-1} + (1 - z_t) \circ \hat{h}_t
\end{aligned}
$$

1. Draw a diagram of this GRU cell.
2. Assume h_t and x_t are column vectors, with dimensions d_h and d_x respectively. What are the dimensions (rows × columns) of the weight matrices W^z, W^r, W, U^z, U^r, and U?
3. Like LSTM cells, GRU cells can tackle vanishing or exploding gradient problem too. By taking a look at the formula for LSTM in Q2, what is the main advantage of using GRU cells over LSTMs for some problems? Give an answer it at most 5 sentences. *Hint: We expect a qualitative answer (deep math proofs are not required) that comes with an explanation of the answer.*

ANSWERS

1. GRU Cell diagram is as follows (drawn with draw.io (http://www.draw.io))



2. Corresponding dimensions:

$$
\begin{gathered}
from\ their\ matrix\ multiplications: \\
d_W^z = a \times d_x, \quad d_U^z = a \times d_h, \quad d_z = a \times 1 \\
d_W^r = b \times d_x, \quad d_U^r = b \times d_h, \quad d_r = b \times 1 \\
d_U = c \times d_h
\end{gathered}
$$

$$
\begin{gathered}
from\ their\ elementwise\ multiplications: \\
d_z = d_h \times 1, \quad a = d_h \\
d_r = d_h \times 1, \quad b = d_h \\
c = d_h
\end{gathered}
$$

$$
\begin{gathered}
therefore: \\
d_W^z = d_h \times d_x, \quad d_U^z = d_h \times d_h \\
d_W^r = d_h \times d_x, \quad d_U^r = d_h \times d_h \\
d_U = d_h \times d_h
\end{gathered}
$$

3. The main advantage is that GRU has 2 gates instead of 3, meaning that the parameters required to train in GRU is less than that of LSTM; assuming they perform with the same accuracy, GRU is more computationally efficient.

Coding Part

In [1]:
```
from hw6_code import *

star_filter = ['NEXTEPISODE']
dataset = MovieScriptDataset('../dataset/startrek/star_trek_transcripts_all_episodes_f.csv',
                             filterwords=star_filter)
train_data, test_data = dataset.split_train_test()
print(train_data.line_list[-1], '\n')
print(train_data.line_list[-3], '\n')
print(train_data.line_list[-5], '\n')
print(train_data.line_list[-10], '\n')
print(train_data.line_list[-50], '\n')
```

```
KIRK: What is it?

PIKE: There's no indication of problems down there, but let's not take chances.

VINA: I think it's time to show the Captain our secret.

KIRK: Spock!

SPOCK: Spock to transporter room.
```

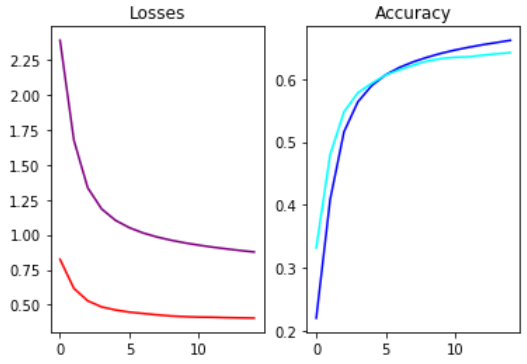
In [2]:
```
lstm_mod = CoveredLSTM(len(charspace), 200, 3, len(charspace)).cuda()
```

In [3]:
```
print("training")
trained_model, train_loss_acc, test_loss_acc = train(train_data, test_data, lstm_mod, resume_from=0, save_model_every=5
,
                                                     learnrate=2.5e-1, batch_size=20, sample_every=5000, epoch=15)
```

```
training
EPOCH 0
      >> Epoch loss 2.38817 accuracy 0.220                    in 169.9061s
      >> Epoch loss 0.82465 accuracy 0.332                    in 36.5304s
      sample line: SCOTT: It's working, sir.
      sample output: SPCC:: I  t toue ng  tote
EPOCH 1
      >> Epoch loss 1.67508 accuracy 0.408                    in 169.8923s
      >> Epoch loss 0.61761 accuracy 0.480                    in 37.3077s
      sample line: KIRK : Clear passageways immediately. The Ambassador will be escorted to his quarters at once.
      sample output: KIRK:: Saatt irrtere  n  tnpitinte.e
Iheycnpiseel r oetl te antett d to tes tuett d .on ander
EPOCH 2
      >> Epoch loss 1.33498 accuracy 0.516                    in 170.1230s
      >> Epoch loss 0.52836 accuracy 0.548                    in 36.6199s
      sample line: ZARABETH: It was not enough that he execute my kinsmen. Zor Kahn determined to destroy our entire fa
mily. He used the atavachron to send us places no one could ever find us.
      sample output: ZANA:ER:: I  ias tot tntugh thet ia inprtsidte cnld.ent
Ieweday  teaar ang. th tettroy tnr pntela tocini
Ie'cne  thercntreyt yu th tee itn arane  toton  oonrd bnen tond tn
EPOCH 3
      >> Epoch loss 1.18458 accuracy 0.564                    in 170.6265s
      >> Epoch loss 0.48539 accuracy 0.578                    in 36.7791s
      sample line: ANAN: We have been at war for five hundred years.
      sample output: ANLN: Ie cave teen tntterpcor tone temdred toars.
EPOCH 4
      >> Epoch loss 1.10320 accuracy 0.590                    in 172.8597s
      >> Epoch loss 0.46330 accuracy 0.593                    in 36.7157s
      sample line: MCCOY: Yes. Yes, in a way it is. The human brain controls the individual's functions.
      sample output: MCCOY: Ies,
Ios, st t cay tn ws
Shercaman teiin tomtrol  aoercntecetetl s tolntion..
EPOCH 5
      >> Epoch loss 1.05065 accuracy 0.606                    in 171.1553s
      >> Epoch loss 0.44813 accuracy 0.606                    in 36.7168s
      sample line: CHEKOV: The new headings will be plotted in a minute, sir.
      sample output: CHEKOV: Iheycex caading  aell be arateid tn tncanute. Cir.
EPOCH 6
      >> Epoch loss 1.01269 accuracy 0.618                    in 169.8611s
      >> Epoch loss 0.43816 accuracy 0.614                    in 36.3688s
      sample line: CHRISTOPHER: I take it that a lady computer is not routine.
      sample output: CHEISTOPHER: I dhle tt toet tncini,oonputer rn aot teoning.
EPOCH 7
      >> Epoch loss 0.98426 accuracy 0.627                    in 172.1203s
      >> Epoch loss 0.42868 accuracy 0.622                    in 40.9652s
      sample line: KIRK: It was not deliberate, I assure you.
      sample output: KIRK: I  was aot teaigerated C wmsure you
EPOCH 8
      >> Epoch loss 0.96139 accuracy 0.634                    in 171.1068s
      >> Epoch loss 0.42056 accuracy 0.628                    in 37.2672s
      sample line: KIRK: to you, friend. Joy and tranquillity.
      sample output: KIRK: Ih aour tooendl
Iul tnd thynsuillisy
EPOCH 9
      >> Epoch loss 0.94235 accuracy 0.641                    in 171.5010s
      >> Epoch loss 0.41566 accuracy 0.632                    in 37.3267s
      sample line: LAZARUS: Both universes, Captain. Yours and mine.
      sample output: LAZARUS: Iunh tniverse.. taptain.
Iou  etnd tand.
EPOCH 10
      >> Epoch loss 0.92621 accuracy 0.646                    in 171.5777s
      >> Epoch loss 0.41269 accuracy 0.634                    in 41.6226s
      sample line: LOSIRA: That is not important. You are Lieutenant Sulu. You were born on the planet Earth. You're he
lmsman for the Enterprise.
      sample output: LAKIRA: Ihet is aot tnportant.
Iou wre aisutenant Cclu.
Tou aire tern tf the clanet.anrth  Iou ve aerpe an.irr the cnterprise.
EPOCH 11
      >> Epoch loss 0.91191 accuracy 0.650                    in 171.2340s
      >> Epoch loss 0.41150 accuracy 0.635                    in 37.1998s
      sample line: MCCOY: How?
      sample output: MCCOY: Iew
EPOCH 12
      >> Epoch loss 0.89950 accuracy 0.654                    in 171.4015s
      >> Epoch loss 0.40863 accuracy 0.638                    in 41.4958s
      sample line: SPOCK: We're being held in place, Captain, apparently from that solar system.
      sample output: SPOCK: Ie re geing deld tn teane. aaptain. bnprrently toom thet thmar dystems
EPOCH 13
      >> Epoch loss 0.88785 accuracy 0.658                    in 172.0734s
      >> Epoch loss 0.40693 accuracy 0.640                    in 42.3164s
      sample line: CLAUDIUS: Prepare food for our friends. They've come from a great distance. A great distance indeed.
```

```
So, this is a Vulcan. Interesting. From what I've heard, I wish I had fifty of you for the arena.
      sample output: CHAUDIUS: Irepare torl.tor aur priends.
Iher re gome from tncreat destrnce  Incaeat destrnce fn eed.
Ihm thes is tnfulcan  I  eresting. Iiom that w me gaard  C walh t cav tonte tf touraor ahe cnea'
EPOCH 14
      >> Epoch loss 0.87731 accuracy 0.661                    in 175.5041s
      >> Epoch loss 0.40540 accuracy 0.642                    in 41.9029s
      sample line: KIRK: You have a talent for understatement, Lieutenant. Without full crystal power, our orbit will b
egin to decay in ten hours. Reamplify immediately.
      sample output: KIRK: Iou mave a frpk t aou tsier tonement  bieutenant
Ielh ut trll poestalsioier  srr pwdit aall be in to sesey tn thn maur..
Iemd leci aspediately
```

In [4]:
```
plot_over_epoch(train_loss_acc, test_loss_acc)
```



purple: train loss

red: test loss

dark blue: train accuracy

light blue: test accuracy

In [5]:
```
# sampling 50 sentences
for i in range(50):
    print('GENERATED:', trained_model.sample(max_length=500))
```

```
GENERATED: ODONA: I would do it.
GENERATED: UHURA: Sensors programmed it. I was arrive and move.
GENERATED: BROWN: I was an indee of the planet surface control.
GENERATED: NOMAD: So the crew work and engineering course.
GENERATED: JARIS: I said I have to be increasing to attend the orcinition. It was an answer. Now. There were an engines in the computer control. For the computer read in the chances. There's no survivors to be a worse.
GENERATED: ELAAN: I wondered the captain. There will be it being we there.
GENERATED: APRIL: What are you doing?
GENERATED: EM3GREEN: I was something to say in the ship.
GENERATED: KIRK: You have a solad has had to be taken on the ship. They were not completely their action of this plane t.
GENERATED: LAZARUS: But the ship is true, Captain.
GENERATED: TAMAR: They were here.
GENERATED: PIKE: They didn't say there's a big blood control of the planet of the basic controls. All centuries?
GENERATED: LEILA: I don't know. He's gone a trained to another computers and the same things to keep one of the same way we can help you. I know what he was something to say they mean. I was the transporter room.
GENERATED: GARTH: At life for anything to be a destroyed by the people of the same transporter room.
GENERATED: HODIN: I was an animal control pants of some mind is permitted. I can't understand this interference with their original condition.
GENERATED: PIKE: Once it is, but we're trying to save the same and officer. I was an instructions are an minor entitle way. I would have an appear to be a few minutes to see.
GENERATED: KIRK: That was an order to be a reason for anything we can keep the ship. Just as it in the right of a thought destroyed the new, but the ship and the same we were going to see the captain.
GENERATED: PIKE: There's no longer crystals are alive.
GENERATED: WALLACE: Then one seruis an order to say sit and death. They can starned them in the same we'll die.
GENERATED: WALLACE: But why do you want to wait to do it?
GENERATED: KIRK: There will be all right. We're going to try to stay here.
GENERATED: UHURA: I am an attack with mine. Detail there and the transporter control report to the computer.
GENERATED: TRELANE: Well, what's the only take him to destroy your ship.
GENERATED: FLINT: The present is commanding cannot answer. What were there?  I am a good and they're still as great dead. Let me position the death, we were say they were to speak of it. Are you sure you can tell you here?
GENERATED: KIRK: Do you realise the computer is reading in this information were another distress call for the transporter room. Computer beams being disturbing anything.
GENERATED: ISAK: There were anything to say in a back six years.
GENERATED: CHEKOV: There's a trick of a reasoning an order to say there in a primitive report to be an action. There is no radio was serious to destroy your crew down there.
GENERATED: FELLINI: They were strong, sir.
GENERATED: ELAAN: We're trying to do. We were trying to get into the creature.
GENERATED: ELEEN: I won't travellate a whole constructions and an ancestors and weak well.
GENERATED: ELAAN: I am ready.
GENERATED: KIRK: Do you know?
GENERATED: DAYSTROM: That did they have any saw the communict we have to leave the word and the ship on Earth answer.
GENERATED: TYREE: No. Captain, the same captain is something wrong.
GENERATED: SPOCK: It is a fight.
GENERATED: WALLACE: They are there to be reading in another mind and control the storage to have any of our own destroying confirmation. The moment that this is the same thing they were men.
GENERATED: VINA: There's an ancient of the transporter room sounds and the sensors field.
GENERATED: BOMA: Here we were traces of the way they can tell them to be an order to be an ancient of an ancient terrior readings. I was coming on an ancestors. I want to know so.
GENERATED: CHEKOV: I am the corridor and remain of an explanation.
GENERATED: DAYSTROM: What's the command is the same answers in the galaxy to protect the galaxy and a repairs.
GENERATED: DARAS: I have to a fact that they were too late.
GENERATED: EVE: You can't find out with you.
GENERATED: HANSON: I am sorry. I know what he was been saved by the same we're the decision. Commodore Spock and you leave the same two and or a moment.
GENERATED: UHURA: Captain, my first officer. I said you were as a little dead. I want to come to me. I want to know what it is, but it is right to see you.
GENERATED: HANSON: Yes, sir.
GENERATED: T'PAU: There is a pattern is the answer.
GENERATED: MCCOY: There is an engineering. She was trying.
GENERATED: DAYSTROM: She's a sense.
GENERATED: T'PAU: There is another dead has considered a bearing officer, there is a balloon computer.
GENERATED: SCOTT: All right. There is an order. I would like to come aboard the Mister Spock.
```