50.039 Theory and Practice of Deep Learning
Coding Homework 7

Joel Huang 1002530

April 8, 2019

1   Formulation

Data wrangling

I manually cleaned the data over 2 hours, using a combination of Python scripting, text editor search and replacement, and manual correction. A sub-sample of the full dataset is shown here:

```
KIRK: Rescan your sensor readings, Mister
Spock.

SPOCK: Readings are still inconclusive. It
 is odd that only the men are affected,
Captain. I suggest

...

UHURA: These figures just don't match up
with Spock's. So far I count three sensor
readings that are off.

KIRK: Fantastic architecture. Only an
incredibly advanced race could have built
it.
```

The final dataset can be found in `clean.txt`. A custom dataset class is also written to support the operations for this dataset, including conversion from strings to tensors and vice-versa, retrieving mappings from labels to characters, dataset splitting with random seeds (we use a random seed of 17), and of course the `__getitem__ ()` and `__len__()` methods.

Model

Following the concept illustrated in Fig. 1, we use a packed padded sequence input into a 2-layer LSTM. Thereafter, the packed sequence returned by the LSTM is unpacked and padded, followed by a single fully-connected layer. The softmax is outsourced to `nn.CrossEntropyLoss`.
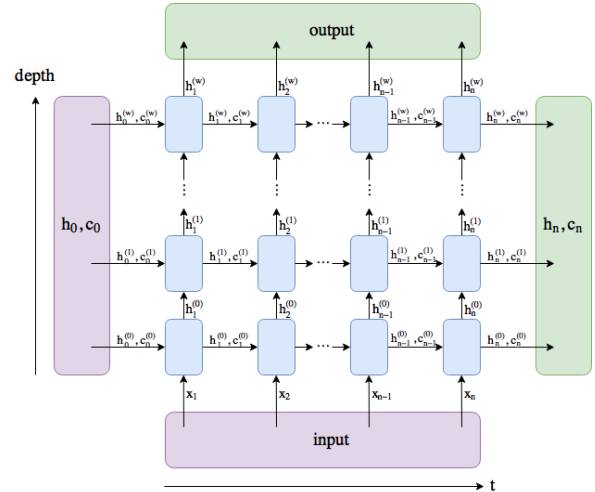
---

[1] Image from: https://stackoverflow.com/questions/48302810/whats-the-difference-between-hidden-and-output-in-pytorch-lstm

Figure 1: Multilayer LSTM[1]

Sampling

We sample with a temperature $T = 0.5$ in all cases. The sampling stops when an end-of-sentence character is encountered. The distribution is given by:

$$P(z) = \frac{e^{z/T}}{\sum_c e^{z/T}}$$

Accuracy computation

We compute the accuracy as follows; for each mini-batch of size N, we have N target tensors. The padded characters in the sequence have a value of 0. We mask this out using `target.nonzero()`, and apply this mask M to both the prediction tensor P and the target tensor T. Any predictions on the padded characters are hence discarded. The accuracy for the n-th sample for each packed sequence index i in the mini-batch is thus:

$$acc(x_n) = \frac{\sum_i (P_{n,i} \circ M_i == T_{n,i} \circ M_i)}{len(M_i)}$$

The implementation is as follows:

```python
def compute_accuracy(output, target):
    """
    Compute accuracy assuming all [PAD]
    characters in
    the target tensor have value 0.
    """
    mask = target.nonzero().squeeze(1)
    _, pred = torch.max(output, dim=1)
    sample_count = len(mask)
    true_positive = torch.sum(pred[mask] ==
    target[mask].long())
    acc = (true_positive.float() /
    sample_count)
    return acc
```

2 Final experiment results

I used a 80:20 split for the training and validation sets. We train with cross-entropy loss using the Adam optimizer. After conducting 5 experiments for hyperparameter selection, the best results are as follows:

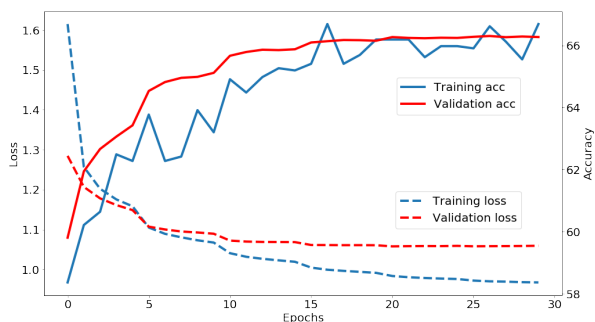| Hyperparameter	Value
Epochs	25
Batch size	16
LSTM Dropout	0.1
Optimizer	Adam
Learning rate	1e-3
Scheduler steps	5
Scheduler gamma	0.5
Loss weights	Ignore [PAD]
Hidden dimensions	200
LSTM layers	2

Table 1: Hyperparameter selection

Figure 2: Loss & accuracies for the final experiment.

Sampling results

Temperature sampling at epoch 1/30

PIKE: The ready, but it would be discussions of a control all the officer of completed the thing of the world to all right in the chance into the power bould of the completed it. We will need to got not a forms to be to better mean to all the alive.
MCCOY: Now we must seen there. The are only through in the Enterprise.
UHURA: I can tood the ship is all of site.
LAZARUS: Yes, sir. We were the scan that we can starm the thing continue, it was in the ship.
ZARABETH: Oh, I might never planet that a minutes of the ship control and whole that was a power and signal are not the ship.
ROBERTA: All right, and the ship. You would be one recert out of the for you were bridge?
JONES: All we are the bridge of the friend to Christopation it.
PIKE: You seem to control the ship and destroyed him.
HANSEN: She must can do every species. The ship is the time to be a problem to make have been all right.
DEHNER: This is a completed of the control of the decks. That has not the speed of themselves.

Temperature sampling at epoch 15/30

SPOCK: I am all right.
COMMANDER: It may be able to have the bridge. The most words, stand by.
PLASUS: The entire starship Enterprise back to Captain Christopher.
HANSEN: Think of all the power is the ship and the doctor.
HANSEN: Status?
LOKAI: I don't know.
MCCOY: What about the descendants are sure to stay in the moment is a fatter beams a court out of the survey of this man.
CHEKOV: We can see the time that I saw some kind of the shields of decision to the bottle contact with you.
DECKER: Well, we're not all right. And I want to know what we beamed down to the same facts. It was still any of the state of the only one.
DEHNER: The power of this has been some signal.

Temperature sampling at epoch 30/30

DEHNER: And now you want to kill you to the planet's reaction. There must be a man who wanted to have the captain to say to see it as a name.
DECKER: The memory capacity to take what they could lose me. The interest will be forced to get her at all.

HANSEN: Sir, we are not a distance. We will
    not be sure they have the ship, and I want
     to know what they are to probe the
    children. They're human in the planet.
ADAMS: She would be an entire species to have
    the problem of the coordinates of the
    tests and the ancient laws are an inert
    ninety.
CHEKOV: Yes, sir. This is the streets. They
    began a representative of the planet
    Commander Boma. That is a men and the
    planet is not a few minutes.
VINA: You should be a design.
WESLEY: I have nothing to see you to be a
    place and who are the remains of the
    planet is the station.
WESLEY: The antimatter descerts of the time
    that the computer is a body. Now my first
    officer Delay, the Captain is stranded. I
    want to know that the creature is aboard,
    Mister Spock. I saw you would be here. The
     food of the ship is a fool, the first
    time they're doing a virus. The new
    starship would be a way to do is back to
    Vulcan.
RILEY: I was the real. The body will continue
    to work. All right, Prefect is one in the
    ship. I can't run a little control.
SPOCK: And you're seeing the transporter room
    and the first time and I are considered. I
    'm afraid I would seem to live on board
    the entire ship. I want to remember the
    matter with the ship.

3   Emmy-winning quotes

ELAAN: I am Spock.
SPOCK: No people are beautiful.
GALLOWAY: The computer is a human woman.
APOLLO: I think you will be more precious than
    the private parts of the ship.