Homework 7

Descriptions & Instructions:

This homework is very different from earlier assignments. It is more open-ended, longer and worth double the points. The goal of this homework assignment is to test your ability to apply the knowledge you have learned from class and your knowledge of PyTorch to new data. In this assignment, you will choose your data source (or choose one of the ones we suggest), visualize the data and collect statistics about the data, read some background papers related to your data, construct a PyTorch model for your data, train your model and conduct experiments.

This homework also gives you the opportunity to go above and beyond to earn extra credit. The point total for this homework adds up to over 100, with the possibility of earning extra points on each section for outstanding work. Examples would be things like: extra work in collecting your own unique data rather than using one from the Internet, a very thorough literature review, particularly extensive experiments, etc.

As with the last assignment, the models may take much longer to train, so please start the assignment early so that you have sufficient time! You have extra time on this assignment but you also are expected to do more than on a normal homework.

To be able to run larger networks, you might also want to look into Colab Pro: https://colab.research.google.com/signup. You should be able to get a free trial as a student by clicking "No cost for students and educators" and signing in with your U login.

Again, please ensure that you have Python as well as PyTorch installed on the computer where you work on the homework. PyTorch can be installed using this link.

You are provided with the Jupyter Notebook, **HW7.ipynb**. Unlike previous assignments, we do not provide much skeleton code. We encourage you to look at code from your last few assignments or find code online (https://github.com/pytorch/examples has several great examples) **as long as you do not copy code or use AI to generate code.** Again, please ensure that the entire code in the notebook executes without any errors.

Please carefully follow the instructions in this document and perform the steps in the colab

notebook. **Submit format:**

As always, you will place your Jupyter Notebook source code and a report (explained in the following) in a folder, compress it into a **zip file** called HW7.zip, and submit it on **Canvas**.

You need to submit a report including your solutions to the coding problems. *You can choose to convert your Jupyter Notebook to a pdf report or take screenshots of your code and results*. **Please submit both,**

the PDF and the .ipynb file on Canvas!
Please mark your solutions to each question correctly while submitting the report on Canvas.
Failure to follow the instructions will lead to a deduction in points!

Part 1: Dataset [20/35 pts]

Task: Collect and analyze your dataset.

In the previous assignments we have provided you with the code to load datasets. For this assignment you will choose one yourself and figure out how to load it.

Keep in mind that the dataset you choose here will then affect later parts of the assignment, so choose carefully. You may also want to consider the size and difficulty of the dataset. It should be hard enough to make training a model interesting but not so difficult that it is hard to train small models that you can easily train in colab.

Some idea of places you can choose your dataset:
https://www.kaggle.com/datasets
https://medium.com/information-expositions/list-of-lists-of-datasets-c9bf5237075
5 https://huggingface.co/datasets/
https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021

If you are not sure what to do, you might choose a simple image classification dataset CIFAR-100: https://huggingface.co/datasets/uoft-cs/cifar100 to explore image classification pipelines. Or the Visual Question answering dataset if you're interesting in vision and language problems: https://huggingface.co/datasets/HuggingFaceM4/VQAv2. Or if you want to explore NLP, you might try a sentiment analysis dataset: https://huggingface.co/datasets/Sp1786/multiclass-sentiment-analysis-dataset.

Once you have chosen a dataset, in your notebook load the dataset. If you need to download the dataset or its data you have collected, make sure to include the dataset in your submission along with the instructions of how to load the dataset in colab.

You may also want to use a dataloader for your dataset rather than put the entire dataset into colab at once as we did previously with smaller datasets. Use any method you like, but you may want to look into PyTorch dataset loaders:
https://docs.pytorch.org/tutorials/beginner/basics/data_tutorial.html

If the dataset does not already have defined train, validation and test splits, you should also do this.

Next, you should do some dataset analysis and statistics on your data. This should definitely include the basic statistics of your dataset:

- A description of the dataset
- Format of the inputs and outputs
- Number of examples in the train/val/test splits

- A visualization of at least 5 examples from your dataset
- Summary statistics (e.g. for language datasets, you might include things such as the average lengths of the inputs or the number of unique words).
- If possible, any visualizations of the dataset statistics.

For more ideas, see how published dataset papers show statistics: https://arxiv.org/pdf/1612.00837

To get base level points for this part of the assignment (30) we expect you to:

- Choose a pre-gathered dataset such as one of the ones in Kaggle or somewhere else online or one of the suggested datasets.
- Perform adequate analysis and statistics gathering on the dataset
- Implement the data loading in colab
- Use or implement a train/validation/test split on your dataset.

To get up to 15 extra points on this part of the assignment you should do one or more of the following: ● Gather or scrape some dataset that is not already easily accessible. This may be some data you collect for a project that is interesting to you personally. Please describe in detail how you collected this data.

- Provide outstanding data analysis on your dataset

Part 2: Literature Review [20/30 pts]

Task: Find papers which train models on your dataset / type of data.

Your next step is to do background research on your dataset or on the type of data you are using. For datasets which already have published papers (e.g. https://arxiv.org/pdf/1612.00837) read those as well as related methods papers using the dataset. For instance, for VQA you might read a hierarchical attention paper like https://arxiv.org/abs/1606.00061. Or for images you might read the original vision transformer paper (https://arxiv.org/abs/2010.11929).

If the dataset is not directly used in published research, look at papers on similar kinds of data (e.g. if you have audio data, look at some deep learning papers on audio). We suggest looking at either published conference papers in venues like neurips and/or papers with more citations to filter for more impactful papers. You may also want to look for older papers to find papers which use methods that are more manageable to train in colab than newer papers. Google Scholar is a really good way to search for papers and will show you the number of citations and conference publications.

Choose at least 2 papers, but more is better.

Then, in your notebook or in your pdf report, summarize the methodology of the paper and in particular the types of deep learning models they used on the data. For instance, for a vision dataset, you might find papers which use CNNs or ViTs to do the classification.

To get base level points for this part of the assignment (20) you should:

- Read at least two papers related to the type of data you chose and describe the types of neural networks they use.

To get extra points (10) you should:

- Read more papers and provide a more thorough discussion of the types of neural networks used.
- Provide insight/trends learned from a larger review of the literature.

Part 3: PyTorch Model Implementation [35/40 pts]

Task: Implement a PyTorch model and then train and evaluate the model on your dataset.

Now that you have your dataset and you have read a few papers, you should now have some idea of what kinds of neural networks can be used on your dataset.

Choose a kind of neural network and implement it in PyTorch.

Hint: the choice here may be crucial for choosing a network that is viable for you data. Try starting with a smaller/simpler version of one of the models used in papers from your literature review. Or choose one of the networks we talked about in class. If you find that it takes more than 30 minutes on colab, try reducing the number of parameters or take a smaller subset of your dataset. You will also want to plot your train/val curves constantly during training rather than at the end so that you can quickly see and debug your network. If training is taking too long, try removing layers or choosing simpler networks.

Make sure that you do not *exactly* copy the implementation from previous homeworks. So for instance, don't just copy your CNN or LSTM code from the previous homework. Even if you end up having convolutions in your model, make sure to change up the architecture somehow by adding different kinds of layers or somehow changing the model.

In your colab notebook, implement a neural network architecture in PyTorch, train the model on your dataset and plot your train and validation accuracy over time during training. Feel free to re use code from previous assignments.

To get full base level points for this part of the assignment (35) you should:

- Have a correctly implemented PyTorch neural network module
- Have correct training loop and show the training and validation error curves plotted (see previous assignments)
- We do not expect anywhere close to state-of-the-art results, just that you correctly implemented a network and can show that your validation accuracy improves during training.

You can also get up to 5 extra points for a particularly creative model.

Part 4: Experimentation and Analysis [25/45 pts]

Task: Perform experiments with your dataset and model.

Now with a correct PyTorch implementation and training procedure, we want you to do experiments with your model or your data.

In your colab notebook perform experiments on your network. At a minimum, you should perform a grid search on different hyperparameters of your network (see the previous assignment). As before, use your validation set, not your test set.

You should then perform other experiments to try to improve your validation accuracy or to analyze your model in some way. Some ideas could include but are not limited to: ● Data augmentation

- Major architectural changes to your base network
- Pretraining on other dataset or using alternative auxiliary training method ● Generating new features for you dataset (e.g. use the output of another model or use some hand-defined features you come up with)
  - Do some visualization of your model (e.g. GradCam https://arxiv.org/abs/1610.02391)

For all of these experiments, plot the validation accuracy of your experiment against your baseline accuracy from Part 3. For the grid search, plot the val accuracy over all of your parameter choices as you did in the last homework assignment.

After all of these experiments, take whatever the best performing set of hyperparameters and experiments is and compute your final test accuracy for your dataset.

To get base level points for this part of the assignment (25) you should:

- Perform at least 2 grid searches over different hyperparameters including learning parameters (e.g. learning rate) or architectural hyperparameters (number of layers, hidden layer size etc).

To get extra points (20) you should:

- Do some experiments beyond just hyperparameter search. See the list above for some examples, or look through some of the papers in the literature for ideas.
- Here we are looking for extra effort and creativity in your experiments.