

Homework 4

*Due: 5:59pm ET Thursday October 9**This problem can be done in teams of up to 2 students.*

1 Window Type Classification [20 points]

In this problem, the goal is to classify window images into one of five categories: “New Awning Window”, “New Bay Window”, “New Fixed Window”, “New Horizontal Sliding Window”, and “New Hung Window”. The training and test datasets can be accessed via the following links:

- https://canvas.wpi.edu/files/7719816/download?download_frd=1 and
- https://canvas.wpi.edu/files/7719811/download?download_frd=1.

To help you get started, a demo code are available at:
<https://colab.research.google.com/drive/1fG0f6LiPnv7a4nDWNptVyo1Y0Xh5vp7l?usp=sharing>.

2 Comparing Vanilla RNN with Variants in Sequence Modeling [20 points]

You will implement and train three different neural networks for sequence modeling: a Vanilla RNN (a simple RNN with shared weights), and two variants of a NN with a similar architecture to the Vanilla RNN but which do not share weights. You will compare their performance on a sequence prediction task and analyze the differences between them.

Prediction task: this is a many-to-one regression task, i.e., a sequence of inputs is used for predicting a single output. In particular, the i -th input sequence $\mathcal{X}^{(i)} = (\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_{\ell_i}^{(i)})$ has length ℓ_i , where $\mathbf{x}_j^{(i)} \in \mathbb{R}^{10}$; the i -th output is a scalar $y^{(i)} \in \mathbb{R}$.

Dataset: You will use a synthetic dataset containing sequences of variable lengths stored in the zip file `homework5_question2_data.zip`. Each sequence consists of input features and corresponding target values. The sequences are generated such that they represent a time-dependent process. Note that ℓ_i may be different than ℓ_j for $i \neq j$. So the (pickled) numpy object \mathbf{X} is actually a list of sequences.

Tasks:

1. (4 points) Implement a Vanilla RNN: Implement a Vanilla RNN architecture (needless to say, weights are shared across time steps). A pytorch starter code is provided in `homework4_starter.py`. **Important: You are not allowed to use an RNN layer implementation from any library.**
2. (4 points) Implement a NN with Sequences Truncated to the Same Length: Implement a NN where sequences are truncated to have the same length before training. In other words, if the shortest sequence in the dataset has length L , all sequences should be truncated to length L before training.
3. (4 points) Implement a NN with Sequences Padded to the Same Length: Implement another variant of NN where sequences are padded to have the same length before training. Use appropriate padding techniques to ensure that all sequences have the same length, and implement a mechanism to ignore the padding when computing loss and predictions.

4. Train and Compare the Models:

- (a) (1 point) Train all three models (Vanilla RNN, Truncated NN, Padded NN) on the provided dataset.
- (b) (1 point) Use a suitable loss function for sequence prediction tasks, such as mean squared error (MSE) or cross-entropy.
- (c) (1 point) Train each model for a fixed number of epochs or until convergence.
- (d) (1 point) Monitor and record performance metrics, such as training loss, on a validation set during training.

5. Evaluate and Compare the Models:

- (a) (1 point) Evaluate the trained models on a separate test dataset.
- (b) (2 point) Compare the performance of the three models in terms of MSE, convergence speed, and overfitting tendencies.
- (c) (1 point) Analyze the results and discuss the advantages and disadvantages of each approach in terms of modeling sequences with varying lengths.

Additional Information:

You can choose the specific hyperparameters for your models, such as the number of hidden units, learning rate, batch size, and sequence length. Feel free to use any deep learning framework or library you are comfortable with, and provide clear code documentation. Note: Be sure to clearly explain your implementation, provide code comments, and present your results in a well-organized manner in the report.

3 Fine-tune a DistilBERT model [20 points + 2 bonus points]

In this project, you will first train a classification head using a pre-trained DistilBERT model on a dataset of social media tweets to classify tweets as containing medical information or not. You are provided with a dataset of social media tweets, where each tweet is labeled as either containing medical information (class 1) or not containing medical information (class 0).

The preprocessing of the dataset, by tokenizing the tweets and converting them into a format suitable for DistilBERT, is already provided in the starter code:

https://colab.research.google.com/drive/17syAcTav5Wtq-n_Rs3P1cQl0szIjLlQS?usp=sharing

Useful documentation for this question can be found here:

<https://huggingface.co/docs/transformers/index>

<https://huggingface.co/docs/transformers/training>

https://huggingface.co/docs/transformers/tasks/sequence_classification

You will need to make the following changes to the existing code:

Q1: (2 points) Add 4+ relevant arguments to the parser.

Hint: Check how `args` is used within `load_pretrained_and_finetune` and think about which other arguments should be added.

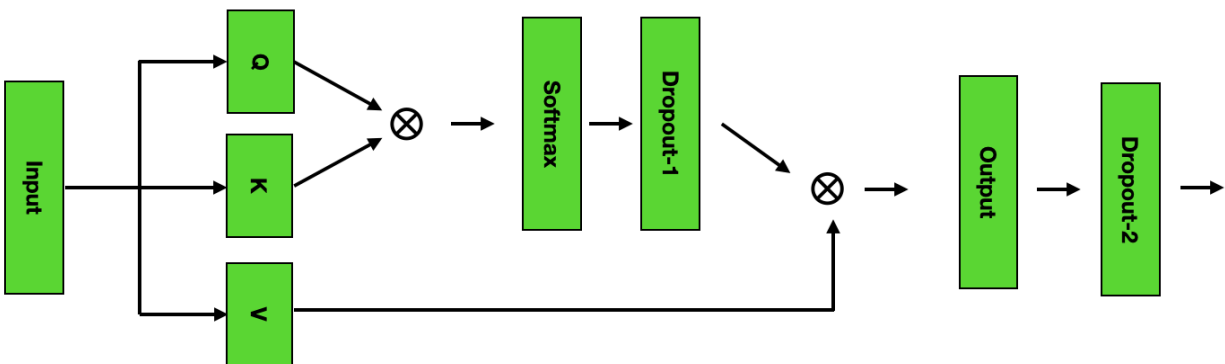
Note: while `argparse` is designed to read arguments from the command line, it is currently adapted to work with a jupyter notebook by passing arguments to `parser.parse_args` as a list.

Q2: (2 points) Split the code in train, val and eval (test) sets stratified by classes.

Q3: (2 points) Convert sets to HuggingFace Dataset and tokenize using function `tokenize_batch`.

- Q4: (6 points) Implement grid search for at least one hyperparameter by training a classification head for a pre-trained DistilBERT model on the training set and evaluate its performance on the validation set. You can use `DistilBertForSequenceClassification.from_pretrained` from the `transformers` library to load the pre-trained model. Note: DO NOT train the entire model, only the classification head. You can initially freeze all parameters except the classification head by setting `requires_grad=False` for all parameters in the base model.
- Q5: (6 points) Run the final training on train+val with best hyperparameters.
- Q6: (2 points) Based on the item above, discuss the performance of the model, any challenges faced during fine-tuning, and potential improvements that can be made to further improve accuracy.
- Q7: (BONUS: 2 points) Apply a principled change to your code in order to achieve F1-macro > 0.50. Explain what you did and why you did it.

4 NOT PART OF HW4: Tensor Shapes in a Transformer Layer [0 points]



The figure above shows the transformer layer. The input size of the transformer layer is $[10, 90, 20]$ (where 10 represents batch size, 90 represents sequence length, and 20 represents hidden size). We consider 5 attention heads in this attention layer. The shape of the (combined) projection matrices is $H \times H$, which are used to project the input data to Q (query), K (key), V (value). Please compute the size of the each output, including:

1. The shape of Q
2. The shape of K
3. The shape of V
4. The shape of Q for each head
5. The shape of K for each head
6. The shape of V for each head
7. The shape of the attention map (output of softmax)

8. The shape of Dropout-1's output
9. The shape of Output
10. The shape of Dropout-2's output
11. Total number of the parameters in this transformer layer

For this homework, you are encouraged to experiment with various models and training strategies. To achieve a full score, your model must achieve at least 60% accuracy on the test set.

Submission

Submit one PDF file that includes your notes for the theoretical problems (scanned or typed) and screenshots of your code for the programming problems. All material in the submitted PDF must be presented in a clear and readable format.