Introduction to Computational Linguistics
Exploring PyTorch for NLP

Reference to this Colab PyTorch tutorial.ipynb

Tokenizer Class

1. What is the purpose of the `word2idx` and `idx2word` dictionaries in the `Tokenizer` class, and how are they initialized?
2. What does the `pad_or_truncate` method do, and why is it necessary when working with neural networks?
3. How does the `tokenize_dataset` method handle unknown words that are not seen during training?
4. What happens if a sentence contains only words that were not seen in the training set? How does the tokenizer handle it?
5. If `seq_length = 10` and an input sentence contains 15 tokens, how will the tokenizer modify the sentence? What if it contains only 5 tokens?

TextDataset Class

1. What is the purpose of the `TextDataset` class, and how does it relate to PyTorch's `Dataset` class?
2. What is the expected shape and data type of the tensors returned by `__getitem__`?
3. Why does `__getitem__` return `lengths[idx]`, and how might this be useful when designing a neural network for text classification?

Deep Averaging Network (DAN)

1. What is the purpose of the `nn.Embedding` layer in this model, and why is `padding_idx=0` specified?
2. How does dividing by `lengths.unsqueeze(1).float()` create an average embedding for each sequence? What would happen if this step were omitted?
3. Why is `Softmax(dim=1)` applied at the final layer, and what does `dim=1` refer to?
4. What happens if `lengths` contains zeros? Why might this cause issues in the computation of `avg_embeds`, and how could you fix it?

Dataset Preparation

1. What is the difference between `tokenize_training_set` and `tokenize_dataset`, and why is `tokenize_training_set` used for the training set while `tokenize_dataset` is used for validation and test sets?

2. What is the role of the `label_dict`, and why are labels converted into integers instead of keeping them as strings?

3. Why do we use dictionary lookups (`label_dict[label]`) when creating `Y_train`, `Y_dev`, and `Y_test`? What would happen if a label were missing from `label_dict`?

4. If the dataset had an additional label, say 'Spam', how would you modify the code to include it?

Training Loop

1. What is the purpose of defining `VOCAB_SIZE`, `EMBED_DIM`, `HIDDEN_DIM`, and `OUTPUT_DIM` before instantiating the model?

2. Why do we use `DataLoader(train_dataset, batch_size=BATCH_SIZE, shuffle=True)` instead of passing the dataset directly to the training loop?

3. What is the purpose of calling `model.train()` before training and `model.eval()` before evaluation?

4. What happens when `optimizer.zero_grad()` is called before computing the loss and performing backpropagation?

5. What would happen if `shuffle=True` were omitted from `train_dataloader`? Would this affect the training process?

6. Why is `torch.no_grad()` used during evaluation, and what would happen if it were removed?