Project Title:

Fine-tuning a BERT model for text classification.

Dataset Description:

The Amazon reviews dataset consists of reviews from amazon. The data span a period of 18 years, including ~35 million reviews up to March 2013. Reviews include product and user information, ratings, and a plaintext review. It provides a set of 3.6 M reviews for training, and 400K for testing. For this assignment, only 1000 rows, drawn from the training and testing subset, were chosen.

- Example:

  An example from the AmazonPolarity test set looks as follows:

  {

  'title':'Great CD',

  'content':"My lovely Pat has one of the GREAT voices of her generation. I have listened to this CD for YEARS and I still LOVE IT. When I'm in a good mood it makes me feel better. A bad mood just evaporates like sugar in the rain. This CD just oozes LIFE. Vocals are jusat STUUNNING and lyrics just kill. One of life's hidden gems. This is a desert isle CD in my book. Why she never made it big is just beyond me. Everytime I play this, no matter black, white, young, old, male, female EVERYBODY says one thing ""Who was that singing ?""",

  'label':1

  }

- Data Fields
  - 'title': a string containing the title of the review
  - 'content': a string containing the body of the document
  - 'label': either 1 (positive) or 0 (negative) rating.
- Size:
  - Size of Downloaded Dataset Files: 688 MB
    - This is the size of the dataset files before any modification or generation.

A brief description of the BERT model I chose to use:

Model Name: distilbert-base-uncased

- Model Type: DistilBERT is a type of transformer-based model for natural language processing (NLP). It is a smaller and more efficient version of the original BERT (Bidirectional Encoder Representations from Transformers) model. DistilBERT retains much of BERT's performance while using fewer parameters, making it faster and more lightweight.

- Base Model: The term "base" in the name refers to the size of the model. In transformer-based models, the size of the model is often categorized into different levels such as "base," "large," or "x-large." A "base" model is usually a mid-sized version, and larger versions would have more parameters.

- Uncased: The "uncased" part indicates that the model was trained on uncased text. In English, this means that all the letters in the input text have been converted to lowercase. This is done to simplify the training process and reduce the model's computational requirements.

- Pretraining: The model has been pretrained on a large corpus of text data to learn contextualized representations of words and sentences.

- Tokenization: It uses sub word tokenization to represent words as smaller units, allowing for a more flexible and efficient encoding of text.

A brief description of network and training setting

Network Architecture:

- Input Layers:
  - token_ids (InputLayer): Accepts sequences of token IDs with a shape of (None, 512).
  - attention_masks (InputLayer): Accepts attention masks with a shape of (None, 512).

- BERT Model (DistilBERT):
  - The pre-trained DistilBERT model is loaded and set as non-trainable.
  - It takes the token_ids and attention_masks as inputs.

- The output from the BERT model is a tuple, and the relevant part (bert_output[0][:,0]) is sliced to obtain the representation of the [CLS] token.

- Dense Layers:
  - dense_layer: A Dense layer with 64 units and ReLU activation is applied to the [CLS] token representation.
  - output: The final Dense layer with 2 units and softmax activation for binary classification (assuming the task is binary).

Model Summary:

- The model has a total of 66,412,226 parameters, divided into trainable (49,346) and non-trainable (66,362,880) categories. Most parameters come from the non-trainable DistilBERT base model.
- Non-trainable parameters come from the loaded DistilBERT model.
- Trainable parameters include weights in the dense layers.

Training Settings:

- Optimizer:
  - Adam optimizer is used with default parameters.

- Loss Function:
  - Binary Cross entropy is chosen as the loss function, suitable for binary classification tasks.

- Metrics:
  - Accuracy is monitored as a metric during training.

- Training Data:
  - Training is performed on the first 1000 examples from the Amazon Polarity dataset.

- Batch Size:
  - Training is done with a batch size of 25.

- Number of Epochs:
  - The model is trained for 5 epochs.

Text Classification with DistilBERT:
A Report on Amazon Polarity Dataset Analysis

1. Objective

The objective of this project is to perform text classification using the DistilBERT model on the Amazon Polarity dataset. The dataset consists of Amazon reviews labeled with polarities (positive or negative). This report details the steps taken to prepare the data, define the neural network architecture, and train the model.

2. Dataset Loading

The "amazon_polarity" dataset is loaded using the Hugging Face datasets library. This dataset contains text reviews along with corresponding polarity labels.

3. Tokenization and Data Preparation

The data is tokenized using a transformer-based tokenizer, and input features and labels are prepared. In this example, the dataset is limited to the first 1000 examples.

4. BERT Model Definition

A DistilBERT model (distilbert-base-uncased) is loaded and set to non-trainable.

5. Model Architecture

The neural network architecture is defined, with BERT embeddings as the base, followed by a dense layer with ReLU activation and an output layer with softmax activation for binary classification.

6. Model Compilation and Training

The model is compiled using the Adam optimizer, binary cross-entropy loss, and accuracy as the metric. Training is performed with a batch size of 25 and for 5 epochs.

7. Model Evaluation

After training, the model will be evaluated on a test dataset to assess its performance on unseen data.

Result of Task 1 and comments:

The reported accuracy gives an indication of how well the model generalizes to unseen data. An accuracy of approximately 82.6% suggests that the model performs well on the test set.

There are two Recommendations:

- It's good practice to also look at other metrics (e.g., precision, recall, F1-score) to get a more comprehensive understanding of model performance.
- Consider evaluating the model on a larger portion of the test dataset for a more representative assessment.

How was accuracy calculated?

The model architecture consists of a pre-trained DistilBERT layer, which is frozen and not trainable. The input layer includes token IDs and attention masks with a sequence length of 512 tokens. The output from DistilBERT is further processed through a dense layer with a ReLU activation function. The final layer is a dense layer with a softmax activation for binary classification.

The model was compiled using the Adam optimizer and binary crossentropy as the loss function. The accuracy metric was chosen as the primary evaluation metric during training.

The model was trained on the training data for 5 epochs with a batch size of 25. Training involved optimizing the model's weights based on the provided tokenized input data and corresponding target labels.

To assess the model's performance, it was evaluated on a subset of the test data. The test data was tokenized using the same process as the training data. The accuracy metric, representing the proportion of correct predictions over the total predictions, was used to evaluate the model's effectiveness on unseen data.

The 3 observations from Task 2 (include the correct and incorrect examples and their predictions):

Observations on correctly predicted examples:
1. Predicted: 1, True Label: 1, Content: My son is very happy to have this game! It was used but the disc had no singe scratch on it! Very awesome game! 2. Predicted: 0, True Label: 0, Content: This book has little to teach that's not for the absolute beginner. To boot, the only time it gets into detail is when it's dealing with Cubase specific programming. 3. Predicted: 0, True Label: 0, Content: Today, February 26, 2009, is the day I cut my losses and buy a new player. After fighting with the Samsung for almost two years, I've had enough. During that time, there was never a moment where all the discs in my collection played properly...or, in some cases, at all. The first firmware update helped, but subsequent attempts seemed to do nothing. AND even if they

had, about 50% of the audio CD's that I try to play on it skip. They always have. And I suspect that no firmware update is going to address that. The most maddening thing about this is that, being an "early adopter" of the format, I paid a fortune for this thing. It's going to be a lonnnnggg time before I consider buying another Samsung product.

4. Predicted: 0, True Label: 0, Content: This book has lots of scholarship & award information but it wasn't useful. It did not give enough information. I needed to consult other sources. Also, the book was faster to use than the CD. Skip this one, I didn't find it helpful at all.
5. Predicted: 0, True Label: 0, Content: if the record companies want to try to hault technology, don't support them. there is no reason why this should not be able to be imported onto an ipod. ipod and other mp3 players are the future (and present) of music, and if the foo's new cd can't keep up with it leave it behind.
6. Predicted: 0, True Label: 0, Content: I honestly do not know what Topo US is good for. I have hunted for hiking trails but haven't found any in my area... or any area for that matter. I have found pipelines and powerlines... whooptie do...Also, I use my Garmin for road navigation... I assumed that TOPO would overlay street maps... WRONG... There are like a million little TOPO maps and I have to go into map setup and individually enable/disable the maps... a major pain when scrolling through 50-200 maps. The TOPO has streets but then are WAY out of date. like 20 years out of date. and you HAVE to look at the old roads when looking at TOPO data.Some friends had me do some searches on underwater wrecks for deep sea diving but they didn't come up... I honestly don't know what this software is good for.2 Stars instead of one because there ARE topolines and mountian peaks.
7. Predicted: 1, True Label: 1, Content: El documental muestra una realidad vivida por hombres y mujeres durante la II Guerra Mundial y el impacto que esta dejo en ellos y en las generaciones siguientes. Lo mejor del documental es que es imparcial y no favorece a ninguna doctrina politica y muestra la trajedia de ambas partes, tanto la trajedia que las personas de los paises del Eje como la de los paises Aliados vivieron durante la II Guerra Mundial. Desgraciadamente el documental no lo venden con sub titulos en español.
8. Predicted: 0, True Label: 0, Content: I have researched this device, and it appears to NOT include the battery, or charger, even though the picture includes a 12v battery.So, buyer beware: unless you already have a compatible system, this is an incomplete product. FYI!
9. Predicted: 0, True Label: 0, Content: This book has absolutely nothing to offer, nor does the included CD. One might expect that a companion CD could provide helpful guidance and teach a beginning drummer what to listen for when tuning a drum - it doesn't. If you already know what a snare drum, bass drum and tom

sound like, don't even bother to unwrap the CD! If I could have rated this book with no stars, I would have! Please don't waste your money.

10. Predicted: 1, True Label: 1, Content: THIS BOOK IS FANTASTIC....I LOVE TO DECORATE WITH MINI QUILTS AND YOU WILL TOO ONCE YOU GET THIS BOOK. EASY INSTRUCIONS AND THE QUILTS CAN BE FINISHED IN LESS THAN A DAY..HAPPY QUILTING!!

Observations on incorrectly predicted examples:

1. Predicted: 0, True Label: 1, Content: The same model has been selling in Korean websites at almost half of the price here. Considering logistics cost and power volt difference, I decided just to buy this cute frog on this site.Pros:Its water tank can hold 1 galon, which is pretty large enough quntity for any room.Its cute design can be friendly to anybody and a good interior accessory.The price is affordable.It performs relatively well.Cons:The ejection holes should be redesigned to prevent any vapors just from falling down to the ground. It need a more adjustable ejection part.The tank holder is week and poorly finished.The LED power light is so bright that bothers night sleep.
2. Predicted: 1, True Label: 0, Content: Originally filmed in 16mm for German television, RIO DAS MORTES is Rainer Werner Fassbinder's darkly comic tale about Mike and Gunther, two aimless men who concoct a half-baked plan to go to South America to hunt for treasure.The guys never quite make it to the treasure site, but they have one hell of an adventure negotiating around the objections of Mike's girlfriend Hanna.The story idea is credited to director Volkor Schlondorf (The Tin Drum).This perverse character study was Fassbinder's eighth film. He went on to make 33 more and established himself as a stylized storyteller with a savage eye for human foibles and follies.Not a great film, but interesting if your writing a term paper on the evolution of the Fassbinder ouevre.
3. Predicted: 0, True Label: 1, Content: They work well when they are not messed with. They do the job that they are supposed to do. I prefer the 7 day med holder, but this one does the job if you do not take a lot of medications.
4. Predicted: 0, True Label: 1, Content: All I knew of the Zombies was "Time of the Season" and Rod Argent.....yeah you know "Hold your head Up"....great song. And then there was "Time of the Season" which I knew of as an original. Yet covered by Brent Bougoies. So, needless to say very uneventful, or at the very least a cute one hit wonder.THIS IS TRULY ONE OF THE GREAT LOST POP/ROCK ALBUMS OF ALL TIME. DO NOT MISS THIS. REMEMBER, ROCK GROUPS HAD TO RELEASE AT LEAST ONE ALBUM A YEAR IF NOT TWO TO REMAIN "RELEVANT" TO THE GODS AT THE EXECUTIVE TABLE OF MUSIC. THIS IS NOT THE ROLLING STONES OR THE BEATLES IT IS A COMBINATION. THIS IS, IN MY OPTINION ONE OF THE GREATEST ALBUMS EVER RELEASED IN THIS DECADE....AND THOSE TO COME!!!!!!!!!!!!!!!!!!

5. Predicted: 0, True Label: 1, Content: Wifey did some research and these seemed to be priced best and were most functional. They come with labels as well but we are using them just as dividers for sorting purposes.
6. Predicted: 0, True Label: 1, Content: This is a fantastic video with the best in the business sharing the stage. I would have given it five stars but the guitar jam at the end was cut off by the rolling credits!!! That absolutely killed me. Imagine Stevie Ray, Eric Clapton, BB King, Albert King, Dr. John, and Paul Butterfield all trading licks together... and we don't get to see it. You'll be sorry if you don't buy this video, but you'll be disapointed at the end.
7. Predicted: 0, True Label: 1, Content: Considering I bought this from Amazon for 29.99 I couldn't be happier. I would, however, be upset if I paid more money. For one, this unit really needs a recharging cradle instead of having to stick in the dinky cord end each time I want to recharge the unit (which is going to be ALWAYS). Second, the bristles on the unit are too short and a little too firm, not really making it ideal for hardwood floors but better for tile. Finally, it's a little on the heavy side which makes it a bit awkward to use in tight spots. Overall though, for 29.99 you can't go wrong. No way would I pay full price, however, with all these deficiences.
8. Predicted: 0, True Label: 1, Content: This was literally the only paper I could find in this size !!!Luckily is does a decent job. Fills a nice niche as 11x14.worried they wont produce it much longer though.
9. Predicted: 0, True Label: 1, Content: I bought this combo pack, planning to show part of the movie to my French classes. (I teach high-school French.) Unfortunatley, the French-language version is only available on the Blu-ray disc and not on the DVD. Naturally, my school does not have any Blu-ray players, so I won't be able to show this movie to my classes. What a shame, as the French version is quite nice.
10. Predicted: 0, True Label: 1, Content: I was really sad. my favorite PC game,petz 4 wasn't working rightIt was broke.when i visted my cousin in New York he wanted to cheer me up. he showed me Zoombinis, a cool new PC game and I LOVED it! ZOOMBINIS RULE!!!!!!!p.s. i got my own Zoombini game!

Observations on Correctly Predicted Examples:

- Positive Sentiment Recognition:
  - The model correctly predicted positive sentiments such as happiness about a game and appreciation for a book.
- Negative Sentiment Recognition:
  - The model accurately identified negative sentiments related to product dissatisfaction, e.g., a faulty DVD player, an unhelpful book, and software that didn't meet expectations.

- Multilingual Understanding:
  - The model demonstrated the ability to correctly predict sentiments in multiple languages, such as Spanish, indicating a degree of language-agnostic sentiment analysis.
- Context Understanding:
  - The model correctly handled nuanced content, such as recognizing the imparitality of a documentary and understanding the context of a negative review about a product's functionality.

Observations on Incorrectly Predicted Examples:

- Price-Related Sentiment:
  - Some examples indicate that the model struggled with sentiments related to product pricing. It incorrectly predicted positive sentiment for an affordable product and negative sentiment for a product mentioned as almost half the price on Korean websites.
- Movie Genre Recognition:
  - The model seemed to struggle with recognizing the genre of a movie. For instance, it incorrectly predicted a negative sentiment for a darkly comic tale, possibly misinterpreting the user's tone.
- Product Feature Recognition:
  - There were instances where the model misclassified sentiments related to product features. For example, it incorrectly predicted a negative sentiment for a product review that pointed out specific design flaws and suggested improvements.
- Inconsistencies in Sentiment Recognition:
  - There were cases where the model incorrectly predicted positive sentiment for reviews that expressed disappointment or dissatisfaction with the product, indicating potential challenges in understanding nuanced sentiments.

These observations suggest that while the model performs well on straightforward positive and negative sentiments, it may face challenges with nuanced language, context understanding, and certain types of product-related sentiments.

The 5 examples of Task 3 and the results along with comments:

Pair 1:

Sentence 1: "The boy is chasing a ball."

Sentence 2: "The girl is napping in the sunlight."

Cosine Similarity Score (boy, girl): 0.7853

- Comment: The high cosine similarity score between the words "boy" and "girl" suggests a semantic similarity in the gender context, even though the actions described in the sentences are different.

Pair 2:

Sentence 1: "A car is racing down the track."

Sentence 2: "A plane is soaring in the sky."

Cosine Similarity Score (car, plane): 0.6864

- Comment: The cosine similarity score between the words "car" and "plane" indicates a moderate similarity, reflecting the shared context of transportation and movement.

Sentence1: "The car wash is near my apartment. "

Sentence2: "A plane is soaring in the sky."

- Comment: The cosine similarity score between the words "car" and "plane" is less (0.55) than the cosine similarity score between the words "car" and "plane" in above example (0.68). Because the embedding for “car” is different

Pair 3:

Sentence 1: "The chef is preparing a delicious meal."

Sentence 2: "The painter is creating a colorful masterpiece."

Cosine Similarity Score (chef, painter): 0.6046

- Comment: The lower cosine similarity score between "chef" and "painter" suggests less similarity between these specific roles, reflecting the distinct actions and contexts in the sentences.

Pair 4:

Sentence 1: "A scientist is conducting experiments in the lab."

Sentence 2: "An engineer is designing new technologies."

Cosine Similarity Score (scientist, engineer): 0.7422

- Comment: The high cosine similarity score between "scientist" and "engineer" indicates a strong semantic similarity, reflecting the shared context of technical and innovative activities.

Pair 5:

Sentence 1: "A musician is playing a melodic tune."

Sentence 2: "A dancer is gracefully moving to the rhythm."

Cosine Similarity Score (musician, dancer): 0.7255

- Comment: The high cosine similarity score between "musician" and "dancer" suggests a strong semantic similarity related to artistic expression and performance, despite different forms.

In summary, considering the specific words in each pair, the cosine similarity scores align with the expected semantic relationships between the chosen words, reflecting the effectiveness of BERT embeddings in capturing word-level similarity.