Eelisa Jones  
COMPSCI723  
Fall 2021

Assignment 3 Report: Deep Learning Movie Review Sentiment Analysis

**Part I: Model Selection, Configurations, & Expectations**

DistilBERT stands for “Distilled BERT.” It is a lighter architecture version based on the original BERT architecture developed by Google in 2018. To create DistilBERT, developers compressed the original BERT model using knowledge distillation – a practice in which a larger “teacher” model infuses a compressed version of its parameters into a smaller “student model”. Compared to its teacher model, DistilBERT reduces the number of network layers by a factor of two.

In evaluating performance, the DistilBERT developers found that it preserves 97% performance against the GLUE benchmark with 40% less parameters compared to its BERT teacher model. The DistilBERT models were also found to run around 60% faster than its teacher model. The scaling down of the original BERT model provides several benefits, including the reduction of environmental costs and the reduction of computational barrier that users may experience when using less powerful equipment. Moreover, the developers point out that the model is small enough to run on mobile devices, expanding the breadth of scenarios eligible for application.[[1]](#footnote-1)

For this assignment, I have experimented with two versions of the DistilBERT base uncased model: the original version and a version that has incorporated an additional linear neural network layer trained on the Stanford Question Answering Dataset (SQuAD) v 1.1 dataset. The SQuAD dataset is a reading comprehension data set designed to train models for question-answering tasks. I opted to compare the original unbased case and SQuAD version to evaluate the performative difference between the two given that the SQuAD version was specifically trained for another task (i.e., question-answering compared to classification). I naturally hypothesized that the SQuAD version would not perform as well as the non-SQUAD version. I also formulated a second (and perhaps less obviously answerable research question) of to what degree the original base uncased model would outperform the SQuAD-trained version. I believed that comparing the two models would provide some insight into how much performative impact the additional SQuAD fine-tuning would have on the model’s prediction, as well as the nature of that impact. I expected the original base uncased version would produce accuracy scores between 15% and 20% higher than its SQuAD counterpart.

The DistilBERT models used for this assignment were downloaded from the Hugging Face transformer library and were not re-retrained. This assignment places a dense two-unit softmax Keras output layer on top of each model for additional fine-tuning on IMDB movie reviews from the ACL dataset. With the additional Keras output layer in place, both the original uncased and SQuAD uncased versions were trained on 1000 training examples over 3 epochs. Both models were tested on 1000 ACL examples and 20 small dataset examples. Each example contained a maximum of 512 tokens.

**Part II: Model ACL Training & Evaluation Comments**

As predicted, the original DistilBERT base uncased model outperformed the SQuAD linear layer model. The degree of performance difference, however, was much less than originally predicted. The original base uncased model produced an accuracy score of 84.8% while SQuAD version produced an accuracy score of 80.5%. The two models only differed in accuracy scores by approximately 4% - much less than the predicted 15-20%. The 4% difference in accuracies suggests that the additional linear layer within the SQuAD version has had a relatively slight impact on the model’s ability to predict reviews given the same ACL training *time*. It should be of note however, that the max training examples were less than the total number of available examples due to computational limitations; as all examples were first randomized and pooled prior to each model’s training, it is likely that the two models were trained on variated *content* and this difference has likely impacted our ability to evaluate their performance to a certain extent.

The original uncased model appeared to make a more significant increase in training accuracy score between the first and second epoch, while the SQuAD version produced a greater increase in accuracy between the second and third epoch. From the first training epoch to the second, the original version increased in accuracy by 13.8% and the SQuAD version increased in accuracy by 6.9%. From second training epoch to the third, the original version increased in accuracy by 0.8% while the SQuAD version increased in accuracy by 3.8%. These results seem to suggest the SQuAD version produces a more robust input re-representation than the original base uncased version. In turn, the Keras output layer receives a stronger/more transformed representation of an input against which to fine-tune. This hypothesis appears harmonious with the fact that the SQuAD model contains an additional linear layer. From the first to second epoch, the additional layer would appear to prevent a greater increase compared to the original due to stronger re-representation prior to the Keras output layer. The additional representation, however, would also seem to provide more room for accuracy increases as epochs continued (as seen in the epoch 2 to epoch 3 accuracy differences).

It is possible that increasing the number of training epochs for the SQuAD model could produce better training accuracy results; however, increasing training could likewise increase the risk of overfitting the data to the training examples. An alternative approach could be adding additional training examples to the SQuAD version, which could serve to improve the SQuAD version training accuracy while lessening the risk of overfitting. Practically speaking, however, working to improve the SQuAD version in training *may* prove inefficient given (1) the availability of the original base uncased model that already produces better results with the same amount of training and (2) the lack of a guarantee that the SQuAD model could ever outperform the original base case in testing/application if both models receive similar amounts of fine-tuning.

Below are the ACL training and testing results.

|  |
| --- |
| **DistilBERT Base Uncased Training & ACL Testing Data**  **Fitting model to ACL training data**....  **Epoch 1/3**  loss: 0.6542 - accuracy: 0.6520  **Epoch 2/3**  loss: 0.5700 - accuracy: 0.7900  **Epoch 3/3**  loss: 0.5177 - accuracy: 0.7980  **Evaluating model on ACL test dataset....**  loss: 0.4715 - accuracy: 0.8480  Accuracy on ACL test dataset: 0.8479999899864197 |

|  |
| --- |
| **DistilBERT Base Uncased w/ Linear SQuAD Layer Training & ACL Testing Data**  **Fitting model to ACL training data....**  **Epoch 1/3**  loss: 0.6610 - accuracy: 0.6155  **Epoch 2/3**  loss: 0.5874 - accuracy: 0.6840  **Epoch 3/3**  loss: 0.5480 - accuracy: 0.7220  **Evaluating model on ACL test dataset....**  loss: 0.4782 - accuracy: 0.8050  Accuracy on ACL test dataset: 0.8050000071525574 |

**Part III: Model Small Dataset Evaluation Comments**

For the sample dataset, I collected 20 IMDB user reviews for the 2004 film Life Aquatic. Ten of the reviews were associated with a rating of 1 or 2 stars (of 10); these reviews were classified as negative reviews. The remaining ten of the reviews were associated with a rating of 10 or 9 stars (of 10); these reviews were classified as positive.

The small dataset test example text files were created with the following naming convention:

* pos\*.txt for positive examples
* n\*.txt for negative examples

The DistilBERT base uncased version produced an accuracy score of 90% on the small dataset test. The SQuAD version produced an accuracy score of 80%. The 10% difference in small dataset testing more closely resembles my originally predicted 15-20% performance difference but was still 5% lower than my lowest estimate. Interestingly, the original base uncased model produced incorrect predictions for negative examples only; in contrast, the SQuAD version produced incorrect predictions for positive examples only.

Below are the small dataset test results for each model along with brief analyses of test results.

|  |
| --- |
| **DistilBERT Base Uncased Small Dataset Test Data**  **Evaluating model on small test dataset....**  loss: 0.4662 - accuracy: 0.9000  Accuracy on small test dataset: 0.8999999761581421 |

The original model produced incorrect predictions for two test examples – n2.txt and n7.txt. Both were negative reviews that the model predicated as positive. Example n2.txt appears to lack a significant number of negative-associated words and instead consists mostly of neutral exposition. The review does contain words such as “marvelous” and “recommend” which may have provided the model with enough weight strength to incorrectly predict the review as positive with a 64.4% confidence level. Similar to the content of n2.txt, the content of n7.txt lacks a significant number of negative-associated words. This example instead contains many positive-associated words like “quintessential”, “eagerly”, “ever-faithful”, “success”, and “pleasure”. It is likely the model used words such as these to weigh its contextual embeddings more heavily toward a positive review when it produced a positive prediction with 74.1% confidence.

|  |
| --- |
| **DistilBERT Base Uncased Incorrectly Predicted Example Content**   * n2.txt   + Prediction for test example 11: [0.6438299 0.35617006]   + Test example text sample:     - “In this modern age we always have a choice in how to measure things. We can take the high-tech approach or we can go low-tech. As an example, let us refer to a common phrase that we all think we understand, yet so very few of us have ever actually measured. Using the low-tech approach would require two things we already have available to us, a common ruler and any old portable timepiece, be it a wrist watch, a wind-up clock or a battery powered clock. Then we would have only to go outside and hunt around for a while until we found a snail, lay the ruler down next to it and observe the passage of time until we had enough data to compute what a snail's pace is. Or, using the high-tech method, we could go see this movie, with its tens of millions of dollars expended on marvelous cameras and the latest in lighting, and I guess some CGI too. This method would also reveal the meaning of snail's pace although it would cost us money and an interminable amount of our time. But at least it would give this film some purpose, a slight usefulness that it lacks standing on its own. It's your choice, but I recommend the ruler-and-watch method. It's cheaper, faster and a lot less infuriating.” * n7.txt   + Prediction for test example 14: [0.74183923 0.25816083]   + Test example text:   + Quintessential Wes Anderson follows the life of fallen aquatic icon Steve Zissou (Bill Murray), who is an apparent homage to Jacques-Yves Cousteau and his famous marine society. While filming another oceanographic documentary, eagerly awaiting by the public, despite the increasingly critical reaction of once ever-faithful viewers, an almost mythical Jaguar shark consumes Esteban, Zissou's best friend and confidant. Set on the destruction of the shark he venture on his research vessel Belafonte to document this feat. Before embarking Steve Zissou comes across Ned Plimpton (Owen Wilson), who presents himself as his unknown son. After joining "Team Zissou", which amongst others include second-in-command Klaus Daimler (Willem Dafoe), his wife Eleanor (Anjelica Huston), Sikh cameraman Vikram Ray (Waris Ahluwalia) and a host of unpaid college interns. Accompanying them is journalist Jane Winslett-Richardson (Cate Blanchett), come to chronicle the ill-fated voyage. Their direst foe is Alistair Hennessey (Jeff Goldblum), the current golden boy of the oceanographic community, currently going from success to success. An almost peremptory respect for Wes Anderson is almost expected from movie aficionados, but despite best intent almost every watching of quirkiness heavy and goofy populated universes of the director invariably are irritable due to the unconvincing absurdist humour, self-referential love affair with his own on-screen weirdness. Practically the only occasion where the much loved Bill Murray is entirely unwatchable. Eccentricity of characters often feels like the sole idea for plotting in an Anderson movie, making audiences interested in something other than quack-pots with odd posturing, reactions and logic detach themselves from any viewing pleasure. Giving Wes Anderson every chance in the world to catch on only one or two pictures made by him have turned out to be watchable, while the remainder seemed like senseless homework for cinephiles. "The Aquatic Life of Steve Zissou" sadly falls into the latter pool of Anderson movies. At times surrealist, at times cartoonish, including a plastic underwater world populated by colourful and odd creatures. Overly forced, partially plot less (albeit this is something of a norm for Wes Anderson) the feature has an artificial feel which permeates throughout, while personally vexatious characters make the viewing pleasure even more questionable. The deadpan delivery of jokes isn't as cool as Wes Anderson believes them to be, instead with each new movie feels increasingly fixed in mannerism. |

|  |
| --- |
| **DistilBERT Base Uncased w/ Linear SQuAD Layer Small Dataset Test Data**  **Evaluating model on small test dataset....**  loss: 0.4791 - accuracy: 0.8000  Accuracy on small test dataset: 0.800000011920929 |

The base uncased SQuAD model produced incorrect predictions for two positive and two negative review examples – n2.txt, pos3.txt, pos7.txt, and n7.txt. The model was fairly confident that the content of n2.txt implicated a positive review despite the content creator’s actual sentiment. It seems as if key positive-associated terms like “marvelous”, “high-tech”, and “go see this movie” may have produced enough neural weighting for a positive review prediction with a confidence level of 70.3%. The model produced predictions with much less confidence for examples pos3.txt and pos7.txt. Both examples appear to contain a mix of negative-associated and positive-associated words. For instance, pos3.txt contained terms like “great movie” and “loved this movie” but also contained terms like “bit disappointed” and “flop.” During cursory human review, I struggled myself to classify these examples. It appears the model similarly struggled and provided negative predictions for pos3.txt and pos7.txt with mild confidence ratings of 51.7% and 50.2%, respectively. In contrast, the model was more strongly confident in its incorrect prediction of n7.txt as a positive review, providing a confidence of 83.6%. This example contains a fair number of positive-associated terms such as “much loved”, “colourful”, “quirkiness”, and “quintessential” which may have contributed to the strength of the model’s prediction.

|  |
| --- |
| **DistilBERT Base Uncased SQuAD Incorrectly Predicted Example Content**  n2.txt   * + Prediction for test example 16: [0.70346254 0.29653743]   + Test example text:   + In this modern age we always have a choice in how to measure things. We can take the high-tech approach or we can go low-tech. As an example, let us refer to a common phrase that we all think we understand, yet so very few of us have ever actually measured. Using the low-tech approach would require two things we already have available to us, a common ruler and any old portable timepiece, be it a wrist watch, a wind-up clock or a battery powered clock. Then we would have only to go outside and hunt around for a while until we found a snail, lay the ruler down next to it and observe the passage of time until we had enough data to compute what a snail's pace is. Or, using the high-tech method, we could go see this movie, with its tens of millions of dollars expended on marvelous cameras and the latest in lighting, and I guess some CGI too. This method would also reveal the meaning of snail's pace although it would cost us money and an interminable amount of our time. But at least it would give this film some purpose, a slight usefulness that it lacks standing on its own. It's your choice, but I recommend the ruler-and-watch method. It's cheaper, faster and a lot less infuriating. * pos3.txt   + Prediction for test example 14: [0.48259723 0.51740277] - Incorrect   + Test example text:   + I was kind of expecting one of those arty films that make you squirm a bit as you labor through it. I was pleasantly surprised! This was a great movie. I was entranced and entertained throughout. On top of being a great movie in and of itself, so to speak, it was also whimsical, and a bit of a fantasy, with a bit of the magical. I had read from critics I respect that it was a 'valiant try', but was a flop, at the end of the day. I know perceptions and opinions vary, and I am prepared to manage that. But oh my--- how could those critics have got it so wrong? 'The Life Aquatic...' is marvelous! The critics had not been overly kind to this film. Maybe they subconsciously felt that writer/director Wes Anderson had gotten enough praise and enough critical 'help' thus far in his career, and that somehow he was due for some comeuppance? Dunno... but I LOVED this movie. The casting was grand. I am not ordinarily a big fan of Bill Murray, but I liked him here. This was, in fact, the only movie in which I found him thoroughly likable, with the possible exception of 'Caddyshack'. I don't think he's the only person in the world who could have played an intriguing Steve Zissou--- picture Johnny Depp in the role??? Depp would also have good. Although Angelica Houston would then have been miscast as his wife, but hey. The way it was cast, with the super hot Cate Blanchett, the always likable Owen Wilson, and all the other quirky characters, played by such luminaries as Jeff Goldblum, Bud Cort, and Wilhem Dafoe--- was superb! One test for a movie's goodness is the watch test. During the film I never, not once, thought of the time, nor checked my watch. And when it was over, I was a tiny bit disappointed. Oh, I was satisfied enough, but I also wished for more. Like a great party that finally comes to an end. I gave this movie a well deserved 10 out of 10. * pos7.txt   + Prediction for test example 2: [0.49794576 0.5020543 ]   + Test example text:   + Every Movie Buff knows that You need Dynamite to Kill a Shark. So with Sticks on board, We're off to Find the Killer Fish, through the wonderful Visionary World of Director Wes Anderson. His Sharp Eye, like always, is a Hyper Focused, Multi-Colored, Surreal, Symmetrical look at the World that is Vibrating on a slightly Different Frequency than that of regular Folks. As the Film Drifts through Uncharted Seas on a broken down Boat, everything seems somewhat Uncomfortable and Edgy, slightly Out of Step, but is always In Tune with its own Symphony of Oddball Characters and even Odder Sea Creatures. It is a Stunning Display of Artful Imaginings with a Shallow Story line but with a Depth of Emotion that can't help but draw You into its own place of Meticulously Placed Actors and Props that are Posed to posses with an Artificial Attraction that says, "This is real but not really". That is Wes Anderson, take it or leave it. The Pacing is at times Ponderous, but it never lingers enough to bring on Boredom, and it is forever Fascinating Visually. If there is one great Flaw, it is the Overuse of the acoustic Portuguese Language Bowie Songs. It is a Quirk that wears Thin very fast and the thread is Woven into the Film regularly and it is a slight Burden to Bare if Not fully Embraced. * n7.txt   + Prediction for test example 15: [0.8362244 0.1637756]   + Test example text:   + Quintessential Wes Anderson follows the life of fallen aquatic icon Steve Zissou (Bill Murray), who is an apparent homage to Jacques-Yves Cousteau and his famous marine society. While filming another oceanographic documentary, eagerly awaiting by the public, despite the increasingly critical reaction of once ever-faithful viewers, an almost mythical Jaguar shark consumes Esteban, Zissou's best friend and confidant. Set on the destruction of the shark he venture on his research vessel Belafonte to document this feat. Before embarking Steve Zissou comes across Ned Plimpton (Owen Wilson), who presents himself as his unknown son. After joining "Team Zissou", which amongst others include second-in-command Klaus Daimler (Willem Dafoe), his wife Eleanor (Anjelica Huston), Sikh cameraman Vikram Ray (Waris Ahluwalia) and a host of unpaid college interns. Accompanying them is journalist Jane Winslett-Richardson (Cate Blanchett), come to chronicle the ill-fated voyage. Their direst foe is Alistair Hennessey (Jeff Goldblum), the current golden boy of the oceanographic community, currently going from success to success. An almost peremptory respect for Wes Anderson is almost expected from movie aficionados, but despite best intent almost every watching of quirkiness heavy and goofy populated universes of the director invariably are irritable due to the unconvincing absurdist humour, self-referential love affair with his own on-screen weirdness. Practically the only occasion where the much loved Bill Murray is entirely unwatchable. Eccentricity of characters often feels like the sole idea for plotting in an Anderson movie, making audiences interested in something other than quack-pots with odd posturing, reactions and logic detach themselves from any viewing pleasure. Giving Wes Anderson every chance in the world to catch on only one or two pictures made by him have turned out to be watchable, while the remainder seemed like senseless homework for cinephiles. "The Aquatic Life of Steve Zissou" sadly falls into the latter pool of Anderson movies. At times surrealist, at times cartoonish, including a plastic underwater world populated by colourful and odd creatures. Overly forced, partially plot less (albeit this is something of a norm for Wes Anderson) the feature has an artificial feel which permeates throughout, while personally vexatious characters make the viewing pleasure even more questionable. The deadpan delivery of jokes isn't as cool as Wes Anderson believes them to be, instead with each new movie feels increasingly fixed in mannerism. |

1. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. [↑](#footnote-ref-1)