NLP Final Project

Jeremy Drouillard

Abstract

Natural language understanding tasks can often be convincingly solved by artificial neural networks. However, the model performance can belie the simplistic understanding of the text and task. During training, models may learn heuristics that are good enough to fit the training data but cause the model to not generalize well on more challenging instances of the task. For this problem, there exists many hypothesized solutions: altering the model, training data, or training procedure. For the NLP final project, I've demonstrated the shortcomings of an ELECTRA model fine-tuned on the SQUAD dataset. To improve the model performance on questions with many plausible answers, I iterate on the ideas from Dataset Cartography to construct a subset that improves performance on a challenging validation set.

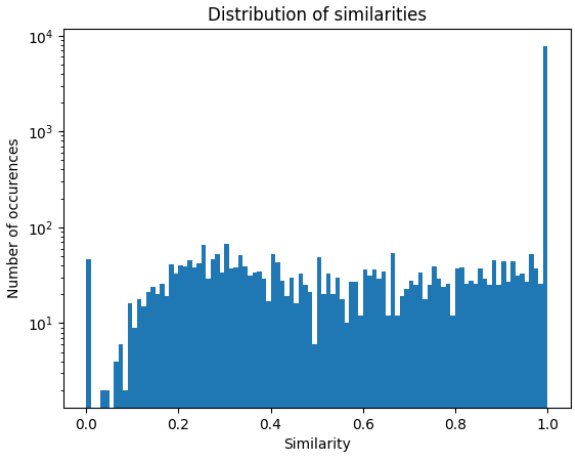

Figure 1: Similarity is the measure of overlap between the predicted answer and gold label. As similarity approaches 1, the answer improves. Qualitatively, anything below .25 was a bad answer.

1 Initial Analysis

I didn't start out with the intention of reapplying dataset cartography (Swayamdipta et al., 2020) ideas. It was only after an analysis of the out-of-dataset errors that I sought out solutions. In my initial analysis, I fine-tuned an ELECTRA model across 10 epochs and then ran it against the validation set. I then ranked the validation predictions based on their similarity to the gold labels. An error was considered more severe the less it overlapped with the gold answers (i.e. 0 similarity meant no overlap). As you can see in the chart below, roughly 100 of the most severe errors would provide an answer that contained no overlap.

I skimmed the 100 most severe errors to better understand what causes a complete miss. Qualitatively, the vast majority of misses are what I'll call plausible distractors. Plausible distractors are answers that satisfy less contextful heuristics. For example, if a question begins "how many..." and the context contains two ideas involving numeric information, one being the answer, then the other would be considered the plausible distractor. It's not a terrible answer, because it is a number that occurred in the context. To be even more concrete, here's a real example.

> Q: How many teams up to Super Bowl 50 have been to the championship game eight times?
>
> Context: . . . The Carolina Panthers became one of only ten teams to have completed a regular season with only one loss, and one of only six teams to have acquired a 15–1 record, while the Denver Broncos became one of four teams to have made eight appearances in the Super Bowl. . .
>
> Predicted: "six"

The predicted answer is indeed a number about the number of teams to have accomplished something,

but it's just a different something than the question asked.

I wanted to further understand the nature of distractions, so I asked the following questions for 45 of the errors:

- Does the context present multiple plausible subjects? In the figures, this is the multiple_subjects tag.
- For the correct subject, are there multiple plausible options? Tag: multiple_options.
- Does the answer match the subject type of the question? Tag: subject_type_match.
- Does the answer match the criteria (e.g. proper noun, adjective)? Tag: criteria_match.
- Ignoring the subject, is the answer's option plausible? Tag: answer_option_plausible.

Sticking with the example, the respective answers to these questions are:

- Yes, multiple teams are mentioned throughout the context.
- Yes, there are many numeric stats associated with all the teams mentioned.
- Yes, "six" refers to the number of teams to have accomplished something.
- N/A. There's no singular correct team, since multiple teams are implied by the question.
- Yes, the predicted answer is numeric.

Across the 45 examples, we can see errors occur because of plausible distractors. Digging further into the structure of the distracting questions and context, we can see in figure 3 distracting subjects occur 90% of the time and distracting options on the correct subject occur 60% of the time.

What this data shows is that there are varied challenges in the SQUAD dataset. In order to perform well, the model must ignore good looking answers that are about different subjects and cannot latch on to the first acceptable answer near the correct subject.

How did the model fare on these challenging problems? In figure 4, we can see the breakdown. Roughly half the time, the correct subject was matched, but the incorrect option was picked.

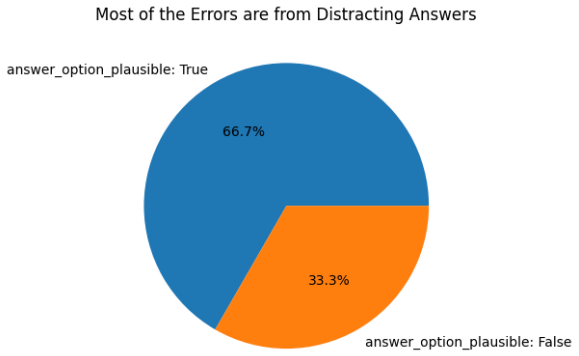

Figure 2: In two thirds of the examined errors, the predicted answer was a plausible distractor.

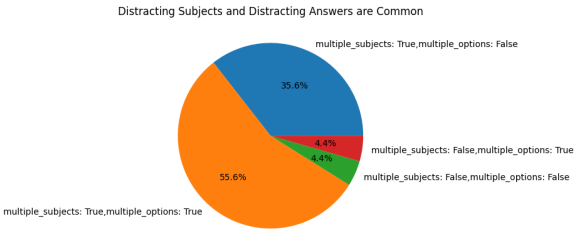

Figure 3: Distraction often takes the form of a context with many related subjects with facts that are somewhat similar to the one sought by the question.

A quarter of the time, the model didn't pick a plausible subject, and another quarter the model was distracted by a plausible subject. These results cast doubt on the model's natural language understanding. A more complex context with more information to sift through can cause the model to produce answers that are about a different subject and unrelated to the given question.

Before developing a possible fix, I wanted to identify if there was a particular type of subject or question that caused the model inaccuracy. Assuming the SQUAD dataset provides good coverage of different types of questions and contexts,

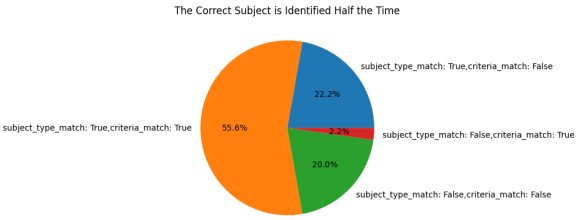

Figure 4: The model has a variable response to distractions. Most of the time it will identify the correct subject, but pick the wrong plausible option.

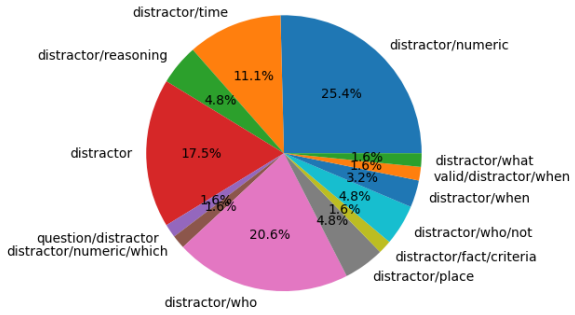

Figure 5: Different types of distraction observed when sampling severe errors.

then failures on a particular subset of questions would be a worthwhile target for improvement. Alas, there was not a particular type of distraction that the model struggled with more than others (figure 5).

The different categories of distraction were created on an ad hoc basis to describe the errors during analysis. No dominant category emerges. The model could be distracted by different times, places, objects, people, options, and quantities depending on the question it was asked.

All of this analysis illustrates how a model misbehaves. In question and answering, there are naturally occurring distractions that are about similar subjects and similar ideas. There are many proposed solutions to this problem of varying complexity and toil. Following this analysis, I was stuck on the basic desire for the model to learn more from the training data. The training data contains questions with a similar level and type of difficulty, so wouldn't focusing on these examples produce a model that can generalize to distracting context better?

2 Dataset Cartography

Motivated by the prior analysis of errors, I wanted to fine-tune the model on subsets of the training data with more distractions. The ideas from Dataset Cartography provide a roadmap for just that. I'd expect training examples that are ambiguous (i.e. variable across training epochs) and hard to learn are more likely to contain distractions. In this section, I will describe the steps I took to apply the ideas from Dataset Cartography to finetuning on the SQUAD dataset.

First, let's recap the main idea from Dataset Cartography. The idea is to train models on subsets of the training data based on how easy it is to minimize the loss for the training examples. There are two dimensions summary metrics guiding us. The first is "confidence... the mean model probability of the true label." In other words, confidence is the mean probability of the correct answer over multiple training epochs. The other dimension is variability, or the variance of the confidence across epochs. High variance examples are considered ambiguous. High confidence is considered easy and low confidence is considered hard.

Before constructing subsets, I computed the confidence and variance from fine tuning. Typically, fine tuning is done in fewer epochs, but fewer epochs makes it harder to separate the ambiguous from the certain. In order to get a better estimate of the confidence and variance, I took a larger sample by fine tuning over ten epochs instead of three. Since this strays from recommended practice, I compared a model trained over all training data for three epochs to one trained over ten. The ten epoch model outperformed the three epoch model on both the training data and the validation set, so I'm not concerned that extending the training cycles undermined the models ability to generalize.

Confidence is nearly synonymous with loss in this case. During training, the loss is the average of the cross entropy loss for the start position and the end position. In other words, the loss is the log probability of the correct start plus the log probability of the correct end. Applying the log product rule, we can take the captured losses during training and convert them into the probability of the correct answer: the product of the start and end probabilities. Now, we have all we need to compute the confidence and variance. The ambiguous third of training data, which trained the ambiguous model, is seen in figure 6.

In addition to the ambiguous set, I computed a random, hard, and easy one-third set. The easy set is the upper third on the confidence spectrum; hard is the bottom. Random is used as a baseline. Out of curiosity, I also created two 66% models using the not-easy and not-low-variance thirds of data to better understand the impact of training set size on model performance.

For the novel piece of my project, I constructed a final 33% dataset to test another hypothesis. I noticed while inspecting the composition of the ambiguous dataset that many uses of the same con-

Figure 6: The ambiguous dataset is shown in red.

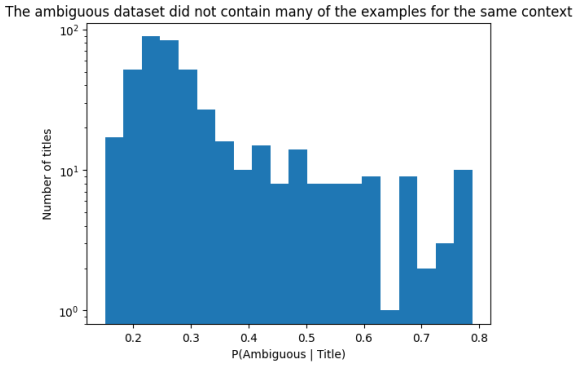

Figure 7: This histogram displays the probability a training example is ambiguous given a title. A value of 1 would mean every training example for the title is ambiguous. As we can see, most titles contain some ambiguous data points, but they're not the majority.

text were no longer included in the dataset (seen in figure 7). In other words, ambiguous examples were separated from their easy and hard counterparts on the same context. This motivated the following idea: a training example's ambiguity is not just a property of itself, but the other examples within the training data. One can imagine it's easier to learn a heuristic for a context that occurs infrequently with unique questions whereas multiple questions of the same context would require a more nuanced understanding ability. This notion is similar to that in Contrast Sets (Gardner et al., 2020) which demonstrated the value of having training examples which force the learning of more nuanced decision boundaries. By training on all the questions of select contexts, the model will ideally learn a more sophisticated decision boundary for the different questions.

In order to test this interdependence hypothesis,

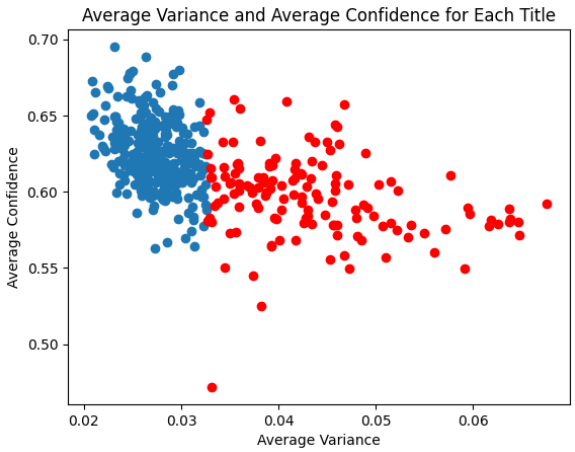

Figure 8: The titles plotted by their average variance and average confidence. Red points had all their examples included in the most-variable-titles set.

I constructed a training dataset based on the average variability for a given title. Figure 8 highlights that we see a similar easy/hard and ambiguous pattern emerge when we average training examples based on their shared context. The algorithm to construct the dataset is simple. Take the rightmost title and all associated training examples not yet included in the dataset until you have 33% of the data. This dataset is called the most-variable-titles dataset.

This is an interesting way to slice the data, but does it find us multi-subject/option problems? Let's look at the most variable title, the University of Notre Dame, to find out. This is the context for some Notre Dame questions:

> The College of Engineering was established in 1920, however, early courses in civil and mechanical engineering were a part of the College of Science since the 1870s. Today the college, housed in the Fitzpatrick, Cushing, and Stinson-Remick Halls of Engineering, includes five departments of study – aerospace and mechanical engineering, chemical and biomolecular engineering, civil engineering and geological sciences, computer science and engineering, and electrical engineering – with eight B.S. degrees offered. Additionally, the college offers five-year dual degree programs with the Colleges of Arts and Letters and of Business awarding additional B.A. and Master of Business Adminis-

tration (MBA) degrees, respectively.

In isolation, this doesn't tell us much; we must consider the questions. Some of the questions of this context are:

- "How many BS level degrees are offered in the College of Engineering at Notre Dame?"
- "In what year was the College of Engineering at Notre Dame formed?"
- "Before the creation of the College of Engineering similar studies were carried out at which Notre Dame college?"

For the first question, the subject is colleges within Notre Dame and plausible options would be numeric. Multiple colleges are mentioned in the context and there are multiple pieces of numeric data. This looks like the hard problems from the beginning. The remaining questions also have multiple similar subjects and options available as well.

How do the Notre Dame examples compare to the least variable "Muslim World" title? Here's one context about Muslim World:

> More than 20% of the world's population is Muslim. Current estimates conclude that the number of Muslims in the world is around 1,5 billion. Muslims are the majority in 49 countries, they speak hundreds of languages and come from diverse ethnic backgrounds. Major languages spoken by Muslims include Arabic, Urdu, Bengali, Punjabi, Malay, Javanese, Sundanese, Swahili, Hausa, Fula, Berber, Tuareg, Somali, Albanian, Bosnian, Russian, Turkish, Azeri, Kazakh, Uzbek, Tatar, Persian, Kurdish, Pashto, Balochi, Sindhi and Kashmiri, among many others.

The associated questions:

- "In how many nations are Muslims in the majority of people?"
- "How many languages are spoken by Muslims?"
- "How many Muslims are believed to live on earth?"

In this context, there are three numeric pieces of data all about the same group of people. Muslims are the subject of the questions and they are the only group of people mentioned in the context. Qualitatively, the Muslim World questions and context bear less of a resemblance to the challenging subset. To test this notion that the least variable titles don't represent the challenging multi-subject/option questions well, I've also created a low-variance-titles model. This model was trained on 33% of the data using the least variable titles.

Now, all the experimental datasets have been defined using ideas from Dataset Cartography. In the next section, I'll examine how each of the fine-tuned models fared on the validation set and the distraction rich subset from the initial analysis.

3 Reviewing the Results

The simplest analysis of the fine-tuned models is just comparing the performance on the validation set. Perhaps unsurprisingly, none of the subset models are able to outperform the "default" model trained on all of the data (table 1). It's still worth pausing here to collect all the observations to gain a holistic understanding of the models and subsets compared to the random baseline. First, there's a clear positive correlation between the size of the dataset and the validation performance. Outside of that, we don't expect a significant change in model performance on this coarse view. If we had seen improvements here over random, that would suggest the training dataset contains many examples that are counterproductive to generalizing (e.g. mislabels). Perhaps on other tasks this is possible, but the SQUAD dataset appears well constructed. Lastly, none of the sophisticated 33% subsets significantly outperform the others or the random baseline. I was aiming to improve the model performance on the distraction-rich examples, so summary statistics across all validation examples won't tell us much.

Subset training doesn't yield a broad improvement, but it may improve performance on the types of errors in the initial analysis. For these predictions, I compare the subset models' performance on the multi_subject and multi_option examples from the initial error analysis. These were the most common distraction types and arguably the hardest types of distraction-prone questions. The results can be found in table 2.

These results are somewhat aligned with my

Model Name	F1	Exact Match
default	85.53	77.54
not-low-variance	84.08	75.5
not-easy	84.42	76.0
most-variable-titles	81.2	72.2
least-variable-titles	82.3	73.3
random	81.6	72.8
ambiguous	81.8	72.8
easy	82.3	73.2
hard	82.1	73.0

Table 1: Validation Results

Model Name	F1	Exact Match
most-variable-titles	27.2	24
least-variable-titles	7.3	4.0
random	22.1	16
ambiguous	16.25	12
easy	19.2	16
hard	25.83	20

Table 2: multi-subject and multi-option validation

expectations. The most-variable-titles model and the hard model outperform the baseline random model. The hypothesis around the most-variable-titles model seems to be supported. Multiple questions of the same context are important for navigating questions and context with multiple plausible subjects and answers. The difficulty of the context and question pairs seems important too, because the least-variable-titles model performed much worse on this task. The ambiguous model also did not perform well. I speculate this the counterpart to the most-variable-titles' success. Contrast sets would have been destroyed during the construction of the ambiguous dataset. Because of the high context-to-question ratio, it was too easy to learn heuristics that fall apart in the face off multiple subjects and options.

3.1 Conclusion

In this final project, I explored the impact of fine tuning an ELECTRA model on different subsets of the training data in an effort to reduce the model's susceptibility to distraction. In the initial analysis, I identified there are numerous ways for a model to be distracted. In it's most extreme form, which I most closely examined, there are multiple similar subjects and potential answers for each subject in the same context. I demonstrated a subset constructed of the most ambiguous contexts during fine tuning can out perform a random baseline and other subsets on the challenging subset. My experiments demonstrated that there is value in picking more ambiguous contexts; it is not enough to pick fewer contexts. The title and ambiguity aware subset construction demonstrates the importance of considering the relationship amongst the training examples when constructing training sets.

References

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of EMNLP*.