Dataset Artifacts and Cartography in Question Answering

Daniel Fernandez
University of Texas at Austin

Abstract

Transformer models achieve high performance on NLP tasks. Unfortunately, due to common patterns in training data annotations, such models often fail to generalize to out-of-distribution (OOD) data. This project explores such data set artifacts on the SQUAD data set and Electra model, including specific failure patterns observed during model evaluation on OOD data. Additionally, it explores using Data Maps (Swayamdipta et al., 2020) to calculate confidence and variability on the SQUAD training data, using the logits of the start of answers. Experiments show that it is possible to reach better generalization on OOD data using the Electra model trained on a filtered subset of hard or ambiguous examples of SQUAD, without the need of augmenting training data to account for adversarial or OOD examples.

1 Introduction

Understanding natural language and being able to answer questions about human text is a very difficult task for computers. Several data sets have been created to train NLP models for such endeavor, one of the most popular ones being SQUAD (Rajpurkar et al., 2016). SQUAD consists of 100,000 examples, which include a context, a question and possible answers. Context paragraphs were sampled from Wikipedia, while question generation was crowd-sourced. However, it is very difficult for such data set to capture every possible context topic or question structure; therefore, generalizing to out-of-distribution (OOD) data is very challenging.

Large pre-trained models, such as Electra (Clark et al., 2020), have made some progress towards OOD generalization (Hendrycks et al., 2020), but adversarial examples or complex grammatical structures non-present on training data still cause such models to fail on such examples.

Expanding on the concept of Data Maps (Swayamdipta et al., 2020), I propose utilizing data set cartography to find the confidence and variability of model predictions during training, with the purpose of detecting instances where the model struggles the most to learn. With this information, the original data can be filtered as desired to find the optimal examples to use during training. I also introduce two new contrast sets: cSQuAD1 and cSQuAD2, based on the work of Gardner et al. (2020), that aim to transform the local decision boundary on such sets, perturbing model predictions.

Different filtering and sampling approaches are also explored and evaluated on the original SQUAD validation set, cSQuAD1 and cSQuAD2, proving that finding hard and ambiguous examples and training on them leads to better OOD generalization in question answering.

2 Data Set Artifacts

Data set artifacts are systematic gaps in data sets that allow models to perform well on such data without understanding it meaningfully (Gardner et al., 2020).

2.1 Detecting Artifacts

Detecting systematic gaps in the SQUAD data set is no easy task. In my experiments I focused on creating contrast sets to find such patterns. Even though it is recommended that data set authors create such contrast sets, creating them ad-hoc (by non-authors) still helps towards evaluating OOD model performance (Gardner et al., 2020). To detect such examples, electra-small-discriminator was trained on the original SQUAD data set as a baseline model.

The first kind of failure observed was multi-step question understanding. Given the following excerpt from a context passage about Super Bowl 50:

> [...] The game was played on February 7, 2016, at Levi's Stadium in the San

> Francisco Bay Area at Santa Clara, California...

the baseline model can easily respond to questions like: "In what city is Levi's Stadium located?", an indication that the model is able to correctly identify that the subject of the question (the stadium) is located at Santa Clara, California. Yet, by transforming the question into: "In what city is the stadium where the 50 super bowl was played at?", the baseline model outputs: "levi's stadium". This example shows that Electra is associating "where" to the stadium, but it is given little significance to the main subject of the question, which is the city.

The next kind of failure detected were small, unclear questions. For example, given another excerpt from a passage context:

> The Broncos took an early lead in Super Bowl 50 and never trailed. Newton was limited by Denver's defense, which sacked him seven times and forced him into three turnovers, including a fumble which they recovered for a touchdown. Denver linebacker Von Miller was named Super Bowl MVP, recording five solo tackles, 2½ sacks, and two forced fumbles.

and the question: "Fumbles?", the baseline model produces no output. This is arguably a hard question, but a human, given such context paragraph, would probably respond: two forced fumbles.

Another class of issues found (similar to the previous one in some way), were questions not using the exact same words or synonyms from words in context. For example, given a passage about the 2006 World Cup with the following context:

> [...] Teams representing 198 national football associations from all six populated continents participated in the qualification process...

the baseline model is not able to answer: "How many countries?". It is not clear in the text the answer, but with some basic knowledge about international sport events, a human could easily deduce that the "198 national football associations" correspond to 198 countries. Again, the baseline model falls short of connecting the subject of the question to the most-likely answer in the context, even though the answer in not explicitly clear.

2.2 Creating Contrast Sets

Before attempting to address such data set artifacts, an evaluation method is required. Based on the work of Gardner et al. (2020), I propose creating two new small contrast sets for SQUAD, which target such systematic failures found on training data.

The first one, cSQuAD1, was created by first sampling 100 examples from SQUAD's test split. Then, both context and questions were perturbed in some examples, in a model-agnostic way. The goal of this data set is to expose models that reach high performance on the original 100 instances sample, but yet fail to answer correctly those small perturbations.

To better evaluate performance on out-of-domain data, I introduced cSQuAD2, a very small set of 40 examples, which was created by finding random articles on Wikipedia and creating questions and answers for them. While this data set does not specifically target common failure patterns found in SQUAD, it is a good benchmark to detect if a model performs well on data the could be introduced by an average user of the model (out-of-domain).

Both of these data sets were manually created by me, and are available through Hugging Face Datasets (dferndz/cSQuAD1 and dferndz/cSQuAD2).

3 Improving OOD Performance

3.1 Data Maps

Swayamdipta et al. (2020) introduced Data Maps, a framework that aims to characterize data sets. Data Maps uses prediction logits and gold label information to compute both confidence and variability across multiple training epochs. Confidence determines how confident the learner assigns the true label of the example, while variability captures the spread across epochs of the standard deviation.

3.2 Training Dynamics

To generate data set cartography for SQUAD, the first step is to generate training dynamics: for each epoch, output the example ID, the prediction logits and the expected gold label. To achieve this, the Question Answering Trainer was modified to also save in the output directory the required data for each epoch.

Question Answering models for data sets such as SQUAD output the probability distribution of a token in the context being the start of the answer,

as well as the probability distribution of a token being the end of the answer. This presents a challenge when computing data maps, as the framework expects a single set of logits and true label. The most simple way to handle this is to assume that the most difficult task is to find where the answer starts. Thus, the modified trainer saves only the answer-start distribution.

3.3 Data Set Cartography

After running the Data Maps framework on Electra training dynamics, a pattern similar to that described by Swayamdipta et al. (2020) is observed (Appendix A). The plot shows the two Data Map metrics: confidence and variability. Examples located to the lower-left end of the plot are hard examples, where the model has a very low certainty of assigning the true label. Examples located to the right of the plot have high variability; that is, the model is indecisive on the label it assigns across multiple epochs. Finally, examples located on the upper-left section have high confidence and low variability. The plot shows a fairly uniform distribution of examples, which gives the same "importance" to all hard, ambiguous and easy instances during training.

Then, the original SQUAD training set is filtered into hard (low confidence) and ambiguous (high variability) examples. On a closer look at hard examples, we can find several classes of obstacles for the model. One is questions for which the correct answer is not very clear and spread across the entire context. For example, given the passage:

> The city council of the city of Bern decided against having twinned cities except for a temporary (during the UEFA Euro 2008) cooperation with the Austrian city Salzburg

the answer to the question "What was the twinning city with Bern in 2008?" is "Salzburg", which is relatively far in the context from cue words such as "Bern" or "twinning". Other hard examples include questions that require multi-step understanding, such as "Where was Namibia ranked on Press Freedom in 2014?", given a context that contains rankings for multiple years.

The ambiguous data set included examples where exists a sentence in the context, containing multiple "cue" words (tokens that are not frequent and also appear in question) and a potential answer, but where the actual answer is in the following sentence. For example, in the passage:

> Shell initiated its US$4.5 billion Arctic drilling program in 2006, after the corporation purchased the "Kulluk" oil rig and leased the Noble Discoverer drillship. At inception, the project was led by Pete Slaiby, a Shell executive who had previously worked in the North Sea.

a "potential" answer to the question "What executive initially led the Artic drilling project?" could be Shell, since it is the entity which "initiated" the program based on the context. But the question asks for a person, which in this case is "Pete Slaiby".

3.4 Enhanced Training

With the goal of reaching better performance on both out-of-distribution and out-of-domain data, the electra-small-discriminator model (Clark et al., 2020) was trained of different samples of hard and ambiguous examples. To sample data, the filtering capabilities of the Data Maps framework were used to generate data sets of the top 50% and 75% hard and ambiguous examples. As Swayamdipta et al. (2020) suggests, training on only hard or ambiguous data yields better generalization.

A new strategy was explored: combining hard and ambiguous examples, while allowing the resulting data to contain duplicate examples (that are both in the top 75% of hard examples and 75% of ambiguous examples. This approach effectively assigns more weight to examples that are both hard and ambiguous, since they will result in more gradient updates during training. By combining hard and ambiguous examples, size of training data is not reduced; thus, degraded in-domain performance should be prevented.

3.5 Evaluation

Appendix B shows the evaluation results from training electra-small-disciminator on the original SQUAD dataset, as well as on the samples described above. The baseline model, trained on the full SQUAD training split, reaches high performance on a sample from the SQUAD test set. On the other hand, both accuracy and f1 are degraded on both cSQuAD1 and cSQuAD2, as a result of small perturbations and out-of domain data. The model trained on the top 50% ambiguous examples

is better at handling perturbations on cSQuAD1, but still fails on cSQuAD2 (out-of-domain data). Training only on hard examples results in the best performance in out-of-domain data, but fails at handling perturbations in cSQuAD1. By combining both ambiguous and hard examples, and assigning higher significance to examples that are both ambiguous and hard (which are over-sampled in the combined data), the model reaches the best performance in both in-domain data and sample with perturbations, while still showing considerable improvements in out-of-domain data (cSQuAD2), with respect to the baseline.

The model trained on both hard and ambiguous example is able to correctly answer the questions described in Detecting Artifacts. For example, given a passage about Las Vegas, which contains weather information during different times of the year, this model is able to answer: "What is the Lowest temperature in Las Vegas?", which is never explicitly mentioned in the context; the baseline model provided an incorrect answer on this instance. Additionally, given the context:

> World War II is generally considered to have begun on 1 September 1939, when Nazi Germany, under Adolf Hitler, invaded Poland. The United Kingdom and France subsequently declared war on Germany on 3 September. Under the Molotov–Ribbentrop Pact of August 1939, Germany and the Soviet Union had partitioned Poland and marked out their 'spheres of influence' across Finland, Estonia, Latvia, Lithuania and Romania. From late 1939 to early 1941, in a series of campaigns and treaties, Germany conquered or controlled much of continental Europe, and formed the Axis alliance with Italy and Japan (with other countries later).

this model correctly answered the question: "Who led Germany during World War II?" (Adolf Hitler), while the baseline model answered: "european axis powers". Once again, the base model could not understand that the question asks for "who", which refers to a person rather than countries.

Conclusion

Modern NLP models are able to solve increasingly complex tasks. While large data sets have been essential for training such NLP models, analyzing and understanding data set artifacts could be used to optimize training and reach better performance on out-of-domain data, without the need of collecting neither additional data nor creating adversarial examples.

This work expands Data Maps to support creating data set cartography for question answering. By filtering hard and ambiguous examples from the training data, models achieve better ODD performance, and by combining both of this classes, degraded performance on in-domain examples is prevented. Additionally, training on only the most difficult half of the data reaches better performance on out-of-domain data than the baseline model, while utilizing half the computational resources.

I encourage all future efforts to create new question-answering data sets to utilize Data Maps to better understand the data and provide contrast and out-of domain sets. This provides more details about how well models generalize the given task, and give a better picture of their performance in production.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

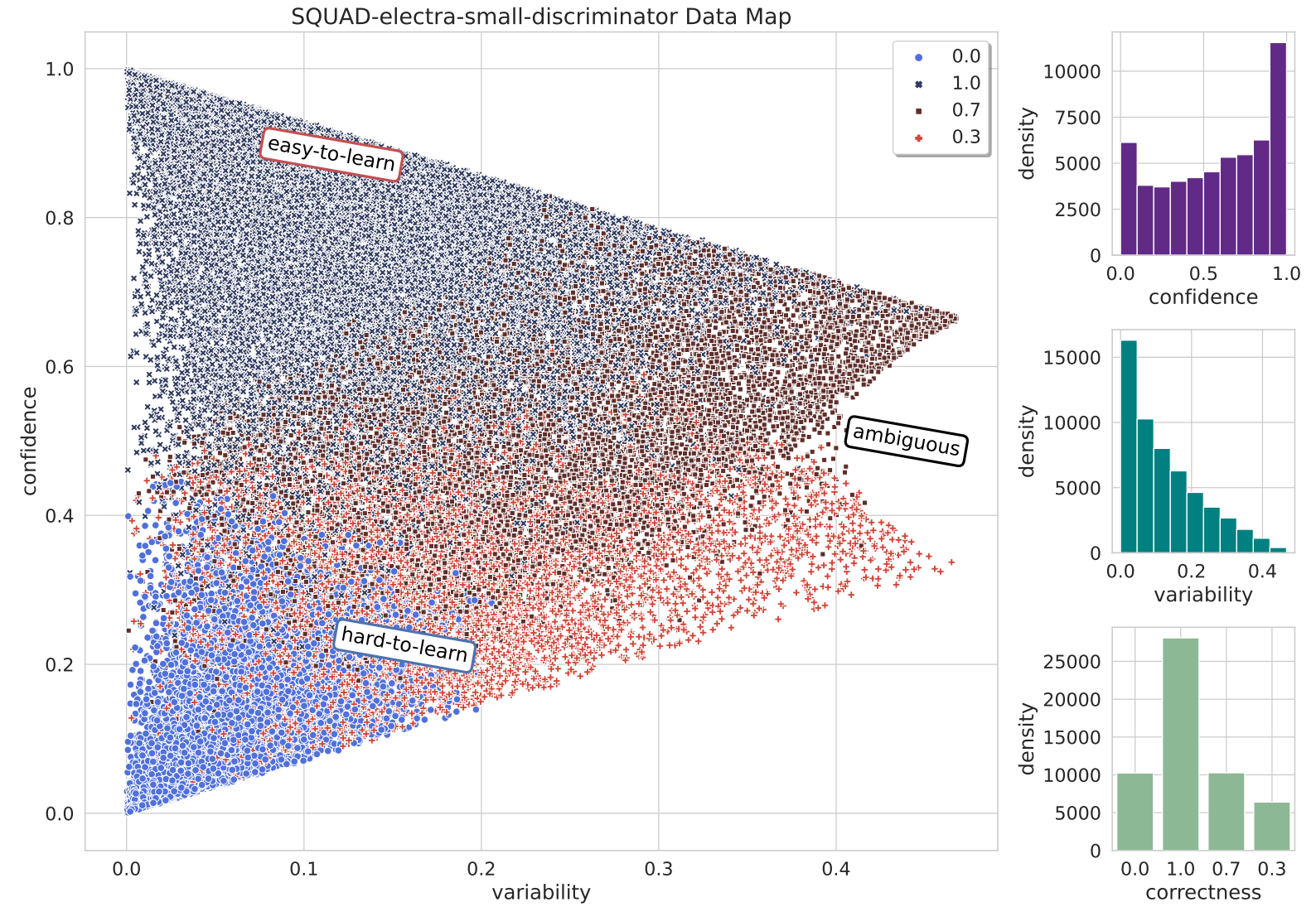
Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah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

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

A Electra-SQUAD Data Map



B Evaluation Results

	SQUAD (100 examples)		cSQuAD1		cSQuAD2	
Training sample	acc	f1	acc	f1	acc	f1
100 % SQUAD (baseline)	84.00	86.18	78.00	80.68	55.00	67.56
top 50% ambiguous	88.00	91.06	84.00	87.87	52.50	66.73
top 75% ambiguous	82.00	86.83	77.00	82.73	60.00	71.72
top 50% hard	77.00	84.10	74.00	81.90	62.50	75.47
top 75% hard	79.00	84.11	76.00	82.01	**62.50**	**75.73**
top 75% ambiguous + top 75% hard	**89.00**	**91.56**	**86.00**	**88.96**	60.00	73.98

Table 1: Exact match accuracy and f1 metrics from electra-small-discriminator trained on the original SQUAD data set and on samples generated using Data Maps