Oregon State University

AI534 - MACHINE LEARNING

HW5: Machine Learning Paper Review

Authors

Melek Derman - 934382167

December 10, 2024

Non-technical background

Pranav Rajpurkar, Robin Jia, and Percy Liang's 2018 study titled "Know What You Don't Know: Unanswerable Questions for SQuAD" was published as a conference paper in the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), in Melbourne, Australia. At the time of publication, all authors were affiliated with the Computer Science Department at Stanford University, Stanford, California, with Percy Liang serving as the Principal Investigator (PI). Pranav Rajpurkar and Robin Jia were graduate students in the Computer Science Department at Stanford University during that period. According to Google Scholar, the paper has received 3,103 citations, while Semantic Scholar reports 2,607 citations, of which 643 are classified as highly influential. The citation count exhibited an increasing trend until 2020 and then stabilized.

This influential paper presents an extended version of the Stanford Question Answering Dataset (SQuAD), a widely used reading comprehension dataset developed using a large collection of Wikipedia articles by a team of collaborators. Building upon SQuAD 1.1, which includes 100,000 questions and was also published by Percy Liang, the study introduces more than 50,000 unanswerable questions to resemble the answerable ones, creating the SQuAD 2.0. The primary objective of the study is to address a key limitation of extractive reading comprehension systems: while these systems are good at answering questions based on a given document, they tend to produce incorrect answers when faced with unanswerable questions instead of indicating the absence of an answer. By identifying this tendency to provide wrong answers as a weakness, the study aims to enhance the robustness of such systems. In their evaluation, the authors describe SQuAD 2.0 as a challenging natural language understanding task, substantiating this claim by demonstrating that powerful natural language systems achieving an 86% F1 score on SQuAD 1.1 only reach a 66% F1 score on SQuAD 2.0.

There are two articles discussing SQuAD 2.0 on Medium.com, an online platform where individuals can publish their blog articles. In this paper, the authors created the dataset using crowdworkers, and it has been released as open source through the CodaLab website. Additionally, the dataset is available in the SQuAD-explorer repository on Pranav Rajpurkar's GitHub webpage. On this repository, the training, development, and test splits are clearly seperated, and individuals who have reproduced the dataset using various methods have published their Exact Match (EM) and F1 scores. Alongside the clearly separated sets, the site also provides the evaluation scripts and sample submission files used to evaluate models. Since 2018, over 100 individuals or groups have made predictions on SQuAD 2.0 on the GitHub Leaderboard, clearly indicating the paper's and dataset's significant influence.

Furthermore, the CodaLab site includes all details of the three models evaluated on SQuAD 2.0, complete with demo codes. Additionally, various posters and slides related to the study are available on Stanford University's website, and some videos about the SQuAD 2.0 dataset can be found on the YouTube online video sharing platform.

In conclusion, the study is well-documented, and all the code and datasets used are released as open

source on platforms like CodaLab and GitHub. These resources have been reproduced hundreds of times by many other studies, emphasizing the study's substantial impact in the field of computer science.

Core

What?

In this work, authors specifically discussed how extractive reading comprehension systems sometimes fail where there is not a direct answer expressed within a given context to the question asked. These models are then fine-tuned on datasets such as SQuAD 1.1, for which an assumption is being made that every question has some sort of valid answer within its context. Because of this, the systems focus on picking text may be relevant rather than deeply understanding whether an answer is correct or wrong.

This is not a new challenge in natural language understanding, but previous approaches to solve it have shown a number of significant limitations. Approaches for automated unanswerable question generation, such as rule-based or distant supervision, tend to create questions that are either completely irrelevant to the context or even simplistic enough that a model would easily be able to determine. Thus, this type of generated question does not test the robustness of comprehension systems. Most of the datasets are focused only on answerable questions; this makes the systems unable to handle real-world scenarios, which have a large number of unanswerable questions. The lack of complexity in earlier datasets has also restricted their usefulness for evaluating advanced models' understanding capabilities.

To overcome these limitations, the authors developed SQuAD 2.0, a dataset containing more than 50,000 unanswerable questions. These are human-generated questions which are designed to be comparable to the answerable ones but for which there is no valid answer in the given context. With SQuAD 2.0, the system faced a new challenge, where it had to skip answering if there was no valid answer. This is the type of situation where we want a model to recognize when a question cannot be answered based on the given context.

This approach helps systems better understand language and context. Adding human-made unanswerable questions makes the dataset more complex and varied. SQuAD 2.0 shows the problems in current systems and gives a starting point for future research to build stronger, more reliable, and smarter models. The authors not only improved the system with SQuAD 2.0 but also set new goals for progress in natural language understanding. This is an important step toward making reading systems more useful in real-world situations.

Why?

The authors chose this problem because it is very important and challenging in NLP. Unanswerable questions are a key issue in making reading comprehension systems more reliable, flexible,

and useful for practical purposes. For real-world applications like virtual assistants or automated customer support, giving wrong answers or failing to recognize when there is no correct answer can harm trust in the system. A system that clearly shows there is no answer is much more reliable than one that gives misleading or incorrect responses, which is important for user satisfaction and system reliability.

Current models trained on datasets like SQuAD 1.1 perform well on answerable questions but struggle with harder or unanswerable ones. This shows that these models often fail to understand the context and lack the ability to properly check their answers. Solving this issue will help systems handle more complex and realistic situations. The large gap in performance between humans and machines also highlights room for improvement. For instance, humans outperform the best models on SQuAD 2.0 by over 20 F1 points, showing significant potential for better results.

Additionally, existing datasets are often too simple to test models effectively. SQuAD 2.0, which includes carefully designed unanswerable questions, sets a higher standard and creates a challenge for models. This problem is particularly hard because it requires many skills, like understanding context, reasoning logically, and validating answers. It also demands recognizing minor differences between incorrect but reasonable answers and true ones. The authors aim to address these challenges to create more reliable NLP systems and advance the field for future technologies.

How?

The authors extended SQuAD 1.1 to create SQuAD 2.0 by adding unanswerable questions along with answerable questions about the context. That way, a model will be able to recognize questions which are not supported by the context and return "no answer."

The crowdworkers created 53,775 questions. These questions were designed to look like they could fit the context text, but their answers did not appear in the context. Reasonable but incorrect answers were also included to mislead the reader. This approach ensures that the systems are tested for their robustness against misleading information.

SQuAD 2.0 combines answerable and unanswerable questions, thus creating a challenging training and evaluation environment for models. This structure is balanced to require systems to generate correct answers and handle contexts filled with incorrect or incomplete information effectively.

Each question was reviewed by at least one other person to check its correctness, thereby increasing the consistency and reliability of the dataset. Questions were divided into different types, such as negation, name changes, and impossible conditions, to enable a detailed analysis of their impact on systems.

Other work was done by the authors who considered few top models (among them BiDAF and DocumentQA) testing on SQuAD 2.0. This set of experiments shows difficulty either at a dataset level or how models overcome these difficulties. In particular, it has been shown that this SQuAD 2.0 represents a far more useful evaluation compared with one based on automatically produced

unanswerable questions.

Overall, the SQuAD 2.0 is not satisfied with just models finding the correct answers; rather, it wants them to also be able to detect questions whose answer is not supported by the context, so that such models avoid generating incorrect answers.

Wow!

What the authors have done highlights the weaknesses of current models and shows where improvements are needed. For example, a powerful neural network model scored 85.8% on the SQuAD 1.1 dataset but only 66.3% on SQuAD 2.0. Humans, on the other hand, scored 89.5% on SQuAD 2.0, making the big gap between human and machine performance clear. This shows there is still a lot of work to be done to improve these models in understanding language and making sense of context.

SQuAD 2.0 makes testing models even harder by including tricky or misleading situations. Human-created unanswerable questions are much harder for models to handle compared to those made by automated methods, like TF-IDF or rule-based systems. This makes the dataset a stronger challenge, pushing models to develop deeper understanding of context instead of relying on pattern recognition.

Further

But...

Although the authors present the SQuAD 2.0 dataset as a way to address a key gap in reading comprehension systems, this work does have some limitations. For example, while unanswerable questions created by crowdworkers add variety to the dataset, there could still be concerns about quality and consistency. The different skill levels, language abilities, and ways of approaching the task among workers might affect how similar the questions are overall. Some workers may create more detailed and complex questions, while others might produce simpler ones, which could cause inconsistencies in how models are evaluated. To address this, a stronger quality control process could help improve the overall quality of the dataset.

More...

If I were to take this line of research further, I would work on building new datasets with more realistic and complicated scenarios. I would create multimodal datasets that incorporate non-textual modalities, such as images and sounds, that will enable language understanding systems to achieve much deeper contextual understanding. Developing a process to incorporate user feedback into system performance measurement can further enhance the reliability of such systems for real-world applications.

All...

This study is very helpful in showing the weaknesses of NLP systems and suggesting ways to improve them. Overall, it is a well-organized and valuable paper. The main take-home message from this work is that any reading comprehension system should be able to find where the answer is or recognize if a question cannot be answered. For NLP systems to work well, this ability has become an important requirement in real-world applications. While the authors point out the limitations of NLP, they also offer a clear and useful guide for future research in this field.

Relevance

What I learned in this course about how NLP systems work definitely helped me understand this paper. More importantly, it made me think deeply about how ML systems are used in the real world and the challenges they face in practical applications. This paper clearly shows how important it is for ML systems to have the ability to say, "I don't know the answer to this question," a skill that even we as humans should value in our lives. It highlights how this ability can be crucial for ML systems as well.

While reading this paper and preparing this review, I explored the authors' websites, forums like StackOverflow, Wikipedia pages, and various online resources for additional information which was helpful for expanding my overall ML knowledge.