Reflection

Research Methods in Software Engineering and Computer Science

Arpit Garg
Faculty of ECMS
The University of Adelaide
Adelaide, Australia
a1784072@student.adelaide.edu.au

1. What was your research project?

My research project was to generate stories from visuals (sequence of images or videos) automatically. There is not much research has been done on this topic. That is why we have the limited resources available for the research. We have several networks and techniques from which we can generate captions of the images, but these captions are not the stories as the caption is generated for every image independently. In this project, we have tried to generate narrations while considering the previous images. We have compared three different best techniques to perform this. We have used all different approaches present and combined them with some heuristic approach to achieve the desired results.

Generating automated stories could be a revolutionary change in the industry. It would help the machine to narrate stories like humans. Narrations are the crucial attribute of any living being to convey information, and providing this behavior and intelligence to machines would bring the machines one step closer to humans. Apart from this, it could be used as an asset in different industries. It could work as an amplification and verification system to the narration; It could be installed with devices to help differently-abled persons and much more.

We have tried to implement this project by following the pipeline. We have first collected the data and make changes in the data according to the requirements. Then we analyze the data and model our network in which we have implemented various techniques. At last, we have applied our heuristic approaches to achieve the scores and desired results.

The pipeline starts with collecting and constructing the data according to our requirements. We have used an open-source VIST dataset (Visual Story Telling dataset) provided by Microsoft for research and trial purposes[1]. In the mentioned dataset, we changed the structure like we have changed all specified names with the gender, and all the videos are converted to frame. Then we start by analyzing the data. In this part, we have compared three different categories to generate stories and to improve the results, and we have included the global and local attention-based mechanism that achieved state of the art for the translation of languages in real life[2]. After generating the results to improve the structure of the narrations, we have implemented a greedy and heuristic approach and compared all using the statistical approach of meteor score [3]. These methods are taken from different research and applied to this project to automatically generate human-like stories.

2. What did you learn about doing research as part of your project that you didn't know before?

There are many things that I did not know are possible before starting this project. I have only heard that machines can generate stories but not sure about its implementation. Through this research, the things that I learned are:

The most important thing I learned is how to implement the project following the necessary steps. We need to start with some research, then data collection, then data analysis, then statistical comparison to see the results at end converting this whole implementation to the research. If we do not follow these steps, we would not be able to achieve the desired results. This project is the first time I have implemented a model systematically and show the real-life usage of the implementation and model.

I have also learned about the different models. I did not know about the attention-based translation techniques and research done by Microsoft and Stanford on automated text-generation. I have implemented all these things together for this research. I learn how to adapt to various terminologies and use them for the specific purpose of your research.

Before this research, I was unaware of the different statistical models available and how to compare these statistical models to select the best model according to the requirements of our research.

I did know mathematical formulas like Bayes' theorem and the law of total probability, but I did not know about its usage in real life. By doing this project, I get to know about the importance and usage of various mathematical formulas.

In this project, I have also learned different types of neural networks possible and their modified versions. Previously I knew about limited neural networks, but now I have implemented various modified advanced neural networks. So, I got an understanding of the working of these neural networks.

Above mentioned are the few main learnings from this research, but there are other learnings also like how to modify and tune models, how to create a pipeline and connect different models, how to optimize the network and improved my ability of research.

3. What would you do differently next time?

In this project, the model is trained on the VIST dataset, so it is giving better results if we work on this dataset. In the future, I would like to train this model on more other sources of data like the dataset launched by Stanford or OpenAI for text generation.

Next, I would try with YOLO (You Only Look Once). They have recently launched its new version, which contains a pre-trained model for object detection. By using this, we can simply detect the objects in a fraction of seconds and compared it with the albums to generate narrations.

In our network, we are using the validation set to tune the hyperparameters. This process is costly in computation and very time-consuming. Recently, OpenAI launched GPT-3, which contains more than 175 billion trained hyperparameters, and they are helping NLP models to reach the state of the art by reducing computation and time. I would prefer to use this on our model to improve the results.

We have used ADAM optimizer for this research due to its popularity in improving scores in various neural networks. I want to implement different advanced and other versions of optimizers (like NADAM.) and select the final optimizer after the comparisons[4].

In our research, our model is generating automated stories, but it requires a minimum number of frames to generate stories. I would try to tune the model so it could generate the story independent of the number of frames.

For the scores, we have used automated scores and improved the results by using a greedy and heuristic approach due to time limitations. This process is not the best approach in real-life. We need to work on this to improve our method next time.

All mentioned scores and findings in this are based on validation data, which is specially constructed in the data collection and construction step for this model. Next time, I try to work on the findings and tune the model accordingly.

After implementing all these mentioned techniques, I would love to implement our model on real-world videos and visuals. For this research, we have tested our model only on the constructed dataset.

We have not considered all these points in this research due to time and computation limitations, but these are the essential steps that need to consider next time.

4. What is your advice to someone who is going to work on a similar project?

Before starting a similar project, I would suggest first to do some prerequisite readings like research and advancements in computational linguistics by Ledeneva et al. [5]. It would help to clear the basics and gives the rough path that needs to follow while performing the project.

If someone is interested in similar projects, I would recommend the one to open to all approaches possible and try to implement a project using various models possible and compare based on the structure of the sentence. One should not restrict themselves with the mentioned model; one should combine various models into one model and should observe the changes in the result.

The whole process is very time consuming and expensive in computation; I encourage them to start their project by using pre-trained models like OpenAI GPT-3 and YOLO and use the cloud services to improve the efficiency of the model and saves time of the user also.

Another suggestion would be to research the related work in-depth as this type of work requires a high amount of tuning the model. If we are not on the right path, the model could get worse each time with the tuning of parameters. It would be great to use some of the recent techniques of benchmarking and random guessing presented by Oller et al. [6].

As we know, the process is very time consuming and needs patience while training models, and we need to compare different modes for similar kinds of projects. It's better to use some of the NLP human language toolkits to ease our work [7].

If one is successful in creating the similar model for the project, I would recommend to train the model on the generalized dataset, rather than only on the one specified dataset and at the time of testing, test the model on real-world visuals and videos to observe the real usability of the project.

Most of these points include the process to reduce computation and time because it is the main thing that we need to keep in mind while creating any model. Any model which takes a significant amount of time and computation power, generally makes the result and usage less efficient.

REFERENCES

[1] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollár, P. Microsoft COCO: Common Objects in Context (2014).

[2] Sutskever, I., Le, Q., Vinyals, O. and Zaremba, W. Addressing the Rare Word Problem in Neural Machine Translation. *arXiv.org* (2015).

[3] Young, P., Lai, A., Hodosh, M. and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2 (2014), 67-78.

[4] Bock, S., Goppold, J. and Weiß, M. An improvement of the convergence proof of the ADAM-Optimizer. *arXiv.org* (2018).

[5] Ledeneva, Y. and Sidorov, G. Recent advances in computational linguistics. *Informatica*, 34, 1 (2010), 3.

[6] Oller, D., Glasmachers, T. and Cuccu, G. Analyzing Reinforcement Learning Benchmarks with Random Weight Guessing. *arXiv.org* (2020).

[7] Zhang, Y., Zhang, Y., Bolton, J. and Manning, C. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv.org* (2020).