A Visual Content-Based Approach for Automatic Evaluation of Student Assignment Reports

Jiajun Huang, Shuo Yang, and Chang Xu.
School of Computer Science, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia
Corresponding Author Email: c.xu@sydney.edu.au

Introduction

The last few years have witnessed a surge of interest in artificial intelligence (AI). There is a significant increase in the number of students who major in or take courses related to AI. This inevitably brings great pressure to unit coordinators. Especially when the deadline of assignment approaches, dozens even hundreds of assignment reports will flood in almost simultaneously. The sharp rising of workload will cause a tremendous burden to the lecturers and tutors who are responsible for marking the reports. As a result, a preliminary evaluation or a rough assessment of the reports before marking is needed to propose some referential suggestions and boost the marking process.

Another severe challenge in marking a large volume of assignment reports is the marking consistency issue which has been previously discussed by Blok (1985). Due to the diversity of report markers, there is no guarantee of a consistent assessment of all the assignment reports, even if a common evaluation criterion is provided. This inconsistency problem is more serious for the novice markers. With relatively less experience, they are prone to make a biased judgment according to their subjective viewpoints. Thus, for a consistent and fair review of the student assignment reports, a more objective approach is urgent to assist the report marking.

“Whoever started the trouble should end it.” The extra burden caused by the fervour of AI should be released by AI. Because of the powerful learning and generalization ability, machine learning (ML) and deep learning (DL) models have been applied to a wide range of applications including those in the education sector. Currently, the examples of intelligent education include but not limited to Intelligent Tutoring Systems (ITS) (Chaudhri et al. 2013), virtual facilitators and learning environments (Swartout et al. 2013) and so on. However, as more time-consuming and tedious jobs, the report assessment and analysis remain under-investigated via ML/DL approaches.

In this paper, we introduce deep learning techniques to analyse the qualities of the student assignment reports based on their visual appearance. By training a deep convolutional neural network (He et al. 2016), we can judge whether a report is of good quality or not. To further explore the influence factor of the report quality, we then resort to Generative Adversarial Networks (GAN) (Jolicoeur-Martineau, 2018) to generate the sketch of good-looking reports. Both experiments received convincing and constructive results. The dataset used for this analysis is composed of marked student assignment reports in units of the University. We evaluate the proposed deep learning algorithms on this collected data, but our approach can be straightforwardly deployed in other units. Besides, because our model is trained over reports marked by multiple tutors, the subjective bias from individual marker is expected to be neutralized. We hope our work could help the report markers to improve the marking efficiency and make more objective evaluations. Meanwhile, we also provide tips for the students to improve the quality of their assignment reports.

Related Work

In this section, we review existing literature related to automated report analysis. Based on our two major research questions, the related work is divided into two parts, which are automatic grading and consistent assessing.

Automatic grading

To increase the efficiency of manual grading, researchers have developed some automatic grading approaches based on lexical, grammatical and syntactic features, such as Chen and He (2013), Attali and Burstein (2004) and Zesch et al. (2015). However, most of these methods have to manually design features to describe the material and do not have a general view of the grading material. Some other methods try to evaluate the paper quality by visual features. Bearnensquash (2010) and Huang (2018) predicted the acceptance of computer vision conferences (e.g., CVPR, ICCV) papers by ML and DL models, respectively. Nonetheless, compared with the papers submitted to these conferences which are under a clear template and length limitation, the student assignment reports are usually accomplished in a freestyle, which could interfere with the automatic marking accuracy.

Consistent assessing

The inconsistency problem generally occurs when multiple markers are responsible for massive marking assignments (Blok, 1985). Bird and Yucel (2013) and Buskes and Chan (2018) outlined some helpful suggestions to promote the marking consistency, such as training the markers before actual marking, establishing detailed marking rubrics and criteria, and exchanging knowledge of standards between different markers. However, the training process or reaching a consensus could be time-consuming and violate efficiency. On the contrary, our well trained deep learning algorithms take only very little time to evaluate new reports. Furthermore, due to the huge capacity of DL models, our approach can learn invaluable knowledge from the training data, which can provide objective referential assessments.

Assignment Report Analysis

This section introduces the details about how we collect the data to train and evaluate the proposed algorithms and further investigates the influence factors of the report quality.

Data collection and processing

We collected student reports and their corresponding marks by tutors from two assignments of one unit in two years. After eliminating the invalid reports, we finally have 220 reports in the form of PDF files as our raw data. However, these PDFs cannot be directly thrown to any deep learning algorithms. Thus, we first converted them to images by using the off-the-shelf tool pdf2image. Because of the high variance of the report length, we also need to resize the report to uniform image size. According to the statistics, the length of a report is typically 10 to 20 pages, so we organize every report image as a 3×5 grid with 3300×4250 pixels. We simply eliminated the redundant pages after page 15 and padded the insufficient with blank pages. In this way, we could keep the major visual information of reports despite a little loss. For different assignments, the marking metrics are slightly different. For consistency, we map all the marks to a hundred-score system. We label the reports whose scores are higher than 85 with rank A, between 84-65 with rank B and the remaining parts with rank C. We randomly selected 10% of the collected assignment reports to evaluate the performance of the algorithm, while the remaining reports were used to train the algorithm.

Ranking assignment reports

Intuitively, the student report assessment problem can be formulated as a regression task given the reports for training and their corresponding scores marked by multiple tutors, i.e. directly predicting the exact marks of reports, or a multi-class classification task, i.e. categorizing reports into predefined classes (e.g. HD and DI) based on their qualities. However, it could be difficult to obtain an accurate predictor for either of these two tasks, given the limited number of assignment reports collected for training the algorithm. Instead, we transformed this problem into a pairwise ranking task which compares two reports. If the model can correctly discriminate the ranking of each report pair, the relative quality of all reports can then be determined.

Suppose that we have collected N student reports of an assignment. Each report is stored as an image, and is associated with its score marked by the tutors. We first orderly encapsulate two randomly selected reports into a sample pair. In this way, the training set size can be expanded to $N(N-1)/2$. By comparing the scores of two reports, the sample pair can be annotated with 0, 0.5 or 1, which implies that the score of the first report in the pair is greater than, equal to or less than the score of the second report, respectively. This kind of annotation can be viewed as the ground-truth labels of report pairs for the purpose of training a ranking model.

The ranking model is taken as a report score estimator that aims to approximate the relative score of the input report based on the visual features extracted from a deep neural network (e.g., Resnet18 (He et al. 2016)). After the relative estimated scores of the report pairs are derived by the ranking model, their relative score difference can be scaled to $(0, 1)$ by the *Sigmoid* function (for the reason of matching the ground-truth ranking score of the report pair). Thus, the discrepancy between the estimated and ground-truth ranking scores will serve as the error signal back propagated to the ranking model via gradient descent method to update the parameters in the deep neural networks.

Generating assignment reports

Generative adversarial networks (GANs) are deep neural networks for generative modelling. The word “generative” implies that the task of the GANs is to create something of its own, mostly images but other modalities including audio and text have been done. There are two deep networks in GANs: the generator and the discriminator. The generator network tries to generate realistic data, and the other discriminator network tries to discriminate between real data and data generated by the generator network. The classification loss of the discriminator can then be taken as a supervision signal for the update of the generator networks to generate data that starts to look more realistic. GANs have been widely studied in a variety of applications, including creating lifelike images, aging or de-aging human faces, and colorizing photos.

In this paper, we introduce GANs to generate the images of student assignment reports, so that we can more easily analyse how a deep neural network processes and understands these reports. In particular, we sample some noise z from a normal or uniform distribution. The generator network G can be used to generate an image $x = G(z)$. Conceptually, z represents the latent features of the images to be generated, e.g. the colour or the shape. The semantic meaning of z does not matter, and we will implicitly discover its meaning by plotting the generated images. Given the newly generated images and real images, GANs build a discriminator network to learn what features make image real. The generator network will receive feedback from the discriminator network to make updates of the parameters in the network.

The major aim of the discriminator in classical GANs is to distinguish whether the input image is real or generated, and thus a binary classification loss is usually applied to train the

discriminator. However, in the analysis of student assignment reports, we care more about what factors make the report better. Instead of following the conventional binary classification loss in GANs, we suggest a ranking approach to re-formalize the GANs. In particular, we sample some excellent reports as positive examples. The discriminator network is expected to rank these real reports above those newly generated reports by the generator network in GANs. On the other hand, the generator network aims to generate a report that would be ranked higher than our sampled real report by the discriminator network. Regarding this ranking loss, the generator network and the discriminator network will compete with each other, and the quality of reports generated by the generator network will be continuously improved as a result.

Results

We examined and discussed the results of the proposed algorithms. The architecture of the neural network used in our experiments was adapted from Resnet18 (He et al., 2009).

Ranking assignment reports

In our experiments, we are most likely interested in identifying the top-ranked excellent student assignment reports. So, it makes more sense to have evaluation metrics over the top k reports instead of all the reports. We used Precision and Normalized Discounted Cumulative Gain (NDCG) at k as the evaluation metrics for the resulting ranked student assignment reports. In particular, Precision at k is the proportion of reports in the top-k set that are HD. In NDCG, to compute how good the returned ranking report list is, each report has a relevance score (e.g. HD, DI, and PS). That is gain. For bad reports, we usually set the gain to zero. By adding up those scores, we have cumulative gain. To see the most excellent reports at the top of the list, therefore before summing the scores we divide each by a growing number (usually a logarithm of the report position) – that is discounting.

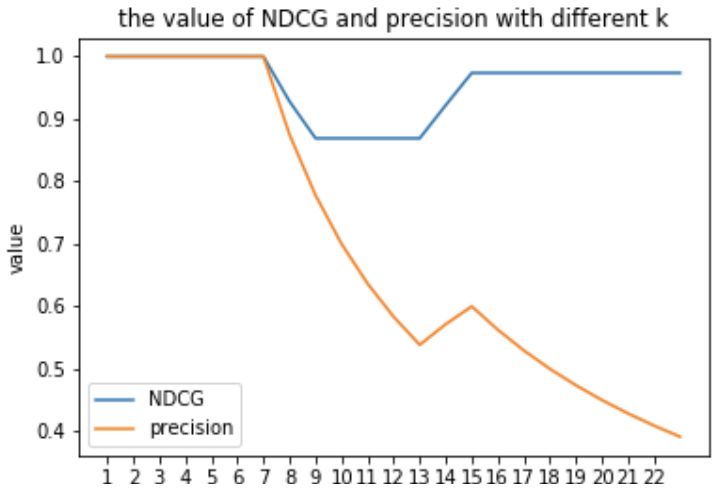

Figure 1: The value of NDCG and Precision at k.

Figure 1 shows the ranking results of student assignment reports in terms of NDCG and precision values with different k. As we can see, our algorithm can achieve very higher NDCG and Precision values when we consider the top 8 or 9 assignment reports. That is to say, our algorithm can well identify the excellent assignment reports (i.e. HD) and successfully rank them at the top of the returned ranking list. NDCG and Precision values begin to drop when we increase the value of k. It may because the delineation between the excellent and the moderate reports is not sharp enough. The precision value will continue to drop, as there will be more bad reports introduced in the list by increasing k. In contrast, due

to the discounting effect, the NDCG score increases slightly again and becomes stable when the k is larger, indicating that the algorithm can well rank those bad reports.

Discriminative regions of reports for ranking

We are also interested in the characters of an assignment report to get a higher mark. In other words, we plan to analyse which part of the report tends to push the neural network to rank the report higher up. We believe that by exploring the important part of the report, it could provide a guideline for students to better prepare a good report.

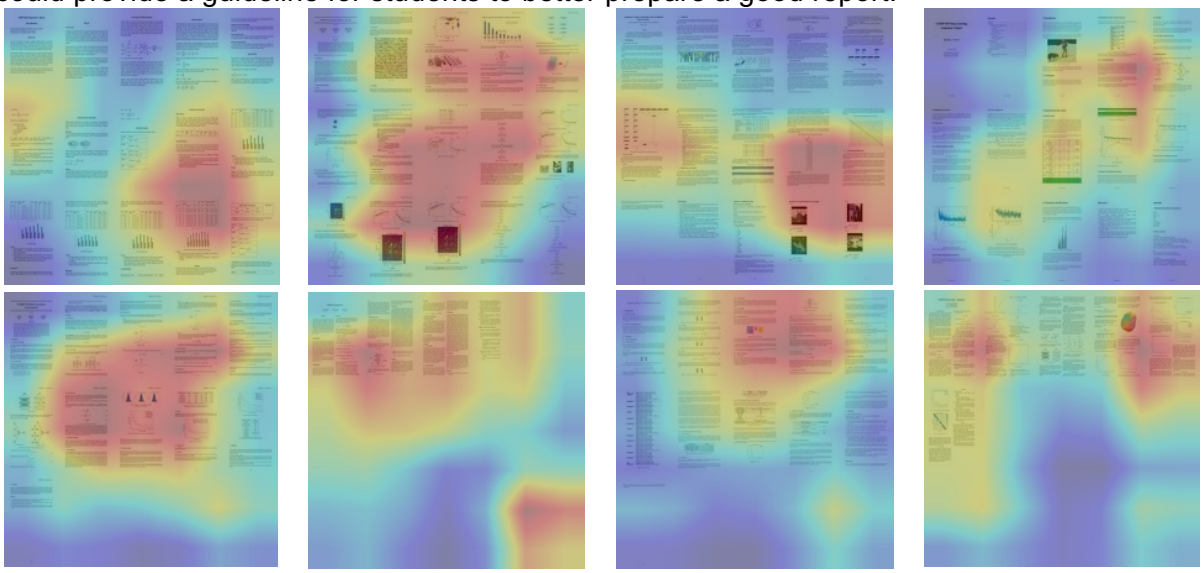

Figure 2: Example heatmaps of assignment reports.

Visualizing the neural network could be a good way to understand how it can make the decision. We selected the Class Activation Map (CAM) (Zhou et al., 2016) as the visualization method. CAM produces the class-specific heat map figures that could indicate which part of the input helps the network to predict the probability that the input belongs to one specific class.

In Figure 2, we report some example heatmaps for the report images. The colour indicates the attention of the neural network to make the decision. Low values tend towards cool blue tones while higher values tend to hotter orange and red tones. In general, our neural network mostly focuses on the area containing figures or tables. Hence in the opinion of the neural network, a report of high quality should contain not only the words but also a variety of figures and tables. In the second row of Figure 2, the neural network also pays attention to the blank area of the reports, which implies that the page length of the report could also be an important feature to evaluate the report.

Sketches of good reports

Beyond ranking student reports and highlighting discriminative regions, we proceed to provide further suggestions to help students enhance their reports. To this end, we resorted to generative adversarial networks to generate visual layouts of a good report. A state-of-the-art GAN method can generate meaningful images of reports. We selected top-ranked reports (i.e. HD reports) as the training data. All the training data were rescaled to 256 x 256 pixels and it took around half a day to finish the entire train process. We then can generate images that follow the same distribution of the given training data (i.e. HD reports).

Figure 3 shows eight random generated reports by the GAN method. We can find that these generated good reports often enjoy a balanced layout of figures, tables, plots or equations. Note that, the generation of fine-grained reports with high semantic information is time-

consuming and difficult. Thus, we only generated sketches here to see what visual characteristics good reports typically have. The visual quality of these generated reports could be further improved if more training data are provided.

Figure 3: Examples of generated good reports.

Discussion

The ranking and generation results show consistent phenomena which indicates that a high-quality report is expected to have elaborate figures and illustrative tables, while the length of the report is also a significant factor for assessment. These make sense in the practical marking of tutors. The students are expected to demonstrate the details of the methods and the modules used in the model. Thus, plotting illustrative figures is an elegant way to describe the major idea of the report. The experimental result is another crucial part of an engineering paper or report. Well-designed charts and tables are clear demonstrations of contrastive results. Besides, the students need to elaborate each of the required elements of the report using words and formulas, and a sufficient page length would be a good proof of the student's substantial work. It is worth noting that while the above points are necessary for a report to be comprehended by the grader so as to get a high score, we emphasize that the grading system can only give the grader preliminary opinions, and the grader still plays a major role in judging the quality of the content of the article.

Although we have achieved convincing results so far, some imperfections still exist and need to be further improved. For example, the final mark of a report usually consists of individual marks of different parts of the report. It is therefore a more reasonable way to predict the separated marks w.r.t. different parts of a report. Besides, our work aims at providing a summary assessment of the overall report quality and we do not expect the model to be confined to few units. Hence, we focus on visual content rather than the textual content.

Based on the discussion above, our model can not only benefit the markers, but also for the students. By analyzing their reports through this system, the students can receive suggestions on how to refine their reports. We also plan to adopt online learning strategy to improve our system constantly.

Conclusion

In this paper, we developed a visual content-based deep learning approach for automatic student assignment report evaluation. For training our model, we collected student

assignment reports with tutors' markings to construct a dataset. Different from a classical neural network for classification, we obtained a deep neural network to rank assignment reports. Besides, with the help of generative adversarial networks, we received a generator network to synthesize assignment reports that follow the same distribution of good reports prepared by the students in the real world. Experimental results demonstrate the effectiveness of our approach, from which we conclude the importance of a balanced layout of figures/tables and page length of the report to influence the marking of a report. Our methods in this paper could be beneficial to improve the efficiency and consistency of large-scale report marking, even for tutors or lecturers who are not familiar with deep learning techniques. Most importantly, by analyzing the experimental results, we conclude useful tips for students to well prepare their reports.

References

Blok, H. (1985). Estimating the reliability, validity and invalidity of essay ratings. *Journal of Educational Measurement 22(1)*, 41-52.

Chaudhri, V. K., Lane, H. C., Gunning, D., & Roschelle, J. (2013). Intelligent learning technologies: Applications of artificial intelligence to contemporary and emerging educational challenges. *AI Magazine, 34(3)*, 10-12.

Swartout, W., Artstein, R., Forbell, E., Foutz, S., Lane, H. C., Lange, B., Morie, J., Noren, D., Rizzo, S., & Traum, D. (2013). Virtual humans for learning. *AI magazine, 34(4)*, 13-30.

Huang, J. B. (2018). Deep Paper Gestalt. *arXiv preprint arXiv:1812.08775*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Chen, H., & He, B. (2013). Automated essay scoring by maximizing human-machine agreement. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1741-1752).

Attali, Y., & Burstein, J. (2004). Automated essay scoring with erater® v. 2.0. *ETS Research Report Series, 2004(2),* i-21.

Zesch, T., Wojatzki, M., & Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. *In Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 224-232).

Von Bearnensquash, C. (2010). Paper gestalt. *Secret Proceedings of Computer Vision and Pattern Recognition (CVPR)*.

Bird, F. L., & Yucel, R. (2013). Improving marking reliability of scientific writing with the Developing Understanding of Assessment for Learning programme. *Assessment & Evaluation in Higher Education, 38(5)*, 536-553.

Buskes, G., & Chan, H. Y. (2018). Implementation of marker training exercises to improve marking reliability and consistency. *In 29th Australasian Association for Engineering Education Conference 2018 (AAEE 2018)* (p. 92). Engineers Australia.

Jolicoeur-Martineau, A. (2018). The relativistic discriminator: a key element missing from standard GAN. *arXiv preprint arXiv:1807.00734*.

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A. & Torralba, A. (2016). Learning Deep Features for Discriminative Localization. CVPR (p./pp. 2921-2929): IEEE Computer Society. ISBN: 978-1-4673-8851-1

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).