Paper Review System

Ruoyu Cheng
Shaghai Jiao Tong University
Email : roy_account@sjtu.edu.cn

Yixiong Wang
Shaghai Jiao Tong University
Email : wyx_ei@sjtu.edu.cn

Jiasong Guo
Shaghai Jiao Tong University
Email : 7_GUO_7@sjtu.edu.cn

ABSTRACT

Over the past few years, the number of papers submitted to conferences such as CVPR, AAAI has surged. The large number of paper submissions and the shortage of reviewers have placed a huge burden on the paper evaluation system. For the fact that some existing methods based on machine learning are weak in interpretability, we propose seven indicators to assess a given paper. They are : number of figures, tables, formulas, which are measured page-wise and concatenated to form a vector, number of references, frequency of the hottest words, paper image as a whole and paper contents in this paper. Experimental results show that we achieve 99% overall accuracy, 6.7% higher than state-of-the-art approaches based on visual features over AAAI dataset. [1]

1 Introduction

[2] In recent years, the number of papers submitted to computer vision conferences has increased significantly. Peer review is an important aspect of disseminating scientific achievements. It is a thorough review of academic work by other experts in the community, but the record number of papers submitted to top computer vision conferences and the insufficient number of qualified reviewers made the peer review process very difficult. The number of the paper submissions to top-tier computer vision conferences has been increased dramatically over the past few years. (See Figure1)

In order to review the papers of all submitters, the conference organizer must expand the scope of reviewers to include inexperienced students and inexperienced reviewers. [2] Therefore, authors who spend a lot of time, even months or years on the research of the paper may end up with a comment that has no reason to be considered badly or unfairly. Therefore, authors who have spent months or years writing papers may end up receiving unreasonable, poorly considered or unfair comments.

At present, there are some works to use the network model such as resnet [6] to extract the features of the paper pages, trying to score the appearance. [7] However, we think that the image features of the page cannot weigh the content of the paper too well, and the interpretability of the image features extracted by CNN is relatively weak.

In this paper, we propose some interpretable features to better evaluate the paper, including number of figures, formulas and tables per page, frequency of the hottest word and numbers of references. We scored these features along with the image features and text features to arrive at the final review.

---

[1] available at https://github.com/7-GUO-7/Mobile-Internet

[2] Jiasong Guo

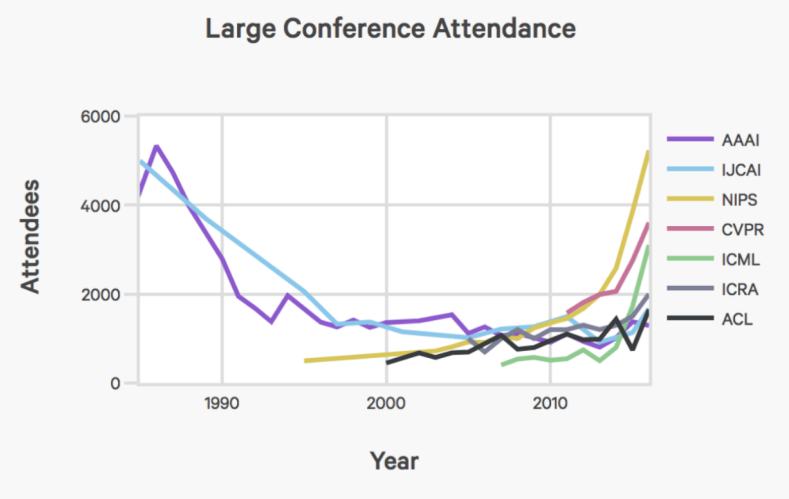


Fig. 1. The number of the paper submissions to computer vision conferences over the past few years.

2 Related Work

[3] **Academic paper rating.** Paper evaluation is a relatively new task in AI. Its basic function is to automatically predict whether to accept or reject papers. Some scholars manually extracted some features, such as the length of the title, whether it is a specific word (such as the latest and new content), the keywords that appear in the abstract, and used machine learning methods, such as logistic regression, decision trees and random forests to learn the relationship. There are also scholars who use modular hierarchical convolutional neural networks (CNN), in which each part of the paper is regarded as a module. For each paper part, they train an attention-based CNN, and apply the attention pool layer to the representation of each part, and then input it into the softmax layer. [5]

Essay scoring. Other popular methods [5] [10] [1] [4] [8]for evaluating the quality of paper articles is to directly analyze the content of the paper articles. One of the simplest solutions is to predict the quality of the article based on the length of the article. Similarly, using content, structure, network and edit history features, Anderka et al. [3] built a binary classifier to predict quality flaws in Wikipedia. The approach is based on the cleanup tags, which are provided by the reviewers who detected the flaws but do not have enough time or expertise to fix them.

Vision-based methods. Computer vision techniques have been applied to accessing the quality of actions [9], surgical skills [12], and images. The work most related to our work is that of the awesome Bearnensquash [11], where the AdaBoost algorithm is used for learning the good/bad paper classifier. Building upon the methodology in [11] that relies on hand-crafted visual features, we revisit the paper gestalt problem with deep learning and learn task-specific representation through a training process.

Administrative methods. Several administrative policies have been proposed to address the surge in the number of paper submissions. Examples include desk-reject by area/program chairs (e.g. violation of anonymity, formatting, or clearly out of scope), mandatory abstract submission one week before the paper submission deadline, expansion of the reviewer pool, and providing training materials for inexperienced reviewers [2].

---

[3]Jiasong Guo

3 Our Approach

3.1 Evaluation Index

[4] Inspired by the novel idea that number of figures, tables and other non-text parts can be an important index for paper quality evaluation. We propose seven indicators to assess a given paper. They are : number of figures, tables, formulas, which are measured page-wise and concatenated to form a vector(in our system, we extract the first ten pages features), number of references, frequency of the hottest words, paper image as a whole and paper contents.

The most important thing for a paper review system is that the evaluation indexed are interpretable, otherwise people may suspect all the results it makes. The features above we propose are convincing enough and we can naturally give comments in regarding to each feature.

3.2 Data Pre-processing

[5] The dataset we obtained contained paper images only. The positive samples are accepted papers of AAAI from 2012 to 2018, negative ones are pre-printed papers on arxiv. So we analyzed their web structures and crawled all the pdf files.

To obtain the features mentioned above, we firstly used python module tika to convert pdf files to texts, then we used regular expression to match reference part and count the total number(some of the papers do not have a reference number and we match the page number range and published year of the referenced one). Then, we deleted some of the meaningless lines including author information, wrong signs to extract the text form of the paper.

For the page-wise features, we used PyPDF2 to read every page and find all the figures, tables and formulas by regular expression. The design of the regular expressions for figures and tables are relatively easy as most of their captions start with "Fig/Figure" and "Table". However, nearly every formula is identical, what they have in common may only be the calculating signs. We score them and calculate the total score of each suspected region. If the total score is greater than the threshold, we assert that it is a formula.

The hottest words are the most frequently mentioned key words in the paper. In the first several attempts, we extracted the most frequent words among the whole texts, but what we derived was some common but useless words, like "model", "algorithm" and so on. Then we realized we need to narrow the selection range. Finally, we fixed the range to paper title and abstract, key words can only appear in this region. Besides, we considered the effect stopwords ,the upper and lower case, parts of speech, non English words and so on. We eliminated all the discrimination. This process was done by python package nltk.

3.3 Data Analysis

[6] As mentioned above, we propose some interpretable indexes and we hope these features can distinguish good papers from bad ones. After extracting all these features, we did the visualization. The page-wise feature for every paper is a 10-dimension vector under one index, indicating number of figures/tables/formulas on each page.

Figure2 shows the distribution 3-D histograms for figures and formulas. We now denote the figure feature vector as $f \in \mathbb{R}^{10}$, $f(i)$ denotes the number of figures at page i. The histogram($i \times f(i) \times n$) indicates there are n samples with $f(i)$ figures at page i. Red and blue points are positive and negative samples respectively. We randomly select 2000 samples for both set. The histogram for tables per page is similar to figures and formulas. It is easy to learn that distribution difference does exist. It preliminarily

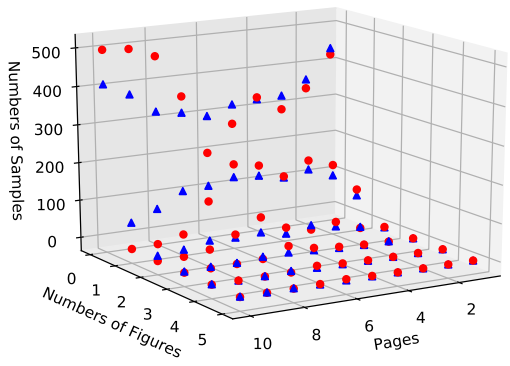
---

[4] Yixiong Wang

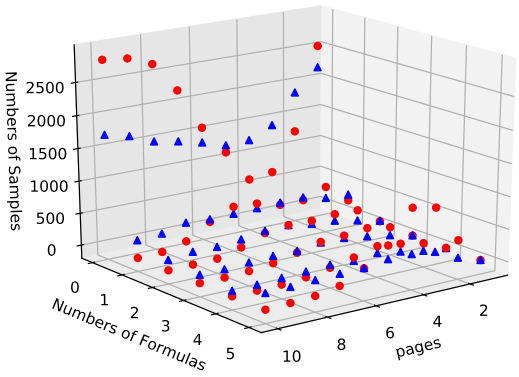
[5] Yixiong Wang

[6] Yixiong Wang

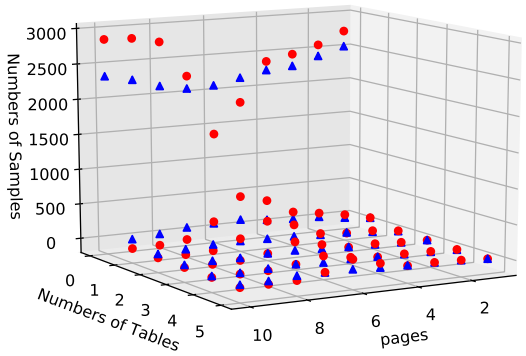
confirms our conjecture.



(a) figure



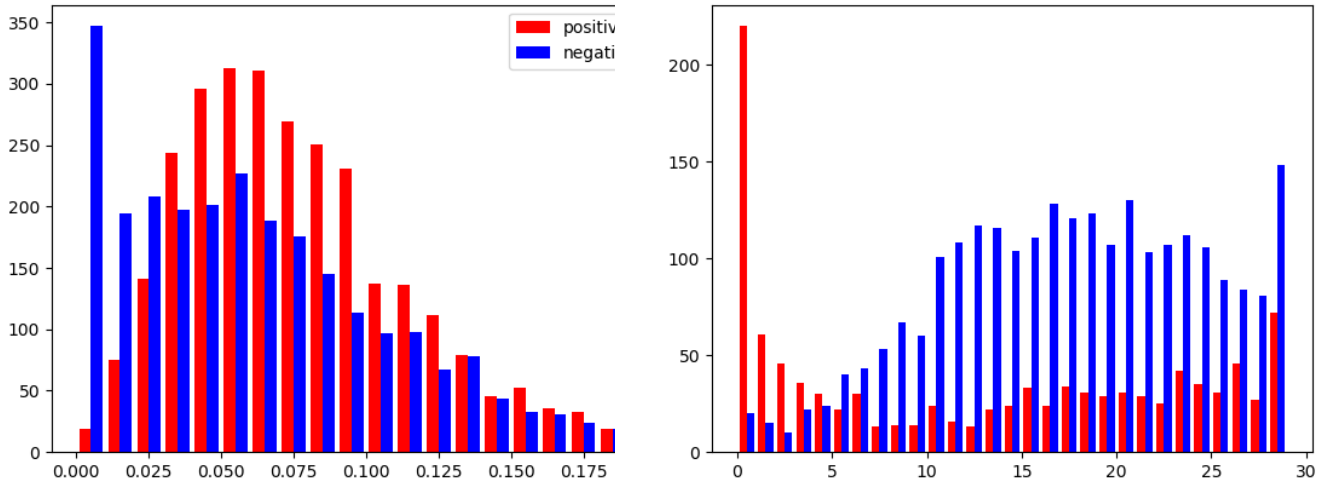
(b) formula



(c) table

Fig. 2. Distribution of figures and formulas per page. Red points are positive samples and blue ones are negative.

For the paper-wise features, we draw the 2-D histograms for hottest words frequency and number of references. As shown in Figure 3, x-axis is the frequecy/number of references, y-axis is the number of samples. Still, we can see the distribution differences between positive and negative samples.



(a) hottest word frequency

(b) number of references

Fig. 3. Distribution of hottest word frequency and number of references. Red is positive and blue is negative. Easy to learn accepted papers mentions key words more frequent and refer to more papers.

3.4 Our Model

[7] As showed in Figure 4, according to the features we extracted, we use 3 MLPs to generate scores for figures, tables and formulas. We use SVM for reference and MLP for frequency. Mention that features of figures, tables and formulas are ten-dimension tensors, whereas reference and frequency are scalars. Text feature is processed with Bi-LSTM and image feature is done by ResNet-18 and VGG-16.

Finally we use an MLP to combine all the results before to generate our final score.

3.5 Review generation

[8] In our system, review generation mainly consists of two parts: comments from different perspectives and the final score, i.e., whether it is recommended to accept the paper. Figure 5 shows the structure of this procedure.

The first part depends on the result of our seven individual models. For example, if the result of image model (ResNet or VGG) is positive, we will assume that visual rendering of the paper is relatively satisfying. Thus, we will draw a positive comment from the corresponding comments pool as our review. Similarly, the text model is corresponding to the logic and grammar perspective and the number of formula is corresponding to the mathematical analysis, and so on. The comments are not always absolutely correct and there is no direct causality between result of models and our review, but we believe that they are strongly correlated.

The second part only depends on the result of our overall model and a threshold is chosen to determiner whether we should accept the paper. Without loss of generality, the current threshold is set as 0.5, while it is a hyperparameter adaptive to the criteria.

---

[7] Jiasong Guo

[8] Ruoyu Cheng

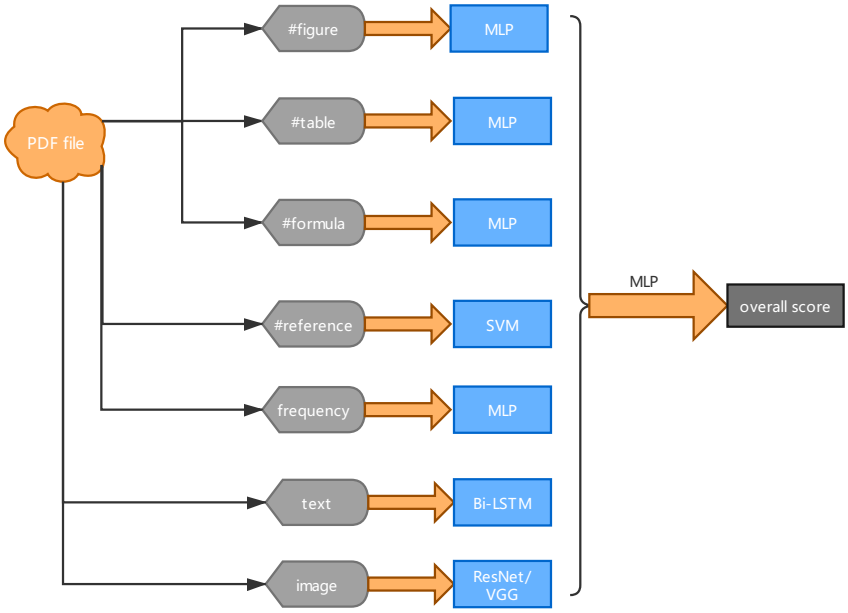
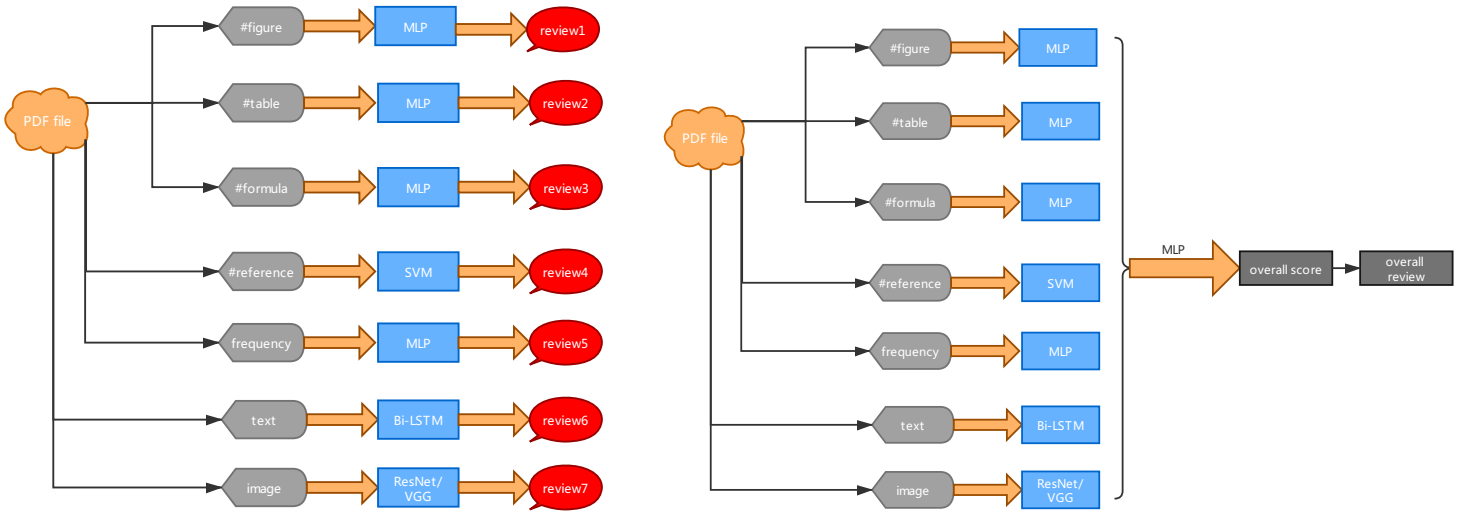


Fig. 4. Our model graph.



(a) comments from different perspectives (b) how we derive the final review

Fig. 5. The procedure to generate review from our models

4 Experimental results

[9] The training set consists of 2860 papers on AAAI Conference from 2012 to 2018 as positive samples and 2650 cs.AI pre-printed papers from arxiv as negative samples, while test set contains 715 positive samples and 696 negative samples respectively. We experiment on seven individual models and the combined model of five hand-crafted features as well as the overall model.

[9]Ruoyu Cheng

4.1 Individual Model

Image model. We adopt VGG-16 and ResNet-18 (both pre-trained on ImageNet) as backbone of the classification network. We replace the 1000 class classification head with two output nodes to distinguish good or bad papers. To finetune the pre-trained network on our dataset, the initial learning rate is set to 10^{-3} and decay it by a factor of 0.5 every epoch. The momentum and weight decay are set to 0.9 and 0.0001 respectively by using SGD optimizer. As expected, both networks achieve superior performance, with 85.3% accuracy on VGG-16 and 92.3% accuracy on ResNet-18. Note that there is a wide gap between two networks for image model, but for the overall model they are equivalent.

Text model. We first derive a sentence representation by averaging across words in a sentence, then feed the sentence representation into a biLSTM layer, and finally a fully connected layer (size = 2 as the number of classes). We use adaptive learning rate with Adam optimizer which is initially set at 0.01. Also, the dropout ratio is set as 0.5 according to recent studies. Our biLSTM model achieves 70.4% accuracy, which is over 15% lower than the image model. A possible reason is that there are too many words in a paper so that the model can not extract the key message.

Reference. We represent each paper by the number of reference and a SVM model is adopted for classification. It is worth noting that this simple model achieves 0.80 accuracy, which verifies our analysis that the number of reference is a good hand-crafted feature.

Frequency. We represent each paper by the frequency of four most frequently used key words from the paper title. A multilayer perceptron (MLP) with two hidden layers is trained for classification. By grid-search to find the optimal hyper-parameters, we set the number of hidden unit as 4 and 2 respectively. The 0.67 accuracy implies that the samples are not always separable by frequency, but it is still of great use for our overall model.

Figure. We count the number of figures in first ten pages and concatenate to a vector. Similarly, we use a MLP model with 5 and 2 hidden units and adam as optimizer. The model achieves 0.79 accuracy individually, which is also a good hand-crafted feature.

Table and **formula** are very similar with **figure**, so we will not repeatedly introduce them here. The result of seven individual models is shown in table 4.1.

	Image(VGG)	Image(ResNet)	Text	Reference	Frequency	Figure	Table	Formula
Acc	0.85	0.92	0.70	0.80	0.67	0.79	0.73	0.82

4.2 Combined Model

Hand-crafted Features. To validate the effectiveness of our five hand-crafted features, we concatenate the output of each individual model (0 or 1) and train a simple MLP with one hidden layer. This combined model achieves 0.90 accuracy, which is much higher that every individual part. It shows that our efficient and interpretable models have comparable performance with the deep learning model.

Overall. Finally, we consider all these factors and train a one-layer MLP as our overall model. For both image model we achieve a significant improvement, with 0.98 accuracy for VGG and 0.99 for ResNet. Therefore, our hand-crafted features can provide the deep neural network supplementary information and make convincing classification. The result is shown in table 4.2.

	hand-crafted features*	overall(VGG)	overall(Resnet)
Acc	0.90	0.98	0.99

5 Conclusion and Future Work

[10] We propose to jointly use neural network models to capture visual features, rnn model to capture textual features and five hand-crafted features to make a more comprehensive and interpretable model. Experimental results show that we achieve a 6.7% higher accuracy than state-of-the-art approaches based on visual features over AAAI dataset.

Although impressive results have been shown, our work suffers from the following limitations. First, our model assume that all the papers are provided the same conference template, thus we may face the generalization problem to other conferences. Moreover, the review we generate from the classification result are not very accurate. In the future, we will make more effort on the review generation step and implement a multi-conference model.

6 Work Division

Jiasong Guo : Put forward the idea of these interpretable features and implement image part(ResNet & VGG).

Yixiong Wang : PDF parsing, data analysis, interpretable multi-dimension feature extraction and data visiualization.

Ruoyu Cheng: Implementation on MLPs and Bi-lstm model, and put forward the structure of overall score.

References

[1] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. *arXiv preprint arXiv:1606.04289*, 2016.

[2] Thomas Wesley Allen. Peer review guidance: how do you write a good review? volume 113, pages 916–920. Am Osteopathic Assoc, 2013.

[3] Maik Anderka, Benno Stein, and Nedim Lipka. Predicting quality flaws in user-generated content: the case of wikipedia. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 981–990, 2012.

[4] Quang-Vinh Dang and Claudia-Lavinia Ignat. Measuring quality of collaboratively edited documents: the case of wikipedia. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*, pages 266–275. IEEE, 2016.

[10]Ruoyu Cheng

[5] Quang-Vinh Dang and Claudia-Lavinia Ignat. An end-to-end learning solution for assessing the quality of wikipedia articles. In *Proceedings of the 13th International Symposium on Open Collaboration*, pages 1–10, 2017.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Jia-Bin Huang. Deep paper gestalt. 2018.

[8] Isaac Persing and Vincent Ng. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, 2013.

[9] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *European Conference on Computer Vision*, pages 556–571. Springer, 2014.

[10] Aili Shen, Bahar Salehi, Timothy Baldwin, and Jianzhong Qi. A joint model for multimodal document quality assessment. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 107–110. IEEE, 2019.

[11] Carven Von Bearnensquash. Paper gestalt. 2010.

[12] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric L Sarin, Thomas Ploetz, Mark A Clements, and Irfan Essa. Automated video-based assessment of surgical skills for training and evaluation in medical schools. volume 11, pages 1623–1636, 2016.