Text & Vision-Fused Framework for Academic Paper Review

Final Project - EECS 498 Deep Learning Winter 2019

Yichen Yang, Tongan Cai, Shuyang Huang, Jiachen Liu
University of Michigan – Department of Computer Science

Abstract

The great boom in Artificial Intelligence these years has been leading to more and more ideas and methods, yet academic paper submissions greatly overwhelmed review committee. This Text & Vision-Fused Framework will exclude paper of lower quality with judgment based on contents, vocabulary usage and image quality with a deep-learning-based model. This framework aims to perform as an efficient and reasonably accurate filter for academic paper review process, and potentially provide scoring factors as suggestions for inexperienced authors.

Initiatives

Background

Taking IEEE Conference on Computer Vision and Pattern Recognition as example, the number of submissions is increasing at a tremendous speed:

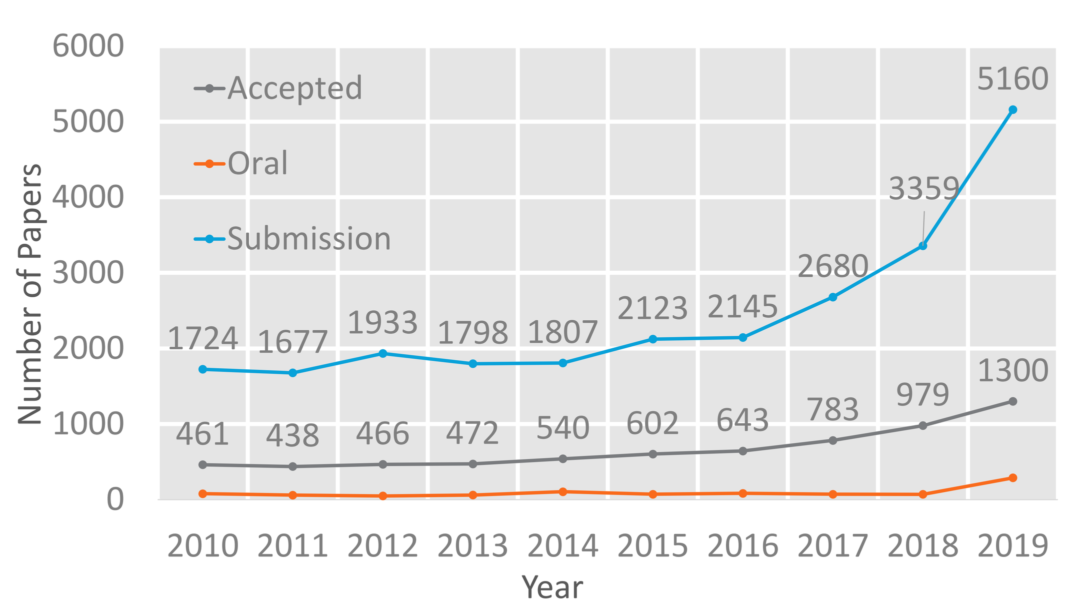


Figure. CVPR2010-2019 Number of Accepts/Orals/Submissions

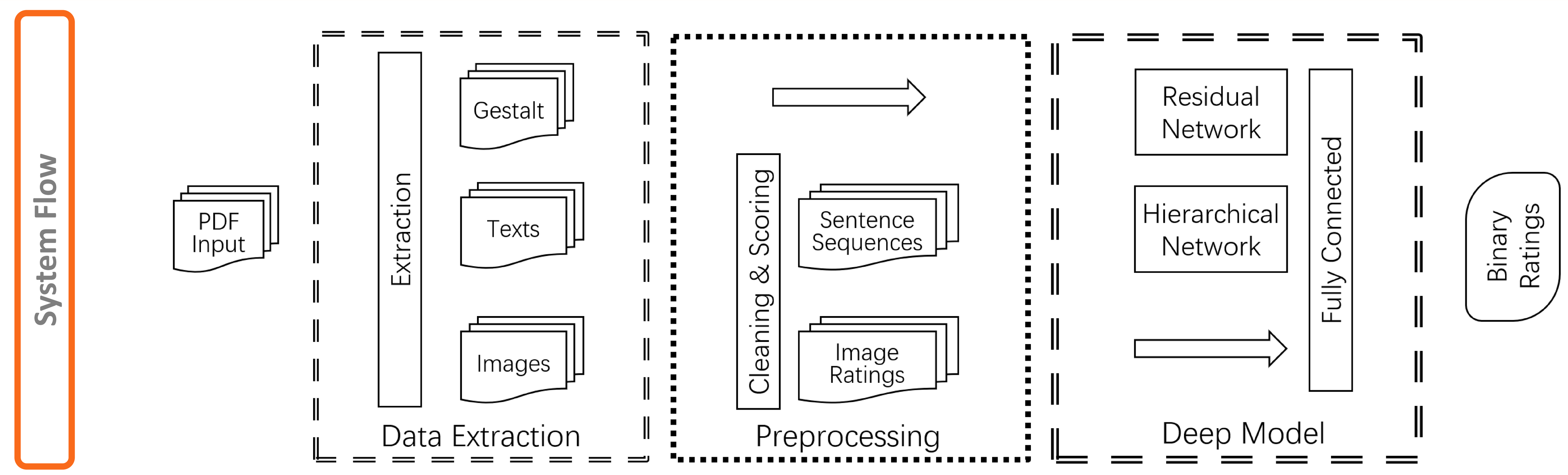
How to efficiently determine the quality of academic paper remains a demanding question. If some papers of low quality could be judged and ruled out ahead of time, the committee members will be greatly relieved.

Assumptions

Our team would like to develop a novel way to judge the quality of academic paper by adopting computer science knowledge. This framework holds 3 main assumptions:

- The quality of an academic paper greatly related to the quality of the texts and images the author uses
- The quality of an academic paper can be reflected by their overall appearance ("Gestalt")
- The quality of an academic paper can be inferred by classifying its pure text content

Methodology



Data

Our dataset covers the submissions of ICLR from 2017 to 2019. The datasets has three labels, "Oral", "Poster" vs. "Reject" based on the review on OpenReview, and we consider "Oral" and "Poster" as "Accept" category. The training set and test set are split with regard to balancing the two classes, with 2414 samples as training set and 1500 samples as testing set.

Preprocessing Logistic

The full PDF files are converted to image "Gestalt", text sequences, vocabulary sets and image sets.

- **pdf2image** convert PDF input to images of 680 × 440 as "Gestalts".
- Parse by **pdfminer.layout** build-in functions. Sequences of strings are cleaned and calculated for vocabulary/sentence statistics. Only sample sentences with more than 50 characters and 13 words.
- **pdfimages** utilities to extract images. To calculate a "Rating of image". Images that are in single color or too small sized are excluded in calculation.

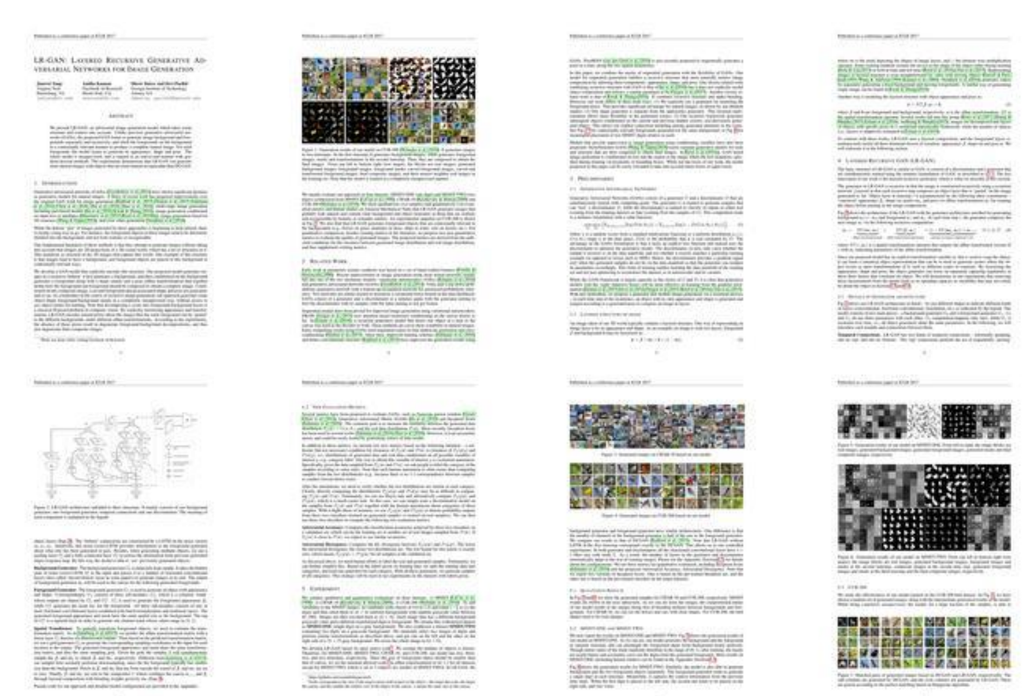
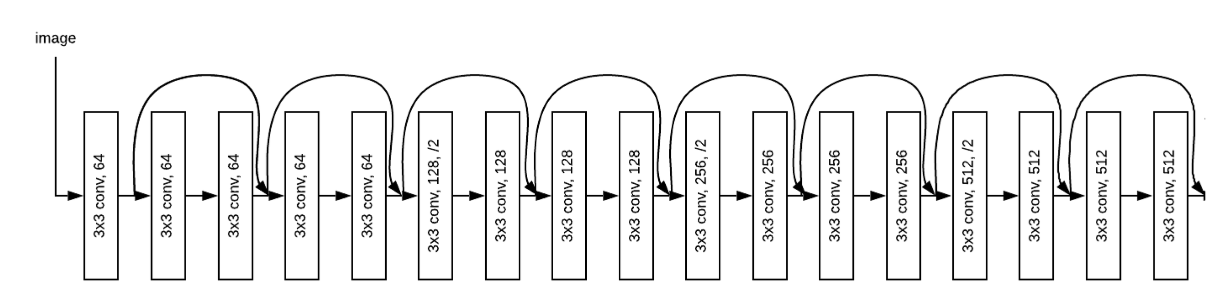


Figure. Extracted "Gestalt" and images from one data sample (ICLR 2017)
LR-GAN: LAYERED RECURSIVE GENERATIVE ADVERSARIAL NETWORKS FOR IMAGE GENERATION

Deep Model

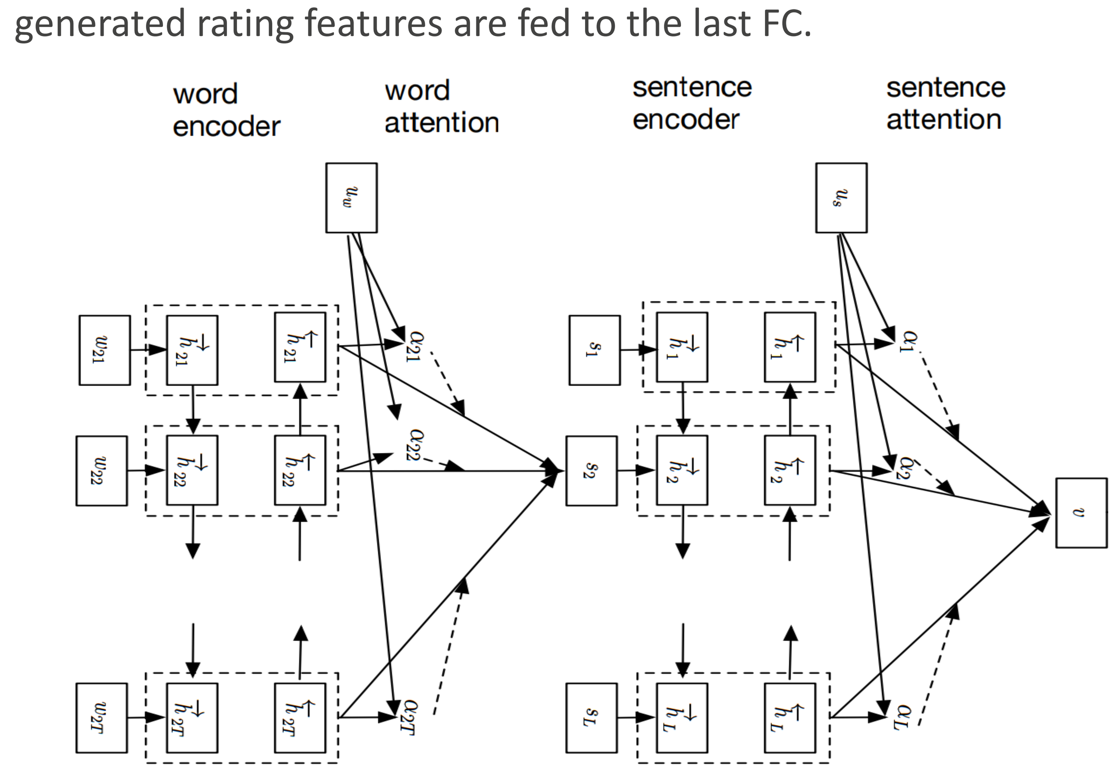
- **Residual network**

For image classifier, ResNet-18 (pretrained on ImageNet) is used for the sake of limited computational resources. We removed the original output layer. This image based classifier reads the low-resolution image "Gestalt" of the PDF which treats the paper as a whole, and generate a sequence of rating features to the last FC.



- **Hierarchical Network**

For text data, we propose a simple improvement on HAN (Hierarchical Attention Network) to synthesizes information from different paper structure levels, including sections, sentences, and words. In this way, our deep model may check the logicality of the context. The last layer (originally a SoftMax) is also removed and the generated rating features are fed to the last FC.



Result & Prospect

We trained our framework on a laptop with 4 Core CPU, 16G Ram and a NVIDIA GTX 1080 GPU. A basic CNN (Conv-ReLU-Pool-FC) and RNN are considered as baseline classifiers. The metrics we're using here are:

***Precision Rate (PR):** correct_accept / (miss_accept+correct_accept)*
***Correct Accept Rate (CAR):** correct_accept / total_accept*
***Correct Reject Rate (CRR):** correct_reject / total_reject*
***Miss Accept Rate (MAR):** miss_accept / total_reject*
***Miss Reject Rate (MRR):** miss_reject / total_accept*

Metrics	CNN (Image)	RNN (Text)	Fused CNN+RNN	Proposed Framework
PR	81.548%	60.317 %	88.158 %	**93.289%**
CAR	94.483%	54.286 %	92.414 %	**95.862%**
CRR	95.969%	64.286 %	97.659 %	**98.700%**
MAR	4.031%	35.714 %	2.341 %	**1.300%**
MRR	5.517%	45.714 %	7.586 %	**4.138%**

Table. Results (best-ever) for proposed model and baselines

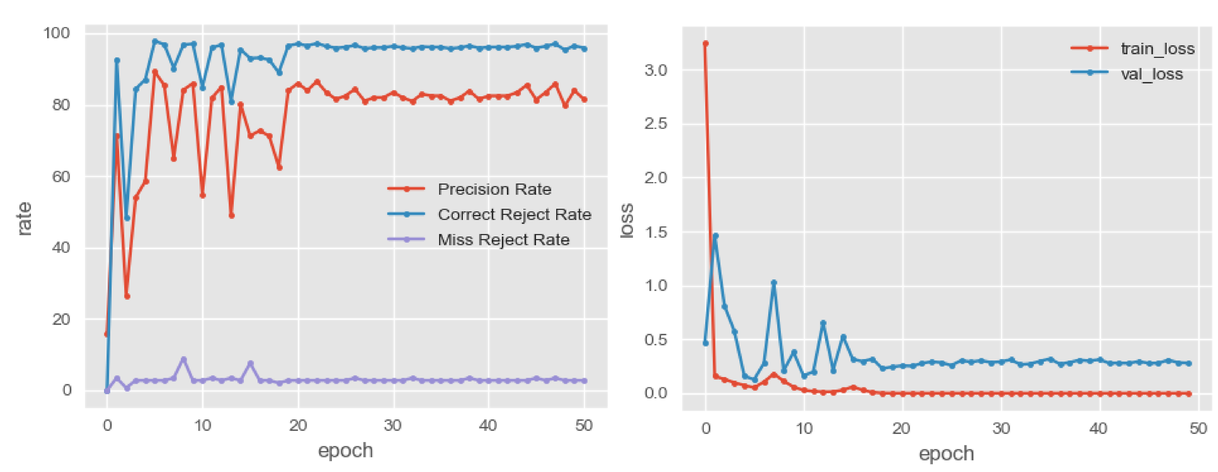


Figure. Accuracy and loss curve for proposed model

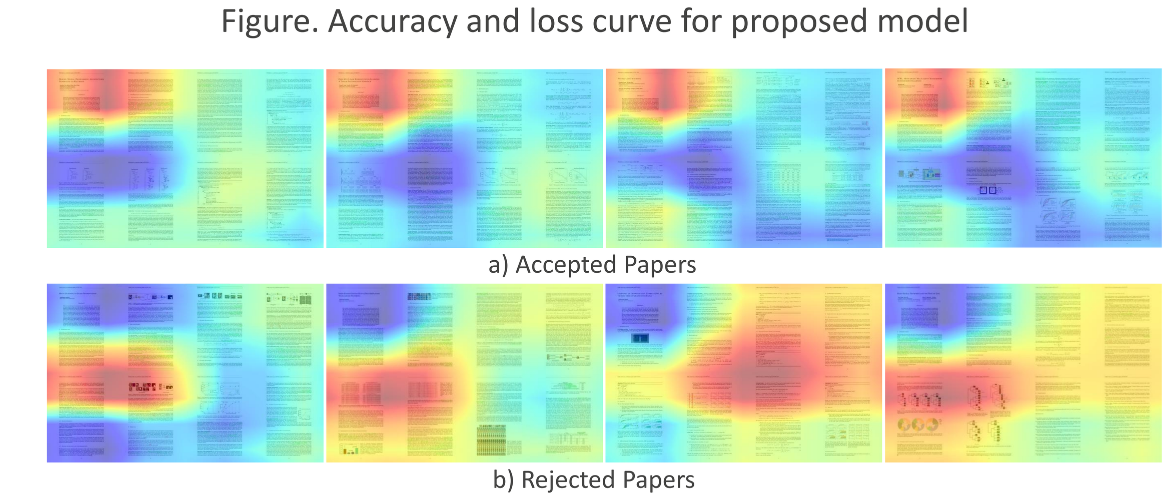


Figure. Samples with Class Activation Mapping

The above results indicates that our framework is valuable to some extent in distinguish the quality of academic papers. **98.7%** paper was correctly rejected while only **4%** was sacrificed. To our best knowledge, this is so far the **FIRST framework to fuse text & vision features of academic papers for acceptance prediction**.

In the mean time, we believe future works can improve the framework in these aspects:

- *Better text extraction quality for structural & grammar analysis*
- *Larger datasets for deeper model & better accuracy*
- *Provide scoring factors & gain interpretability*
- *Better visualization of Neural Networks*