Automated Examination Grading Using Deep Learning Categorization Techniques

Lehan Yang

March 27, 2019

Automated Examination Grading Using Deep Learning Categorization Techniques

Lehan YANG

Chengdu Experimental Foreign Language School, China
sugar.yang@xsourse.cc

In previous time, students whatever from junior or senior have lots of examination in school. As the competition for learning grows, the frequency of exams increases. It takes time for teachers to mark these exam papers, students also need to wait for a long time to get their score. This paper is focus on estimating the score by using deep learning methods. This method of scoring is based solely on the image characteristics of the answer sheet, rather than scoring through text-based methods after recognizing the text. Experimental results shows that We are able to achieve an accuracy of approximately 85% over a 10 scores error and approximately 95% accuracy over a 20 scores error. And we can use the deep convolutional neural network we trained as the discriminator of Generative Adversarial Network to generate high score papers and improve student achievement.

1 Introduction

Before taking higher education, especially in China, middle school students will receive thousands of stage exams and then enter the university with the final SAT, TOEFL or Chinese college entrance examination. More and more students join these test (see Figure 1), it takes time for teachers to rate more papers. In these exams, students usually have to wait a few days or even a month to wait for the grades. This is very aggressive for correcting mistakes. In the process of waiting, many of them will be forgotten in the exam questions that can make up for the previous knowledge.

In this paper, we inspired by the CVPR paper Deep Paper Gestalt[6] , we address these problems by the deep learning method. We trained the deep convolutional neural network by using the PEPD (Private Examination Papers Dataset) based on its visual appearance. Then, we built a tool for students to estimate their examination scores quickly, our deep network based on classifier are able to achieve an accuracy of approximately 85% over a 10 scores error and approximately 95% accuracy over a 20 scores error. So our tool can 1) estimate the approximately scores and find where is the highest probability of error or 2) summarize the high score answer mode to help students improve their scores by writing typography or some tricks.

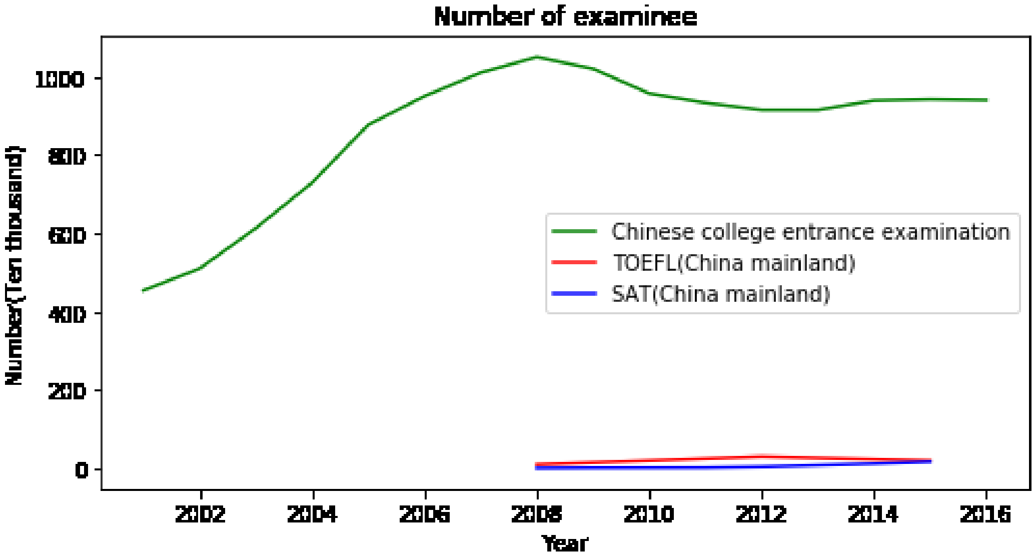

Figure 1: The increase of examinee

2 Related Work

Classification method. It was amazing that CVPR's Paper Gestalt work[6][10] where use AdaBoost for the good/bad paper classifier. Based on the classic computer vision algorithm[5], we use an end-to-end depth model training progress. ResNet-18[4] is selected for our deep convolutional neural network. Because ResNet passes the residual method, it greatly improves the accuracy of modern computer vision and pattern recognition classification tasks.

Vision-based methods. Computer vision has been able to perform well on many reverse sides[1][11] in recent times. The biggest inspiration for our work is the paper that identifies the paper[6] in CVPR. They deal with the paper gestalt problem with deep learning and learn task-specific representation through an end-to-end training process.

Administrative methods. In the TOEFL test of the past few years, the essay written by the candidate will use a preliminary screener to process the essay, check out the obvious grammatical errors in the essay or suspicion of plagiarism, but the TOEFL test official did not publish the technology. The specific steps may be due to the fairness of the score.

Text-based methods. Earlier this year, some scholars proposed a machine learning subjective answer classification system based on MaxEnt text classifier[9]. There are also paper grading methods based on text and traditional machine learning methods[7] proposed by scholars in earlier years. There is also an innovative structure of scoring paper by Two-Stage Learning[8]. But they all ignore important feature information based on text images.

3 Learning to Rate Examination Papers

We trained the ResNet-18 deep residual convolutional neural network on the Imagenet pre-training model based on image data. In the following, we will describe 1) the problems, 2) our processing of the dataset, 3) the details of the deep network training, 4) the way the predictions were scored, 5) and the evaluation of the performance

3.1 Problem Description

The problem we face can be essentially converted into a multi-class task. We label our dataset as a single image with a score range label, $\{(x_1, y_1), (x_1,y_1),\cdots, (x_N,y_N)\}$. It is necessary to train the classification model as much as possible to score future papers.

3.2 Dataset construction

Data source. We signed a confidentiality agreement with a high school and got all the answer sheets and score data from the school in recent years, and obtained a high-resolution answer image using a Canon scanner. The data set includes images of more than 200,000 exams in various subjects in six grades from junior high school to high school.The data set imbalance is also a big problem. The scores of almost every subject are similar to the Gaussian distribution. The number of middle scores is large, and the number of people at both ends is small. We randomly selected the scores of the exams from the dataset, and based on the student number as the x-axis, the score is the y-axis as shown in Figure 2. Since some of the answer sheets are double-sided in the data set, each piece of data corresponds to the same label. In the dataset picture, the top left corner of each test paper answer sheet has the student's test number name and other information and the exam notice reminder, these metadata are not good for the classification of the network. These circumstances have caused the data set to be unbalanced and the data is not clean.

Data acquisition and preprocessing. Here we cite our construction plan for the dataset.

Labeling: The images we obtained were all original images, only the chaotic order pictures and the non-corresponding score labels. We identified the test number of each test paper through the barcode recognition tool and the machine-readable card recognition tool, and took each test paper. Corresponds to its score.

Splice: Because each answer sheet has both positive and negative sides, both sides have important information, so we combine the answer sheets by the upper and lower stitching methods, and randomly extract some pictures in the training dataset as shown in Figure 3.

Pre-processing: In order to avoid the interference of useless futures, we cut off the test number of each image head and some instructions and padding on a white background.

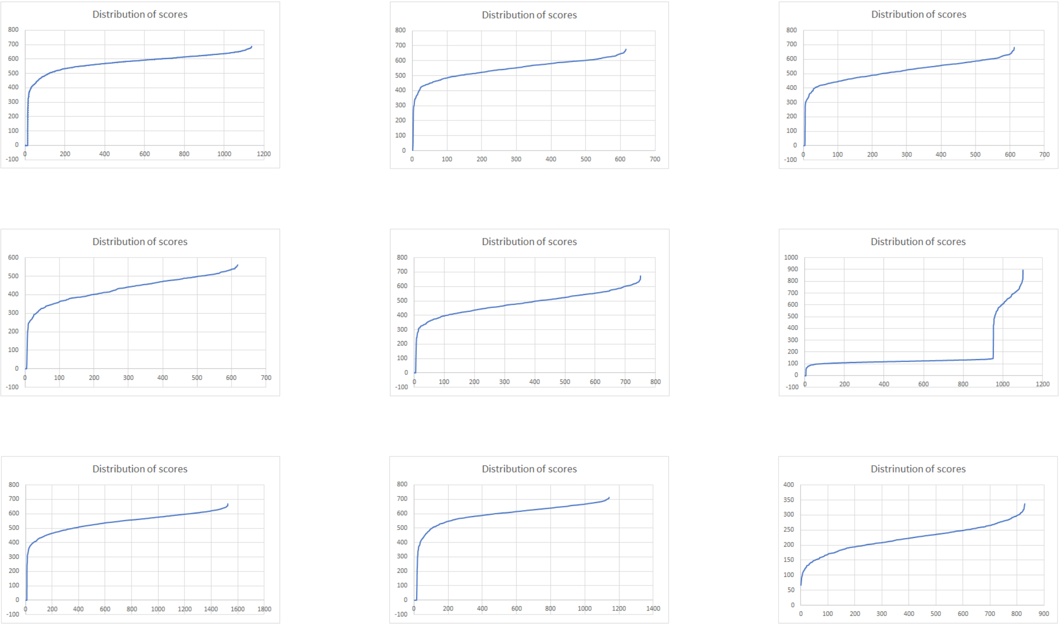

Figure 2: The Distribution of scores.

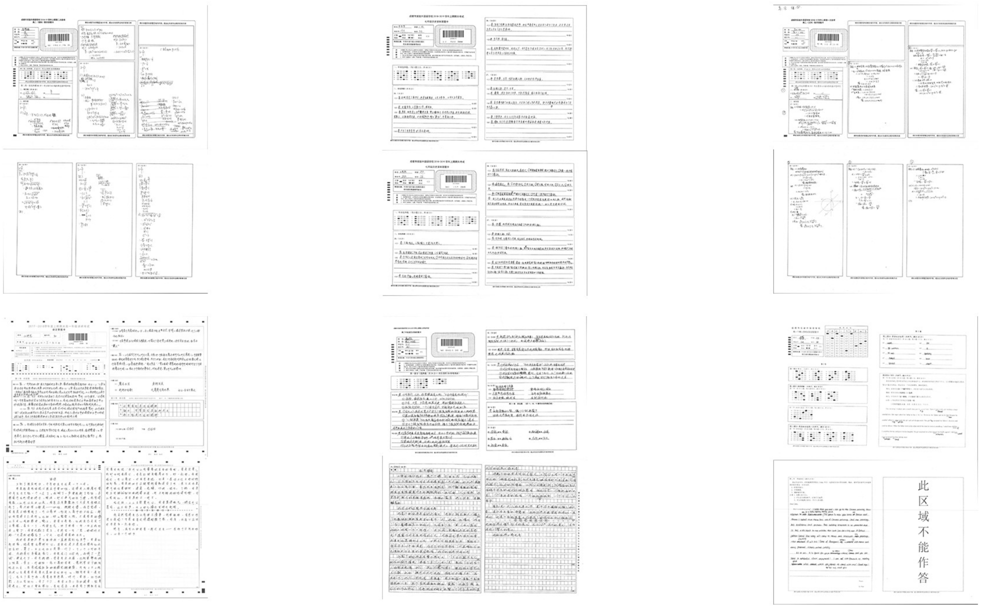

Figure 3: The example we randomly chose from PEPD dataset that after splicing.

3.3 Scoring as image classification

Because of the imbalance of data, we group data by segment by tag information. Because the data at both ends is small, we divide the data at both ends according to a larger interval, and the more intermediate portions are divided by smaller intervals, thereby improving the accuracy of scoring. We have divided the data into 10 categories (according to the score label). Of course, if we collect enough data in the future, we can divide more classes, so that we can accurately predict inter-cells with high accuracy in ground score. We use ResNet-18[4](pre-trained on ImageNet[2]) as our classification network. We replaced ImageNet's 1000 classes with 10 categories and fine-tuned them on the web. In order to speed up the training, we resize the HD image with the original size of 2482*3504 to 512*512 resolution. In the gradient descent section, we chose the SGD optimizer and trained 100 epochs through learning rate decay. Due to the very structured data, we did not do any image enhancement methods similar to adding noise, enhancing contrast, flipping, etc., so we saved as much original visual information as possible. This deep network took less than a day on an NVIDIA Tesla K80 GPU with 100 epochs training.

3.4 Experimental results

Evaluation. In the testset divided by our PEPD dataset, in general, we are able to achieve anaccuracy of approximately 85% over a 10 scores error andapproximately 95% accuracy over a 20 scores error. For each subject's accuracy rate, there are differences in the performance of the deep network. In Table 1, we enumerate the testset accuracy of the answer to the six subjects of Chinese mathematics English physical chemistry, biology, political history and geography.

Subjects	CHI	Maths	ENG
Accuracy(10 points error)	0.895	0.854	0.906
Accuracy(20 points error)	0.949	0.936	0.960
Subjects	Phy	Chem	Bio
Accuracy(10 points error)	0.827	0.815	0.804
Accuracy(20 points error)	0.948	0.941	0.919
Subjects	Politics	Hist	Geog
Accuracy(10 points error)	0.863	0.872	0.855
Accuracy(20 points error)	0.970	0.969	0.954

Table 1: Model Evaluation

AnalysisOur model can achieve an accuracy of about 0.85 or 0.95 on the PEPD dataset, but there are differences in the accuracy of the test data in various disciplines, as shown in Table 1. Among them, the accuracy of Chinese and English political history geography papers is relatively high, the mathematics test paper data is at a medium level, and the accuracy of physical and chemical biological test paper data is low. This is because the Chinese English political history and geography are biased towards the liberal arts. There are many subjective questions on the test paper, which requires the candidates to write a large number of words, and the deep convolutional neural network classified only by image features that it can learn more robust. Physical and chemical biology is biased towards science. Most of the test paper data are numbers and formulas and calculations, and there are many fill-in-the-blank questions. The image features of good answers and bad answers are not very different, so the learning of deep convolutional neural networks is not powerful as the other. The mathematics test paper data is somewhere in between. There are both fill-in-the-blank questions and big questions to be justified. Deep convolutional neural networks can learn some features and have certain robustness.

4 Limitations and Future to Do

In this paper, we use a deep residual convolutional neural network to classify test papers, but only by image features to score the test papers, it can also be said to be a metaphysical method, the scores are not highly reliable. Sex, because we don't really identify the text and formula content on the test paper and use the Text-based method to score accurately. With regard to the interpretability of this classification score, we can use the analysis of the feature map or the generation of the heat map to see what the deep convolutional neural network has learned, and what kind of features appear in a picture will the network The pictures are sorted into the high score category. While scoring through classification, we should also think about how to generate images that are favored by the deep convolutional neural network or the teachers who score the test papers, and analyze these images to summarize some non-additional subjects for students that the knowledge increasing the score of the tricks. The generator can be generated using the Generative Adversarial Network[3] method, but may result in greater computing resource consumption.

References

[1] J. A. Carballo, J. Bonilla, M. Berenguel, J. Fernández-Reche, and G. Garcıa. "New approach for solar tracking systems based on computer vision, low cost hardware and deep learning". In: *Renewable energy* 133 (2019), pages 1158–1166.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "Imagenet: A large-scale hierarchical image database". In: (2009).

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pages 2672–2680.

[4] K. He, X. Zhang, S. Ren, and J. Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pages 770–778.

[5] G. E. Hinton and R. R. Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *science* 313.5786 (2006), pages 504–507.

[6] J.-B. Huang. "Deep Paper Gestalt". In: *arXiv preprint arXiv:1812.08775* (2018).

[7] L. S. Larkey. "Automatic essay grading using text categorization techniques". In: *SIGIR*. Volume 98. 1998, pages 90–95.

[8] J. Liu, Y. Xu, and L. Zhao. "Automated Essay Scoring based on Two-Stage Learning". In: *arXiv preprint arXiv:1901.07744* (2019).

[9] A. Sakhapara, D. Pawade, B. Chaudhari, R. Gada, A. Mishra, and S. Bhanushali. "Subjective Answer Grader System Based on Machine Learning". In: *Soft Computing and Signal Processing*. Springer, 2019, pages 347–355.

[10] C. Von Bearnensquash. "Paper gestalt". In: *Secret Proceedings of Computer Vision and Pattern Recognition (CVPR)* (2010).

[11] M. Winter, J. Bourbeau, S. Bravo, F. Campos, M. Meehan, J. Peacock, T. Ruggles, C. Schneider, A. L. Simons, and J. Vandenbroucke. "Particle identification in camera image sensors using computer vision". In: *Astroparticle Physics* 104 (2019), pages 42–53.