COMMUNICATIONS IN
SCIENCE AND TECHNOLOGY
Homepage: cst.kipmi.or.id

Comparison of text-image fusion models for high school diploma certificate classification

Chandra Ramadhan Atmaja Perdana*, Hanung Adi Nugroho, Igi Ardiyanto

Department of Electrical and Information Engineering, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

Article history:
Received: 29 April 2020 / Received in revised form: 17 May 2020 / Accepted: 20 May 2020

Abstract

File scanned documents are commonly used in this digital era. Text and image extraction of scanned documents play an important role in acquiring information. A document may contain both texts and images. A combination of text-image classification has been previously investigated. The dataset used for those research works the text were digitally provided. In this research, we used a dataset of high school diploma certificate in which the text must be acquired using optical character recognition (OCR) method. There were two categories for this high school diploma certificate, each of which has three classes. We used convolutional neural network for both text and image classifications. We then combined those two models by using adaptive fusion model and weight fusion model to find the best fusion model. We came into conclusion that the performance of weight fusion model which is 0.927 is better than that of adaptive fusion model with 0.892.

Keywords: Text classification, image classification, text-image classification, convolutional neural network

1. Introduction

Information extraction from a high school diploma certificate are necessary for the uninversity admission commitee during new students admission. Submitting a high school diploma certificate are one of many requirements that must be fullfilled by the school leaver. In this digital era the high school diploma certificate are commonly submitted electronically using file scanned document that will later be classified by the admission commitee member of the university.

Many manual labors are needed to verify and classify the submitted documents and it raises a question whether it is posible to automate the process. The high school diploma certificate contain texts and images. Information extraction from text and images in documents for classification is explained as follows.

File scanned document's texts can be extracted using OCR. There are no guarantee that OCR accuracy will be 100%. The error in OCR usage raises a question whether the OCR results will affect the accuracy of text classification in folowing process.Taghva's research using a small collection of documents with long paragraph proved that OCR errors do not affect the accuracy of text classification [1].

However, in the following years, Taghva proved that using automatic correction on the documents with OCR error would improve text classification accuracy [2]. Another reasearcher [3] concluded that OCR error will affect greatly on text classification accuracy if the disturbed words are significant for specific classes. In [3], three methods of documents representation were introduced to improve the accuracy of text classifcation in which texts were acquired through OCR. The three method introduced in [3] were the elimination of stop words, lemmatization, and n-grams of character.

Another aproach to improve OCR accuracy is background elimination [4]. This method worked by comparing three OCR software and aplying the background elimination. The research proved that background elimination can improve OCR accuracy.

Image resolution also affects OCR accuracy. A research to improve OCR accuracy of low resolution image has been done and showed good results [5]. The research used three steps method namely resizing, sharpening and blurring to improve OCR accuracy.

There are many research works on text classification model. In [3], there were four methods mentioned for text classification, centroid, support vector machine (SVM), k-nearest neighbor (k-NN), and Naive Bayes (NB). Some researchers agreed that decision three also was a feasible method for text classification [6]. An application of term weighting matrix in SVM proved an improvement of SVM performance [7]. The most recent research shows a trend of convolutional neural network (CNN) usage in text classification and proved better performance [8,9].

CNN achieved a good performance not only for text classification, but also for image classification. Although CNN consumes a great computation resource and requires long time to train, some method are still available to solve those problems [10]. CNN is not quite good for image classification if the

* Corresponding author.
Email: chandra.atmaja@mail.ugm.ac.id

image contains many objects with the variation of shapes and sizes, and cluttered [11]. However, all images in high school diploma certificate are at the same shape and size and are not being cluttered; thus, CNN is still a feasible method.

Recently, the combination of text-image classification has been a new developed model. There are two papers in text-image classification one by Guo Li and Na Li [12] and another one was by Fangyi Zhu *et al* [13]. Guo Li and Na Li used an adaptive fusion model, while Fangyi Zhu *et al* used a weight fusion model with decision strategy.

Although Guo Li and Na Li claimed that the proposed adaptive fusion model were compared with the weight fusion model, it was not clearly explained whether the weight fusion model applied the decision strategy. Both text-image classification models use dataset in which text data have already been digitally provided. In our dataset of high school diploma certificate the text must be acquired using OCR.

Our contribution are the addition of OCR pre-processing in the text classification sub model of the text-image fusion model, and clear comparison between adaptive fusion model and weight fusion model on our dataset.

2. Materials and Methods

2.1. The Dataset

The dataset consisted of 1555 files, splited into three, 870 for training, 218 for validation, and 467 for test. Fig. 1 shows the image examples from each class.

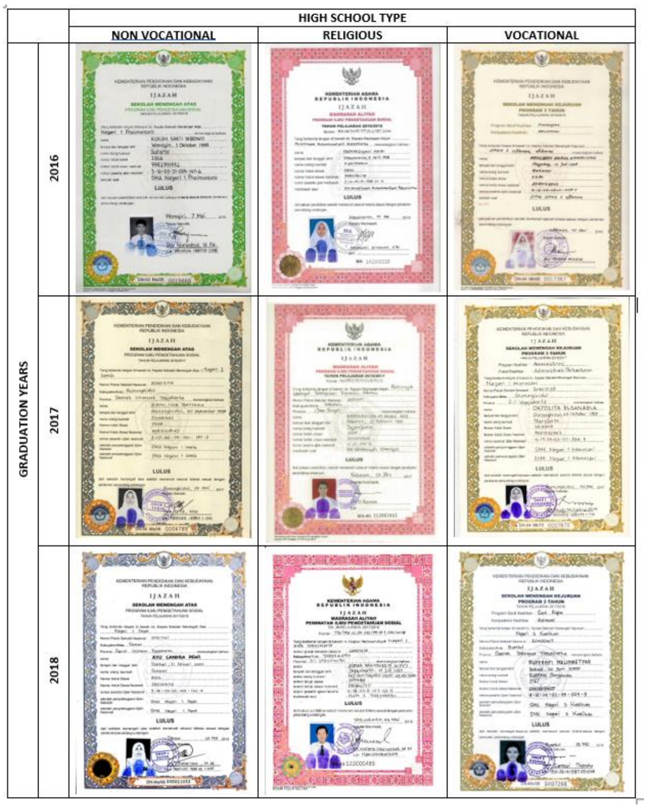

Fig. 1. Image examples from each class in the dataset

The acquisition of this dataset has been approved by the admission committee of the university considering for research purpose only and for the development of automated high school diploma certificate classifier for the admission system. The dataset will be kept from public access with a purpose to keep the privacy of its owner.

2.2. Text Classification Model

We trained text classification model separately from the image classification model. Text pre-processing after OCR process included converting to lower case, removing stop words, converting numeric to letter, removing word with one letter only, and removing multiple spaces. Converting numeric to letter was deemed necessary with a consideration that graduation year information is written in numeric, and we use word embedding vector. This conversion was able to ascertain that all numeric information was properly embedded with vector values.

As seen in Fig. 2 we made branch layer for graduation year categories with 3 classes (2016, 2017, and 2017) and for high school categories with 3 classes (non-vocational, vocational, and religious). The model consisted of 1-dimensional convolutional layer with kernel size 3, 1-dimensional max pooling layer, followed by hidden layer, ReLU activation layer and output layer with node for each class.

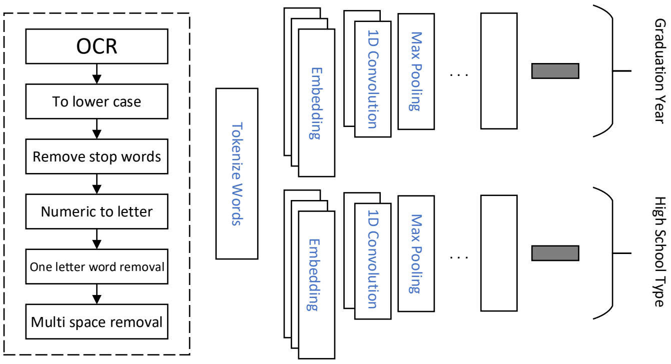

Fig. 2. Illustration of text classification model

2.3. Image Classification Model

Input image for our image classification model was resized to 160 × 128. This resizing process was necessary to reduce the computation process to make model training faster, but adequate to ascertain that there were no missing information from the image.

The image classification model had 3 2-dimensional convolutional block and 1 fully connected block. First convolutional block had 32 filters, 11 × 11 kernel size, and 4 × 4 strides, and 2 × 2 max pooling. The two following convolutional block had 64 filters, 3 × 3 kernel size, 1 stride, and 2 × 2 max pooling. The flatten block had ReLU activation layer and output layer with node for each class. Fig. 3 shows the illustration of this model.

2.4. Fusion Model

Table 1 presents the difference between Guo Li and Na Li [12] and Fangyi Zhu *et al* [13] proposed fusion model.

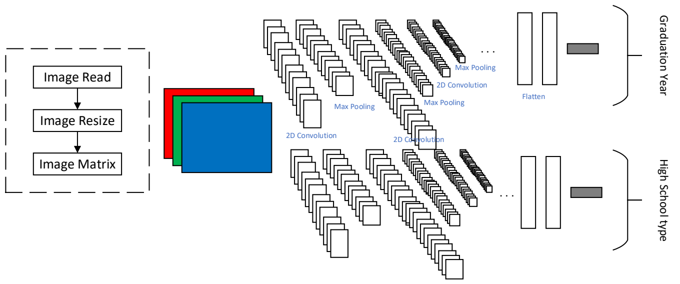

Fig. 3. Illustration of image classification model

Table 1. Difference method between Guo Li and Na Li and Fangyi Zhu *et al*

	Guo Li and Na Li [12]	Fangyi Zhu et al [13]
Text Classification	CNN	CNN
Image Classification	CNN	NN
Fusion Model	Adaptive Fusion	Weight Fusion with Decision Strategy
Dataset	Crawled from a Chinese global trade website in September 2018	LabelMe and UIUC-Sports
Accuracy Results	Proposed Model: 0.9383 Compared Weight Fusion Model: 0.9370	Proposed Model: LabelMe: 0.9775 UIUC-Sports: 0.9951

The adaptive fusion proposed by Guo Li and Na Li can be explained as follows. A data (*m*, *t*) contained images and texts. The developed image classification model had training accuracy $a_{img}(i)$ and probability $p_{img}(m, i)$ for *i* class. The text classification model had training accuracy $a_{text}(i)$ and probability $p_{text}(t, i)$ for *i* class. For data *x* the combined probability *p(x, i)* for *i* class could be calculated using (1), and the data *x* was classified to a class with the largest *p(x, i)* value.

$$p(x, i) = w_{img}(i)p_{img}(x, i) + w_{text}(i)p_{text}(x, i), \qquad (1)$$

where $w_{img}(i) = \frac{a_{img}(i)}{a_{text}(i) + a_{img}(i)}$, $w_{text}(i) = \frac{a_{text}(i)}{a_{text}(i) + a_{img}(i)}$

The weight fusion model used regularization parameter λ to control the balance between text classification and image classification model. The value of λ was set between 0 and 1. After the probability *p(x, i)* of data *x* for each *i* class has been calculated, data *x* were classified to the largest *p(x, i)* value. The formula for this weight fusion model can be examined in (2).

$$p(x, i) = \lambda p_{img}(x, i) + (1 - \lambda)p_{text}(x, i) \qquad (2)$$

Fangyi Zhu *et al* added a decision strategy because in the text classification model, if there are indiscriminative words it cannot get the correct class. Thus, the decision strategy are, if the text classification model cannot get the correct class, the results from image classification model will be used directly and discard the results from text classification model. With data (*m*, *t*) the decision function can be seen in (3). Image features are *θ*(*m*) dan text features are *φ*(*t*).

$$f(m,t) = f(\theta(m)^T\varphi(t)) = \begin{cases} 0, & \text{if } \theta(m)^T\varphi(t)=0 \\ 1, & \text{if } \theta(m)^T\varphi(t)\neq 0 \end{cases} \qquad (3)$$

The equation for the final decision strategy are written in equation (4).

$$p(x, i) = \lambda p_{img}(x, i) + f(m, t)\ (1 - \lambda)p_{text}(x, i) \qquad (4)$$

Fig. 4 depicts the structure of fusion model for our research. We compared fusion model from [12] and [13]. We were not able to apply the decision strategy because our dataset were different from [13]. The difference with [13] is that the indiscriminative words were always available in our text classification model.

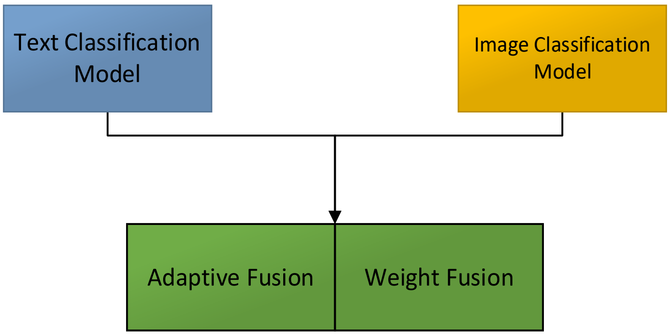

Fig. 4. Structure of fusion model

3. Results and Discussion

3.1. Text Classification

For the adaptive fusion model, we need to get the accuracy results on training dataset for each class. This accuracy on training dataset was required to calculate the adaptive weight. Table 2 shows the accuracy of text classification model on training dataset.

Table 2. Text model accuracy for each class on training dataset

Class	Accuracy
2016	1
2017	0.999
2018	0.999
Non-vocational	0.988
Vocational	0.999
Religious	0.999

The text classification model was not overfitting nor under fitting. This can be confirmed by the learning curve in Fig. 5. This text classification model was trained for 60 epochs. Table 3 shows the performance of the text classification model on test dataset with the model accuracy of 0.925.

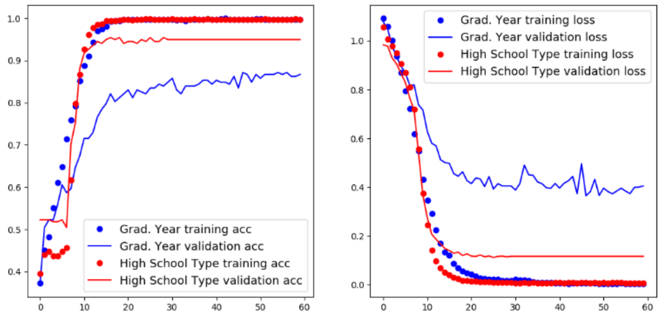

Fig. 5. Learning curve of text classification model

Table 3. Text classification performance on test dataset

Class	Precision	Recall	F1-score
2016	0.927	0.912	0.919
2017	0.839	0.918	0.876
2018	0.874	0.776	0.822
Non-vocational	0.971	0.971	0.971
Vocational	0.967	0.978	0.972
Religious	0.975	0.952	0.963

3.2. Image Classification

Table 4 and Fig.6 respectively show the image classification model accuracy on training dataset for each class and the learning curve on for 60 epochs. The image classification model performance overall was not better than the text classification model as seen by in Table 3 and Table 5. This model accuracy was 0.886.

It can be seen that the image classification model performance for graduation year categories was far below the text classification model. Precision for class 2018 was 0.641 far below the text classification model with 0.874.

Although the image classification model performance overall was below the text classification model, the precision for religious high school type was 1 with recall 0.952. It was better than the text classification model.

Table 4. Image model accuracy for each class on training dataset

Class	Accuracy
2016	0.997
2017	0.964
2018	0.968
Non-vocational	1
Vocational	1
Religious	1

3.3. Fusion Models

The implementation of adaptive fusion model required us to calculate the adaptive weight as shown in (1). Table 2 and Table 4 were used to calculate the $w_{img}(i)$ and $w_{text}(i)$. The results of adaptive fusion model on test dataset are shown in Table 6. Adaptive fusion model accuracy is 0.892.

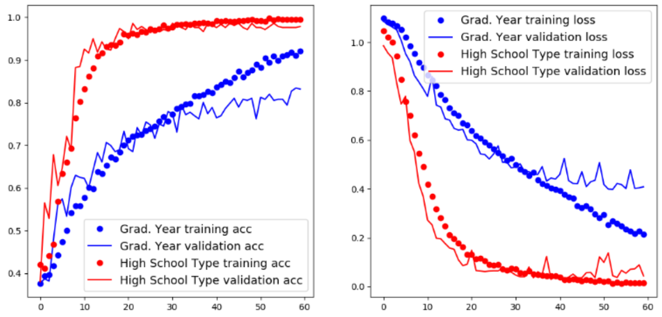

Fig. 6. Learning curve of image classification model

Table 5. Image classification performance on test dataset

Class	Precision	Recall	F1-score
2016	0.907	0.912	0.909
2017	0.836	0.688	0.755
2018	0.641	0.802	0.713
Non-vocational	0.961	0.971	0.966
Vocational	0.961	0.972	0.967
Religious	1.000	0.952	0.975

Table 6. Adaptive model performance on test dataset

Class	Precision	Recall	F1-score
2016	0.927	0.912	0.919
2017	0.839	0.918	0.876
2018	0.874	0.776	0.822
Non-vocational	0.971	0.971	0.971
Vocational	0.967	0.978	0.972
Religious	0.975	0.952	0.963

Table 7 shows that the best accuracy of weight fusion model was 0.927 with λ value 0.02. The performance of weight model with λ value 0.02 is shown in Table 8. Comparing Table 6 and Table 8 it shows that the weight fusion model outperformed the adaptive fusion model. Precision for each class in the weight fusion model was better than the adaptive fusion model.

Table 7. Weight model accuracy on test dataset

Model	Accuracy
$\lambda = 0$	0.925
$\lambda = 0.02$	0.927
$\lambda = 0.04$	0.921
$\lambda = 0.06$	0.920
$\lambda = 0.08$	0.919
$\lambda = 0.10$	0.916
$\lambda = 0.12$	0.916

The best model for high school diploma certificate classification was found in the weight fusion model. Paper [12] claimed that adaptive fusion model were better than weight fusion model, but our research proved the opposite. The accuracy comparison of each model is shown in Table 9.

Table 8. Weight model performance (λ = 0.02) on test dataset

Class	Precision	Recall	F1-score
2016	0.918	0.934	0.926
2017	0.878	0.806	0.840
2018	0.764	0.836	0.798
Non-vocational	0.995	0.985	0.990
Vocational	0.983	0.994	0.989
Religious	1.000	1.000	1.000

Table 9. Weight model performance (λ = 0.02) on test dataset

	Text Model	Image Model	Adaptive Fusion	Weight Fusion
Accuracy	0.925	0.886	0.892	0.927

4. Conclusion

In this paper we have implemented OCR pre-processing to the dataset to acquire digital texts. We used text classification model with digital text from OCR as input. We also used image classification model, which was trained separately. We compared two fusion models from previous research works [12,13]. We trained text classification model and image classification model with fewer epoch than [13]. This research found that the accuracy of weight fusion model with 0.927 outperformed that of adaptive fusion model with 0.892. The limitation of our research is that decision strategy from [13] could not be implemented for our dataset, because the indiscriminative words in the dataset were always available. For the future research, it is suggested that the development of general-purpose decision strategy are not dependent on text features dataset.

Acknowledgements

We would like to thank the admission committee for providing the dataset for our research. Chandra Ramadhan Atmaja Perdana is thankful to Ministry of Education and Culture of the Republic of Indonesia for providing him Magister Scholarship under General Directorate of Learning and Student Affairs.

References

1. K. Taghva, T. A. Nartker, J. Borsack, S. Lumos, A. Condit and R. Young, *Evaluating text categorization in the presence of OCR errors*, *8th SPIE Conference on Document Recognition and Retrieval, San Jose, CA, USA, 2000* pp. 68–74.
2. K. Taghva, N. Thomas and B. Julie, *Recognize, Categorize, and Retrieve*, *Sympo-sium on Document Image Understanding Technology, Columbia, MD, USA, 2001* pp. 227--232.
3. D. Zelenika, J. Povh and A. Dobrovoljc, *Document categorization based on OCR technology : An overview*, Recent Adv. Inf. Sci. (2013) 409-414.
4. M. Shen and H. Lei, *Improving OCR performance with background image elimination*, *12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Zhangjiajie, China, 2015* pp. 1566–1570.
5. M. K. Ugale and M. S. Joshi, *Improving optical character recognition forlow resolution images*, Int. J. Comput. Sci. Netw. 6 (2017) 18-20.
6. V. Renganathan, *Text mining in biomedical domain with emphasis on document clustering,* Healthc. Inform. Res. 23 (2017) 141–146.
7. M. Haddoud, A. Mokhtari, T. Lecroq and S. Abdeddaïm, *Combining supervised term-weighting metrics for SVM text classification with extended term representation*, Knowl. Inf. Syst. 49 (2016) 909–931.
8. P. Li, F. Zhao, Y. Li and Z. Zhu, *Law text classification using semi-supervised convolutional neural networks*, *30th Chinese Control and Decision Conference, Shenyang, China, 2018* pp. 309–313.
9. S. Song, H. Huang and T. Ruan, *Abstractive text summarization using LSTM-CNN based deep learning*, Multimed. Tools Appl. 78 (2019) 857–875.
10. K. Park and D. H. Kim, *Accelerating image classification using feature map similarity in convolutional neural networks*, Appl. Sci. 9 (2018) 108.
11. P. Tang, X. Wang, B. Shi, X. Bai, W. Liu and Z. Tu, *Deep FisherNet for Image Classification*, IEEE Trans. Neural Networks Learn. Syst. 30 (2019) 2244–2250.
12. G. Li and N. Li, *Customs classification for cross-border e-commerce based on text-image adaptive convolutional neural network*, Electron. Commer. Res. 19(4) (2019) 799–800.
13. F. Zhu et al., *Image-text dual neural network with decision strategy for small-sample image classification*, Neurocomputing 328 (2019) 182–188.