ADCAIJ: Advances in Distributed Computing and Articial Intelligence Journal
Regular Issue, Vol. 6 N. 1 (2017), 5-12
eISSN: 2255-2863
DOI: http://dx.doi.org/10.14201/ADCAIJ201761512

Classification of Two Comic Books based on Convolutional Neural Networks

Miki Ueno[a], Toshinori Suenaga[b], and Hitoshi Isahara[a]

[a]Information and Media Center, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580, Japan
[b]Graduate School of Computer Science and Engineering, Toyohashi University of Technology, 1-1 Hibarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580, Japan
ueno@imc.tut.ac.jp, suenaga@lang.cs.tut.ac.jp, isahara@tut.jp

KEYWORD

Computational model of comics; Comic engineering

ABSTRACT

Unphotographic images are the powerful representations described various situations. Thus, understanding intellectual products such as comics and picture books is one of the important topics in the field of artificial intelligence. Hence, stepwise analysis of a comic story, i.e., features of a part of the image, information Convolutional neural features, features relating to continuous scene etc., was pursued. Especially, the length network and each scene of four-scene comics are limited so as to ensure a clear interpretation of the contents. In this study, as the first step in this direction, the problem to classify two four-scene comics by the same artists were focused as the example. Several classifiers were constructed by utilizing a Convolutional Neural Network (CNN), and the results of classification by a human annotator and by a computational method were compared. From these experiments, we have clearly shown that CNN is efficient way to classify unphotographic gray scaled images and found that characteristic features of images to classify incorrectly.

1. Introduction

Actually, unphotographic images are useful way to describe various situations regardless of ages and countries. Previously, we had focused on modeling the pictures and analysis features of comics . Despite rapid advances of image recognition by deeplearning (Quoc et al., 2013), numerous challenging tasks must be addressed in order to understand comics and animation. Generally, there are difficult problems to recognize these unphotographic images. Especially, comic books are composed of complex objects such as characters, natural languages, and several marks. In addition, they are usually drawn by using only gray-scaled images. Namely, the aforesaid tasks are broadly classified into the following two types.

1. **Recognize deformed pictures** Classification of deformed pictures is difficult (Eitz et al., 2012), while a convolutional neural network (CNN) (Fukushima et al., 1982) (Krizhevsky et al., 2012) remarks the potential to achieve the same.

2. **Recognize each scene and the whole story** Recognize the complex meaning of the story in the following three steps and construct a suitable layered classifier.

   A part of each scene The name of the object and facial expressions.

Scene Social relationships between the characters and the emotions related to each character.

Inbetweening scenes Interpretation of the story by generating intermediate frames between two images for sequential transition of the first frame to the second.

The previously used image recognition method involves machine learning based on appropriate manual vectors utilizing SIFT, HOG and Haar-like features. Then, a part of each scene is recognized, for example face detection, object recognition, etc. On the other hand, CNN, which is one of the methods of deep learning, especially for images, is applied to the image and feature vectors are automatically constructed, so that scene recognition is possible. There are a few studies on the feature vectors for comics (Tanaka et al., 2010) because of copyright problems and less information than coloured photographic images. Thus, it is difficult to design the problems to be solved and datasets to be prepared; adequate discussion of the feature vectors is hence needed.

In this study, as the first step toward understanding comics by using computers, several classifiers are constructed for basic problems; to classify two comic books by the same artists. Detailed discussion are described between by hand and by a computational method.

The rest of the paper is structured as follows. Section 2 describes the features of comics. Section 3 shows the preliminary experiment carried out to classify images by hand, while Section 4 shows the experiment for the computational method. Section 5 describes the detailed comparison of the two experiments. Additional experiments undertook in Section 6 and discuss features of images in Section 7. Finally, section 8 concludes this research and gives brief insights into a future study.

2. Four-Scene Comics

Numerous genres and structures of comics exist across the globe. In this study, four-scene comics, which are structured with four continuous scenes, are considered. The length of four-scene comics is limited so as to ensure clear interpretation of the contents. Figure 1 shows the general structure of each page of the four-scene comics.

Story four-scene comics is one of the styles used in popular Japanese comics. The notable feature of this type of comics is that the characters are common among various short stories, and continuous small stories result in a whole story in the book. Therefore, it is easy to classify two stories by considering the whole series of sequential images. Although the fourth scene of a small story plays an important role to interpret stories, it may be difficult to classify two stories focused on the fourth scene that is selected randomly, because some of characters may be identical between stories; features of characters are similar, and new characters may have appeared in the middle of a story, in the case of works by the same artist.

3. Preliminary Experiment

In this preliminary experiment, two *story four-scene comics* by same artist are used in order to consider interpretation of comics by human. In order to prevent interpret stories by the title of a small story and inbetweening meanings, only the fourth scene of each small story is given to the annotator.

Dataset:

Two classes, «*Konpeito 1*» (Fujino et al., 2007) and «*Ringo no ki no shitakko de*» (Fujino et al., 2007) by the same artist are used, which have the number of 182 and 178 images respectively. Six small stories of from the beginning of «*Konpeito 1* were not included in this dataset because the numbers

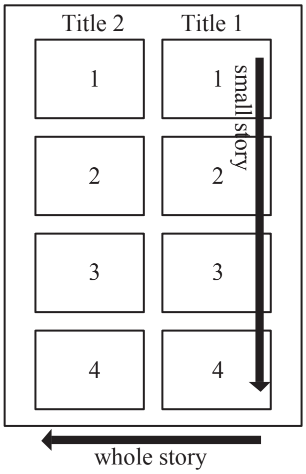
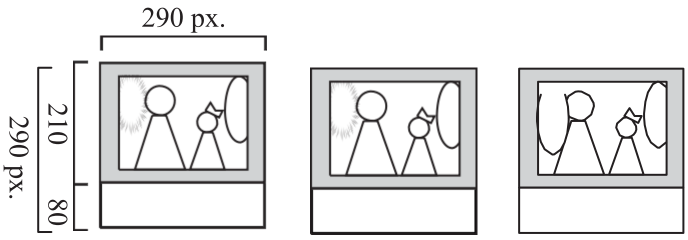


Figure 1: The structure of four scene comics



(a) Original pictures

(b) Without words

(c) Same shaped balloon

Figure 2: Each example of three types dataset

of images in the dataset should be limited to 360 which can be divided by ten. The fourth scene of each small story with 290 × 210 pixels are resized to 290 × 290 pixels added 80 blank pixels to the bottom of the image. Subsequently, they are reduced to 32 × 32 pixels. All the images are in grey scale format.

Comics contain complex information based on both pictures and natural languages. We assumed that picture information was more important for understanding comics. To confirm this, we created three types of datasets by reducing information about natural languages as follows.

- Original pictures

Table 1: The mean accuracy rate by the human annotator

Size of image	Mean accuracy rate
32 × 32 px.	0.68
64 × 64 px.	0.83
128 × 128 px.	0.93

Table 2: Parameters of neural network of exp. 1

The number of output units	2
Batch size	20
The number of epochs	100
Dropout ratio	0.5
Activation function	ReLU function
Loss function	Softmax cross entropy
Optimizer	Adam. default parameter

- Without words inside balloons

- Same shaped balloons without words; shapes of all balloons are replaced into same ones manually.

Figure 2 shows each example of these three types of dataset. A preliminary experiment with «same shaped balloons» dataset type is carried out as follows.

1. Give 90% of the dataset with the name of the story to the annotator. The size of the image is 32 × 32 pixels.
2. Annotator labels 10% of the dataset as test data without the name of the story. The size of the image is 32 × 32 pixels.

The accuracy rate is calculated as: the number of the correct label of the book title is divided by the number of test data. As a result, the mean accuracy rate obtained after repeating the experiment three times. It is difficult for human to recognize what objects appear in image with 32 × 32 px. Thus, the accuracy rate of three types of the size in the same dataset are compared. Table 1 shows the mean accuracy rate for three sizes of images in the same datasets. As larger the size, the higher the accuracy rate. We found that characters appearing is the most important information for the human annotator.

4. Experiment 1

The experiment involves classifying two works by the same artist among three types of datasets by Chainer (Tokui et al., 2015), which is a flexible framework of neural network, written in Python. Table 2 shows the parameter of the neural network. Table 3, Table 4 shows the CNN layer parameters, the pooling layer parameters respectively.

The network architecture and the data are described below.

Table 3: CNN layer parameters of exp. 1

	CNN1	CNN2
Filter size	5 × 5	5 × 5
Padding size	0	0
Stride	1	1

Table 4: Pooling layer parameters of exp. 1

	POOLING1	POOLING2
Filter size	2 × 2	3 × 3
Padding size	0	0
Stride	2	2

Table 5: The mean accuracy rate by the CNN of exp. 1

Types of datasets	Mean accuracy rate
Original pictures	0.72
Without words	0.83
Same shaped balloon	0.84

Architecture:

Input - CNN1 - ReLU - MAX POOLING1 - CNN2 - ReLU - MAX POOLING2 - LINEAR1 - ReLU-Dropout (Hinton et al., 2012) - LINEAR2 - Output

Data:

The same dataset is used in Section 3. Each of the three types of datasets is randomly divided into two groups as follows.

Training data 90 % of the number of the dataset

Test data 10 % of the number of the dataset

Table 5 shows the result of the mean accuracy rate of 10 times of 100 epochs of learning.

5. Discussion 1

Comparison of the results presented in Section 3 and 4 indicates that the accuracy in the computational method is higher than that with the human annotator under the condition of the 32 × 32 pixels. Thus, it can be said that the CNN accurately obtains features of scenes of comics.

The human could learn the features of the two types of books focused on a certain character. In the books prepared for the experiment, numerous characters appeared in one scene, while there were a few scenery images and abstract images. Therefore, the human annotator found that the female protagonist of each story appeared frequently in the scenes. However, the size of image is so small that it is difficult to recognize the characters that appear. In addition, the annotator sometimes cannot identify the characters appearing in each book given that the characters are similar due to the same artist. However, the accuracy rate even for 32 × 32 pixels is quite good. It indicates that human classified two books even if they cannot obtain what objects appear. Thus, other information might contributes to classify.

On the other hand, the result might indicate that the computational method learned based on the arrangement of objects such as characters and balloons. The size of the scene in four scene comics is the same, and it is smaller than that in other types of comics. The balloon object is generally located at the end of the frame. Thus, the location of the other objects is limited.

Table 6: Computer spec of exp. 2

OS	Ubuntu 14.04 LTS
GPU	GeForce GTX TITAN X (12 GB) × 2
CPU	Intel (R) Core (TM) i7-5930K CPU @ 3.50 GHz
Memory	32 GB
Chainer	1.10.0
Java	1.8.0_91
Python	2.7.6

Table 7: Parameters of neural network of exp. 2

The number of output units	2
Batch size	20
Linear 1 node size	256
Linear 2 node size	256
The number of epochs	600
Dropout ratio	0.5
Activation function	ReLU function
Loss function	Softmax cross entropy
Optimizer	Adam:alpha=1e-6

Considering the effect by reducing natural words, the mean accuracy rate for same shaped balloon dataset is the highest, while that of original pictures dataset is the lowest. From this result, it can be said that pictorial information is sufficient to classify these two works by deep learning, because the size of the image is too small to read the character words. Namely, information about the character word is regarded as noise. In the future research, information about pictures and natural languages will be considered in detail.

6. Experiment 2

From above experiments (, 2016), it has been confirmed that CNN is the efficient way to classify comic books by the same artists. Thus, detailed features to classify were described in this experiment. The network architecture were improved and the numbers of learning epochs were increased.

Table 6 shows the spec of computers. Table 7 shows the parameter of the neural network. Table 8, Table 9 shows the CNN layer parameters, the pooling layer parameters respectively. Figure 3 shows the network architectures and the data are described below.

In this experiment, the «original pictures» dataset; the numbers of images are used as shown in section 4. However, the size is different. The input data for CNN is the fourth scene of two books with 290 × 210 pixels without blank pixels. Training and test data are created by three-fold cross validation. Thus, the set contains 360 images; 240 images for training phase, and 120 images for the test phase.

Table 8: CNN layer parameters of exp. 2

	CNN1	CNN2	CNN3	CNN4	CNN5
Filter size	5 × 5	5 × 5	5 × 5	5 × 5	5 × 5
Padding size	0	2	1	1	1
Stride	2	1	1	1	1

Table 9: Pooling layer parameters of exp. 2

	POOLING1	POOLING2	POOLING3
Size	3 × 3	3 × 3	3 × 3
Padding size	0	0	0
Stride	2	2	2

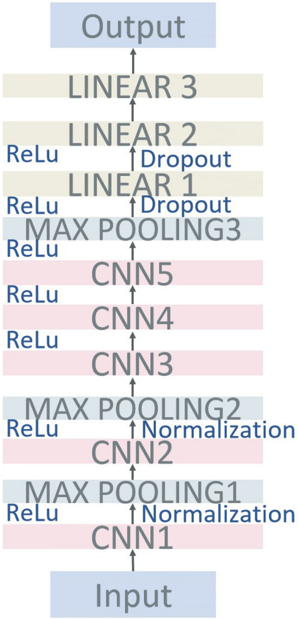


Figure 3: The architecture of Alex net for exp. 2

7. Discussion 2

The result of the mean accuracy rate of 600 epochs of learning is 0.83. The mean accuracy rate was obtained after repeating the experiment two times; each accuracy rate is calculated by three-fold cross validation. 40 common images were incorrectly classified in both times; 62 and 61 failures respectively. The images which were classified incorrectly tend involve following features.

Only supporting characters There is a few images including supporting characters in the whole dataset.

Only based on such scenes, it is difficult to classify.

Similar characters between two books Two books contains similar male characters; they have similar coloured and styled hairs. Scenes of these books usually contains three or more characters. Thus, regional character identification will be useful to classify.

Grayish colours of background images Scenes except characters and some objects are filled with gray as background. The contours of characters are blurred.

8. Conclusion

In this study, as the first step toward understanding comics by using computers, we focused on the example problem to classify images in two comic books by the same artists. From the viewpoint of important objects in comic books, three different types of datasets were prepared. By using CNN approach, several classifiers were constructed. Comparing the result of classification of two books by the human annotator and by computational method, we found that the CNN is efficient to be applied to grey scaled unphotographic complicated images such as scenes of comics. Although CNN approach imitates functions of human sights, we indicated the differences of efficient features to interpret comics between humans and computers, and we discussed the effect of reducing information of scenes of comics; i.e., size and words. In order to interpret story of comics, it is important to recognize structure of comics and sociograms of characters. Therefore, we continue to develop different classifiers and combine them to understand whole stories of comic books.

Future studies would involve the following:

- Consideration of suitable hyperparameters.
- Detailed analysis for features by visualization approach
- Investigation for different features among artists

Acknowledgment

This research is partially supported by DAIKO FOUNDATION. We wish to thank FUJINO HARUKA for providing her comic book as dataset.

9. References

Eitz, M., Hays, J., and Alexa M., 2012. How Do Humans Sketch Objects?, *ACM Trans. Graph. (Proc. SIGGRAPH)*, Vol. 31, No. 4, pp. 44:1-44:10.

Fujino, H., 2007. *Konpeito ! 1 (Confetti ! 1)*, Houbunsha.

Fukushima, K. and Miyake, S. 1982. *Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position, Pattern Recognition*, Vol. 15, Issue 6, pp. 455-469-

Krizhevsky, A., Sutskever, I., and Hinton G. E., 2012. Imagenet classification with deep convolutional neural networks, *In Advances in neural information processing systems*, pp. 1097-1105.

Quoc, V. Le., 2013. Building high-level features using large scale unsupervised learning, *In Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8595-8598.

Tanaka, T., Toyama, F., Miyamichi, J., and Shoji, K., 2010. Detection and Classification of Speech Balloons in Comic Images, *The journal of the Institute of Image Information and Television Engineers*, Vol. 64, No. 12, pp. 1933-1939

Tokui, S., Oono, K., Hido, S., and Clayton, J., 2015. Chainer: a Next-Generation Open Source Framework for Deep Learning, *In Workshop on Machine Learning Systems at Neural Information Processing Systems (NIPS)*.