AUTOMATIC RESTAURANT IMAGE TAGGING

[1]TEERASIRITEERAKUL, [2]KUNLABUTSURIYAKANON, [3]SOFIA SARIDEH

King Mongkut's Institute of Ladkrabang Technology
E-mail: teera.si@kmitl.ac.th

Abstract - Image tagging provides crucial information in many situations. This paper provides investigation of deep learning techniques for this task. We applied these techniques on images taken from restaurant scenes where people tag information only on a few of them. Our aim is to learn from the tagged images and then automatically tag the rest of images. We investigated two deep learning techniques: AlexNet and GoogLeNet. The two techniques are measured first without any pre-trained and then pre-trained with ILSRVC-12 and Places365 dataset. The result of these two techniques provided relatively high tagging accuracy even without pre-train. However, with pre-trained model, the accuracy risen significantly.

Index Terms - Deep Learning, Image Classification, Image Tagging,

I. INTRODUCTION

With booming number of smart phone with image capturing capability, there are a significantly rising number of images. However, the capability of image classification or identification techniques are still improving. With an introduction of deep learning, the task of identify object in an image (object recognition) has risen significantly. Thus, it open up possibilities of many automatic system.

In consumer-reviewed website, such as Yelp or TripAdvisor, we have a lot of images taken by customers in their reviews. These sites typically allow the reviewers to tag each image into category such as atmosphere, outside view, inside view, sign, menu, food, etc. These tagging would help other customer in their search in the future. However, most images are left untagged by reviewers, which makes it hard for develop future system. The task of object recognition has risen significantly in 2012 with an introduction of AlexNet [1] which drop the error rate down by almost 10 percent. Since then, several deep learning architectures, including GoogLeNet [2], has been introduced and lowering the error rate down by approximately 13 percent.

Although the object recognition systems based on deep learning are sufficiently accurate, we still must investigate it for more specific tasks, especially with human tagging where sometimes images are categorized subjectively. Our proposed system aims to follow a tagging system of Wongnai, a customer-reviewed based website for restaurant in Thailand. There 5 tagging categories. In the dataset, some image can be classified in more than one category. However, the proposed system will only classify such image into only one category.

The contribution of this paper is to investigate how well deep learning performs with subjectively categorizedimages. As the result will show, the accuracy result is very high in the pretrained system.

II. RELEVANT THEORIES AND TECHNIQUES

In this paper, we will utilize to deep convolution network, AlexNet [1] and GoogleNet[2]. AlexNet is the firstdeep convolution neuron networks proposed in 2012. The system was trained on 15 million labeled high resolution images with 22 thousand categories. The system consists of 7 hidden layers, 650 thousand neurons, 60 million parameters and 630 million connections. This large-scale network utilizes multiple GPU to speed up the training process. Data augmentation and Softmax techniques are used to reducing the overfitting.

GoogLeNet is also a deep convolution neuron network which enter ILSVRC and won in the year 2014. The architecture of this system is different from the other previous architectures. It used Inception Module with there are 9 modules and more than 100 layers. GoogLeNet has 12 times fewer parameter comparing to AlexNet which save a lot of computational power.

III. RESEARCH METHODOLOGY

Data and Preprocessing Process

There are all together 12,304 images from Wongnai.com. There are 5 classes: 1) inside, 2) food or drink, 3) store front, 4) menu, and 5) profile which are customer's images. Out of 12,304 images, there are only 2,385 tagged images where 9,919 images left untagged. Images are manually check for incorrect tagging or unclear category images.

Model Parameters and Evaluation

Images are resized to 256x256 pixel by squashing. Training epoch is 30. We are experimenting on 4 solvers which are Stochastic Gradient Decent (SGD), Nesterov Accelerated Gradient (NAG), Adaptive Moment Estimation (ADAM), and RMSprop. Performances of all the techniques are evaluated using accuracy and F-1 score.

IV. EXPERIMENTAL RESULT

Model without pre-training

Data cleansing is a process where we manually check for incorrect ambiguous tagged images. Fig.1 shows how data cleansing effect overall accuracy. As illustrated, the data cleansing raised the overall accuracy approximately 3%.



Fig.1 Data cleansing effect.

We evaluated four solvers in each network in our experiments. In Fig.2 top, the result from AlexNet shown that using ADAM as solver yields the highest accuracy of 78.66 where RMSprop yields the lowest accuracy of 74.64 (4% lower). However, the standard deviation of RMSprop is the lowest (1.798) while the standard deviation of ADAM solver is the highest (2.684).

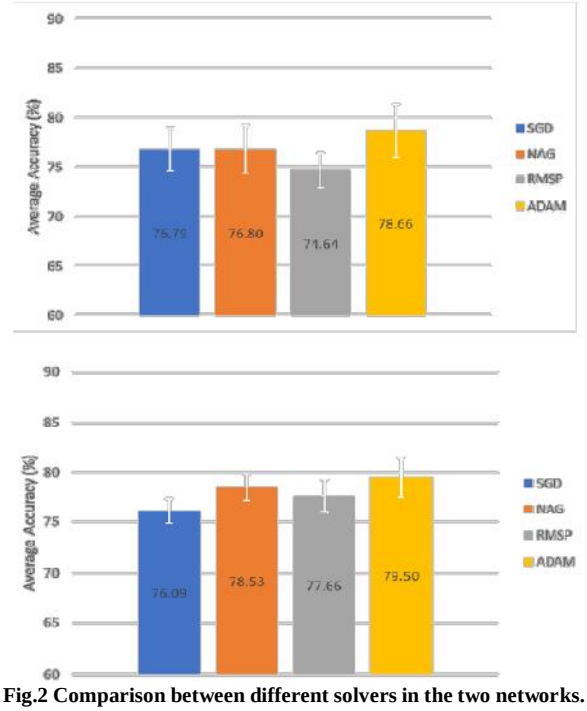


Fig.2 Comparison between different solvers in the two networks.

For the GoogLeNet, see Fig.2 bottom, ADAM solver produced the highest accuracy which is higher than AlexNet network. On the other hand, the lowest accuracy of GoogLeNet came from using SGD solver. Also, the overall standard deviations in GoogLeNet are lower comparing to AlexNet.As shown in Fig.5, F-Score measurements were also follow the same trend. The highest F-Score also yields by ADAM solver in both network where GoogLeNet shown the higher F-Score. The other F-Score and the standard deviations are generally the same as accuracy.

Model with pre-training

In these experiments, we compare the network which are not pre-trained and are pre-trained using dataset from Places365 and ILSRV-2012. With pre-trained model, the overall accuracy was raised for almost 10%. Although the models pre-trained with ILSRV-2012 dataset shown higher accuracy than the models pre-trained with Places365, the standard deviation suggest that the accuracy differences are not significant. See fig.3 for illustrated details.

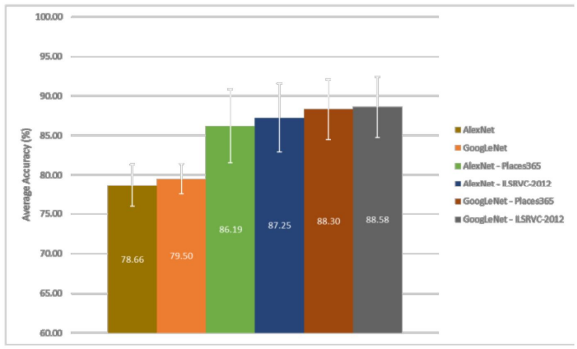


Fig. 3

Fig.6 shown that F-Score of pre-trained networks and none pre-trained network are practically the same comparative relation as in accuracy results. GoogLeNet pre-trained withILSRV-2012 dataset also yield the highest F-Score among the others.

Classification Example

Fig.7 to fig.10 show top 9 classification examples from our experiment. Note that the last category, profile images of customers, are left out due to privacy issue.

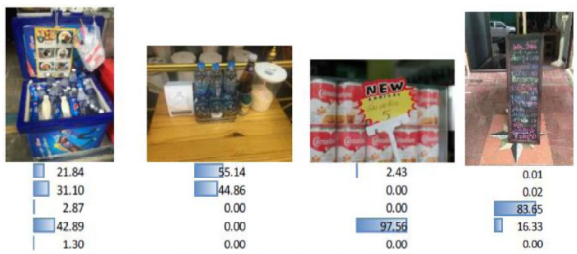


Fig.4 examples of wrongly classified image.

Fig. 4 shows images are our classifier wrongly classified. Under the pictures are confident value of each category as mentioned in section III. The first 3 images from the left in fig.4 should be classified as store inside. However, we can clearly see that it can be easily mistaken as food or drink or menu. On the last image, it is a menu on a blackboard written in chalk, but it is understandable to confuse it with store front image.

	Accuracy (%)								Mean F-Scores (%)							
	AlexNet				GoogLeNet				AlexNet				GoogLeNet			
TCV	SGD	NAG	RMSP	ADAM	SGD	NAG	RMSP	ADAM	SGD	NAG	RMSP	ADAM	SGD	NAG	RMSP	ADAM
1	80.51	82.80	78.31	86.16	75.84	80.78	80.95	83.86	0.8049	0.8279	0.7823	0.8621	0.7564	0.8065	0.8101	0.8394
2	74.69	74.51	73.37	77.25	77.34	79.89	76.72	78.66	0.7473	0.7450	0.7338	0.7720	0.7724	0.7985	0.7677	0.7863
3	76.98	76.81	73.19	77.34	77.25	78.75	77.78	76.81	0.7723	0.7686	0.7336	0.7744	0.7717	0.7884	0.7734	0.7684
4	78.57	76.46	76.54	79.01	76.37	77.07	77.34	79.10	0.7866	0.7642	0.7654	0.7920	0.7643	0.7714	0.7756	0.7929
5	77.78	79.10	74.78	79.54	74.87	77.07	78.31	80.69	0.7775	0.7910	0.7477	0.7964	0.7460	0.7707	0.7846	0.8078
6	72.40	74.34	72.49	76.19	75.49	78.04	76.81	78.66	0.7262	0.7455	0.7262	0.7638	0.7567	0.7813	0.7714	0.7905
7	75.22	74.34	72.49	77.16	73.10	76.46	74.51	76.81	0.7526	0.7445	0.7234	0.7713	0.7304	0.7637	0.7471	0.7687
8	78.57	77.07	75.04	78.92	77.43	78.92	78.66	80.25	0.7873	0.7724	0.7491	0.7903	0.7753	0.7899	0.7885	0.8029
9	76.90	75.49	74.16	77.25	76.90	78.48	76.72	80.60	0.7701	0.7555	0.7400	0.7737	0.7692	0.7854	0.7675	0.8068
10	76.28	77.07	76.01	77.78	76.37	79.81	78.84	79.54	0.7636	0.7710	0.7616	0.7791	0.7666	0.8010	0.7901	0.7965
AVG	76.79	76.80	74.64	78.66	76.10	78.53	77.66	**79.50**	0.7688	0.7686	0.7463	0.7875	0.7609	0.7857	0.7776	**0.7960**
SD	2.182	2.462	1.798	2.684	**1.279**	**1.324**	1.618	1.956	0.0215	0.0243	0.0179	0.0268	**0.0133**	**0.0134**	0.0160	0.0196

Fig.5 Details accuracy and F-Score of models trained with Wongnai dataset.

	Accuracy (%)						Mean F-Scores					
	Pretrained				Non-Pretrained		Pretrained				Non-Pretrained	
	Places356		ILSRVC-2012				Places356		ILSRVC-2012			
TCV	AlexNet	GoogLeNet	AlexNet	GoogLeNet	AlexNet	GoogLeNet	AlexNet	GoogLeNet	AlexNet	GoogLeNet	AlexNet	GoogLeNet
1	99.82	99.56	99.91	99.74	86.16	83.86	0.9982	0.9956	0.9991	0.9974	0.8621	0.8394
2	84.48	86.77	85.63	87.30	77.25	78.66	0.8444	0.8679	0.8561	0.8736	0.7720	0.7863
3	84.48	85.98	85.10	86.68	77.34	76.81	0.8445	0.8595	0.8508	0.8674	0.7744	0.7684
4	85.63	87.65	85.54	86.95	79.01	79.10	0.8561	0.8775	0.8552	0.8694	0.7920	0.7929
5	85.80	87.74	87.39	88.71	79.54	80.69	0.8585	0.8775	0.8738	0.8877	0.7964	0.8078
6	82.54	86.24	84.30	86.24	76.19	78.66	0.8258	0.8638	0.8445	0.8637	0.7638	0.7905
7	83.69	86.33	84.92	85.27	77.16	76.81	0.8354	0.8628	0.8480	0.8537	0.7713	0.7687
8	85.01	87.74	86.95	88.27	78.92	80.25	0.8498	0.8777	0.8697	0.8830	0.7903	0.8029
9	84.22	87.48	86.07	88.36	77.25	80.60	0.8414	0.8746	0.8609	0.8835	0.7737	0.8068
10	86.24	87.48	86.68	88.27	77.78	79.54	0.8620	0.8754	0.8671	0.8830	0.7791	0.7965
AVG	86.19	88.30	87.25	**88.58**	78.66	79.50	0.8616	0.8832	0.8725	**0.8862**	0.7875	0.7960
SD	4.658	3.808	4.317	3.862	2.684	**1.956**	0.047	0.038	0.043	0.038	0.027	**0.020**

Fig.6 Detail accuracy and F-Score of models comparing between pre-trained and not pre-trained dataset.

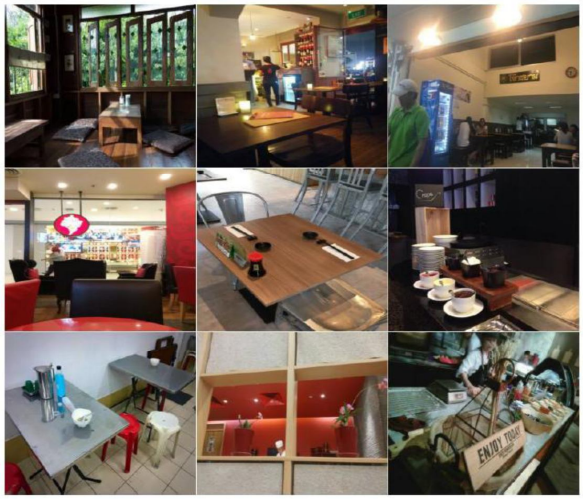


Fig.7 images classified as inside



Fig.8 images classified as food or drink



Fig.7 images classified as store front



Fig.7 images classified as menu

DISCUSSION AND CONCLUSION

This paper provides empirical evident for evaluation of Deep Convolution Neuron Network in the task of image tagging where tagging can be very subjective. We show that data cleansing has some impact on accuracy (3%). On the other hand, having pre-trained model using large dataset such as ILSRVC-2012 can have significant impact on accuracy (almost 10%). In our experiment, we concluded that GoogLeNet yields slightly better performance comparing to AlexNet (<1%).

REFERENCES

[1] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.

[2] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," IEEE Trans. Electron Devices, vol. ED-11, pp. 34-39, Jan. 1959.

[3] C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," IEEE Trans. Image Process., vol. 10, no. 5, pp. 767-782, May 2001.

[4] K. Alex, S. Ilya, H. Geoffrey, "Image Net Classification with Deep Convolutional Neural Network", NIPS, 2012.

[5] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... &Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9), 2015.

★★★