JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 3, June 2023, hlm. 521 - 528 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i3.2277
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

PERFORMANCE ANALYSIS RESNET50 AND INCEPTIONV3 MODELS FOR CAPTION IMAGE GENERATOR

Hasif Priyambudi[1], Arifiyanto Hadinegoro[1]
[1]Department of Informatics, University of Amikom Yogyakarta
email: [1]hasif.priyambudi@students.amikom.ac.id , [2]arifiyanto@amikom.ac.id

Abstract: Generating caption image automatically is one of the challenges in computer vision. This field can be very helpful in many ways, for example search engines. Currently there are many image classification algorithms that we can use to create a caption image model. In this article, we will compare performance between the Resnet50 and InceptionV3 models for text images. We will use 2000 (1800 train & 200 validation) image data and each image has 5 example captions to train the model. After the model is successfully created, we evaluate the model using 100 images and each image has 5 examples of additional captions that are not used in the training and validation process. The result of this research is that the InceptionV3 model is better than Resnet50. BLEU-1 is 0.53, BLEU-2 is 0.35, BLEU-3 is 0.18, BLEU-4 is 0.09, and METEOR is 0.35 for InceptionV3 model. While Resnet50 model has a value of BLEU-1 is 0.51, BLEU-2 is 0.31, BLEU-3 is 0.16, BLEU-4 is 0.06, and METEOR is 0.33.

Keywords: caption image; inceptionv3; LSTM; resnet50

Abstrak: Membuat gambar teks secara otomatis adalah salah satu tantangan dalam computer vision. Bidang ini bisa sangat membantu dalam banyak hal, misalnya mesin pencari. Saat ini banyak sekali algoritma klasifikasi citra yang dapat kita gunakan untuk membuat model teks citra. Pada artikel ini, kami akan membandingkan performa antara model Resnet50 dan InceptionV3 untuk gambar teks. Kami akan menggunakan 2000 (1800 train & 200 validation) data gambar dan setiap gambar memiliki 5 contoh caption untuk melatih model. Setelah model berhasil dibuat, kami mengevaluasi model menggunakan 100 gambar dan setiap gambar memiliki 5 contoh caption tambahan yang tidak digunakan dalam proses training dan validation. Hasil dari penelitian ini adalah model InceptionV3 lebih baik dibandingkan dengan Resnet50. BLEU-1 0.53, BLEU-2 0.35, BLEU-3 0.18, BLEU-4 0.09, dan METEOR 0.35 untuk model InceptionV3. Sedangkan model Resnet50 memiliki nilai BLEU-1 0.51, BLEU-2 0.31, BLEU-3 0.16, BLEU-4 0.06, dan METEOR 0.33.

Kata kunci: caption image; inceptionv3; LSTM; resnet50

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 3, June 2023, hlm. 521 - 528 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i3.2277
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Currently, technology is developing rapidly. For example artificial intelligence, artificial intelligence (AI) is profoundly changing our lives, and it is critical to understand these advances to predict future development strategies [1] [2] [3]. For example, artificial intelligence for caption image generators that can help us understand images. Although it is a very difficult endeavor, being able to automatically describe the substance of an image using well constructed English phrases could have significant effects [4][5].

The main objective of the sub-field of computer vision known as image captioning is to produce an accurate and natural text description of each scenario depicted in an image [6][7]. Although this is difficult to do, it can have a very good impact, for example helping search engines find relevant images. In this study, we will compare Inceptionv3 and Resnet50. So this research is based on transfer learning, the reason we use transfer learning is to train transformer models in language pairs with high resource availability, transfer learning can replace the need for a warm-up phase [8].

Resnet50 is great for image classification. For the detection of brain tumors, Resnet50 has an accuracy of 92% and 90% [9]. If added with LSTM, then Resnet50 is also good for image captions. The experimental results show the resnet50 model can produce quality image captions automatically [10]. Inceptionv3 is also great for classifying. The experimental results obtained a sensitivity score of 95.41% and a specificity of 80.09%, these results outperform other methods in terms of lung image categorization [11]. If added with LSTM, then Inceptionv3 is also good for image captions. Inceptionv3 can generate captions with a good BLEU score [12].

We will perform feature extraction using these two models. After that, we will insert it to LSTM. We choose LSTM because of its popularity and capacity to remember long-term dependencies in the created word sequence [13]. We will measure the quality of the output from LSTM in human language, we will use BLEU (Bilingual Evaluation Understudy). BLEU can significantly increase the correctness of a final translation [14][15]. For more convincing which model is better, we will also use the METEOR algorithm.

The final result of this article is which model is the best (Resnet50 and Inceptionv3) to caption images.

METHOD

An important element of research process is research methodology. Image 1 shows the methodology's flow for this study.

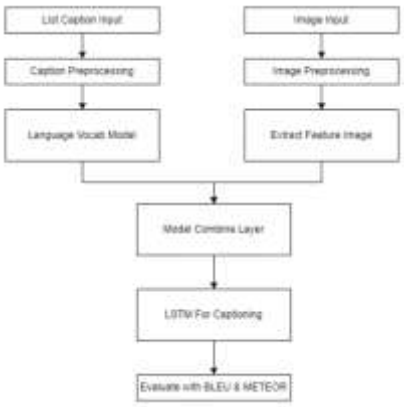


Image 1**.** Methodology Flow

Image Preprocessing

We used the Kaggle Flickr8K

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 3, June 2023, hlm. 521 - 528 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i3.2277
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

dataset for this research, but we only used 2000 image to see which model is better at handling less training data. After that, we divided the data into 1800 training data and 200 validation data. Image preprocessing is the first thing we do, the image is scaled to the necessary size.

We must execute two conversions with different sizes, because we will be comparing the performance of two different models here. The Resnet50 model will employ the sizes (224, 224, 3). Inceptionv3 will make advantage of size (299, 299, 3).

Caption Preprocessing

Text preprocessing used for data selection to make it more structured [16]. In Flickr8K dataset has 5 captions for each image. We will use the caption as the language vocab model. But we need to do preprocessing first. This stage includes case folding, remove special character, remove number and adding special token at the beginning and ending caption.

Case folding is the process of changing letters in a text into lowercase without changing the meaning or structure of the text. Case folding is a technique used to streamline text processing and boost efficiency. For an example see Image 2.

Image 2. Case folding

After that, the next process is to remove special characters and remove number. The purpose of this process is to improve the quality of captions.

For the process of adding special tokens, we add *startseq* at the beginning of the sentence, and *stopseq* at the end of the sentence.

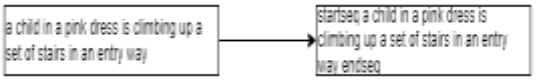
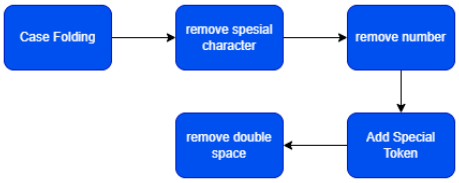


Image 3. Special Token

The final step of caption preprocessing is to replace double spaces with single spaces. This can happen because of the previous process.



Images 4. Caption Preprocessing

Image Feature Extract

There are two steps required to generate an automatic caption for an image [17]. First step is extract information from the image and save it to vector.

We will do two extraction processes using the Resnet50 and InceptionV3 models. Because the goal of this study is not images classification, we remove the last layer of this model for the feature extraction process. We don't need that layer because this one is in charge of classification images into one of 1000 possible groups.

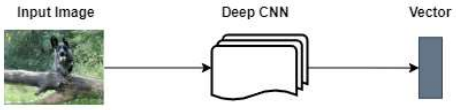


Image 5. Extract Feature Image

Language Vocab Model

This process includes counting the number of unique words in captions and looking at the maximum possible caption length.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 3, June 2023, hlm. 521 - 528 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i3.2277
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Model Combine Layer

This layer is useful for combining the image feature extract output layer and the language vocab model. The output of this process will be entered into the LSTM layer.

During training, we use a batch 128 and for optimization we use ADAM. We use a large batch size because a large batch size model can minimize loss more effectively than a small batch size model [18]. We train models for 20 epochs. But if the loss validation is no longer reduced, we will stop training process.

LSTM

RNN is a method in deep learning that is used to process sequential data such as sentences [16], [19]. One of the well-known models is the LSTM. LSTM (Long Short Term Memory) is very useful for making sentences from each image feature.

The vector obtained from the image feature extraction process will be entered into the LSTM. After that, LSTM will make a sentence and sentence will be transferred to the next LSTM layer and finally we get the sentence generated for the image.

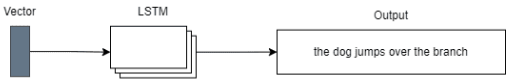


Image 6. LSTM process into sentences

Evaluate Model

The metric standard used for testing is called BLEU [14]. Because BLEU bases its decisions on n-gram precision, it harshly penalizes lexical deviations even when candidates are synonymous: No credit is given if an n-gram's subsequence does not exactly match the reference.

After getting the BLEU value, we will also measure the model output using METEOR. Besides considering the resulting semantic accuracy, METEOR also considers the resulting recall [20].

We will compare the accuracy of machine translation of both models (Resnet50 & InceptionV3) against human reference translation using BLEU & METEOR. After that, we can see which model is better.

RESULT AND DISCUSSION

After the training process, here is the history of training process (accuracy & validation) from Resnet50 and InceptionV3.

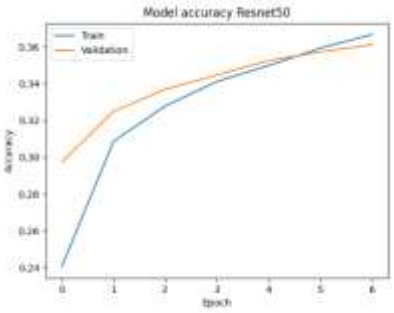


Image 7. Training Accuracy Resnet50

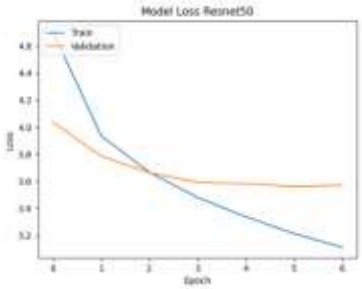


Image 8. Val Loss Resnet50

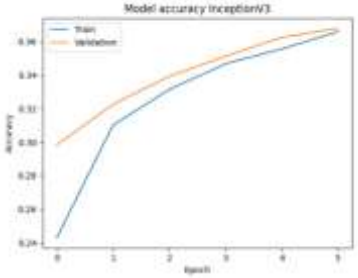


Image 9. Val Accuracy InceptionV3

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 3, June 2023, hlm. 521 - 528 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i3.2277
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

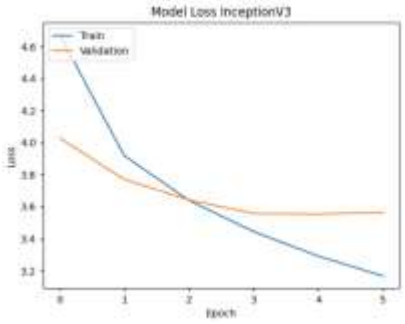


Image 10. Val Loss InceptionV3

We tested a model built using 100 images from the flickr8k dataset that were not used for training & validation. The BLEU & METEOR algorithm is used to see which model has the better caption output.

The BLEU-N (N = 1, 2, 3, 4) scores can be calculated using length of reference sentence, created phrase, uniform weights, and modified n-gram precision [10].

Table 1. BLEU-SCORES Report

Model	BLEU			
	Bleu1	Bleu2	Bleu3	Bleu4
Resnet50	0.51	0.31	0.16	0.06
Inception V3	0.53	0.35	0.18	0.09

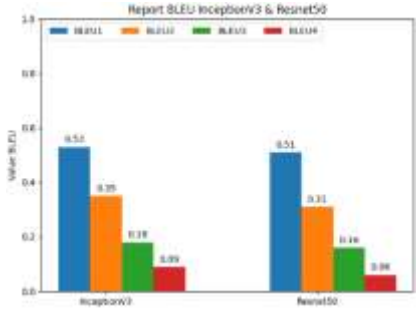


Image 11. Bar Graph BLEU-SCORES

Here we also include measurement results from the METEOR algorithm.

Table 2. METEOR-SCORES Report

Model	METEOR
InceptionV3	0.35
Resnet50	0.33

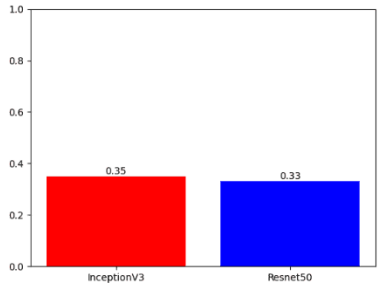


Image 12. Bar Graph METEOR-SCORES

Few samples of captions produces by Resnet50 model are illustrated in Image 13 and InceptionV3 model is illustrated in Image 14.

Image 13. Testing Resnet50 Model

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. IX No 3, June 2023, hlm. 521 - 528 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v9i3.2277
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Image 14. Testing InceptionV3 Model

CONCLUSION

In this research, we try to compare the Resnet50 and InceptionV3 models for text image generators. We evaluate the generated text using BLEU & METEOR. Resnet50 has a BLEU-1 is 0.51, BLEU-2 is 0.31, BLEU-3 is 0.16, BLEU-4 is 0.06, and METEOR is 0.33. While InceptionV3 has a BLEU-1 is 0.53, BLEU-2 is 0.35, BLEU-3 is 0.18, BLEU-4 is 0.09 and METEOR is 0.35. With a batch size 128 and ADAM optimization, the InceptionV3 model is better for image captions than the Resnet50 model. This can be proven by the scores of BLEU-1, BLEU-2, BLEU-3, BLEU-4, and METEOR InceptionV3 which are higher than Resnet50.

BIBLIOGRAPHY

[1] C. Luchini, A. Pea, and A. Scarpa, "Artificial intelligence in oncology: current applications and future perspectives," *British Journal of Cancer*, vol. 126, no. 1. Springer Nature, pp. 4–9, Jan. 01, 2022. doi: 10.1038/s41416-021-01633-1.

[2] A. Rahman *et al.*, "On the ICN-IoT with federated learning integration of communication: Concepts, security-privacy issues, applications, and future perspectives," *Future Generation Computer Systems*, vol. 138, pp. 61–88, Jan. 2023, doi: 10.1016/j.future.2022.08.004.

[3] G. Pinto, Z. Wang, A. Roy, T. Hong, and A. Capozzoli, "Transfer learning for smart buildings: A critical review of algorithms, applications, and future perspectives," *Advances in Applied Energy*, vol. 5, p. 100084, Feb. 2022, doi: 10.1016/j.adapen.2022.100084.

[4] M. Bhalekar and M. Bedekar, "D-CNN: A New model for Generating Image Captions with Text Extraction Using Deep Learning for Visually Challenged Individuals," *Engineering, Technology & Applied Science Research*, vol. 12, no. 2, pp. 8366–8373, Apr. 2022, doi: 10.48084/etasr.4772.

[5] O. Vinyals Google, A. Toshev Google, S. Bengio Google, and D. Erhan Google, "Show and Tell: A Neural Image Caption Generator."

[6] R. Castro, I. Pineda, W. Lim, and M. E. Morocho-Cayamcela, "Deep Learning Approaches Based on

Transformer Architectures for Image Captioning Tasks," *IEEE Access*, vol. 10, pp. 33679–33694, 2022, doi: 10.1109/ACCESS.2022.3161428.

[7] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-Based Image Captioning," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 1, pp. 539–559, Jan. 2023, doi: 10.1109/TPAMI.2022.3148210.

[8] A. Fikri Aji, N. Bogoychev, K. Heafield, and R. Sennrich, "In Neural Machine Translation, What Does Transfer Learning Transfer?" [Online]. Available: http://www.panl10n.net/english/

[9] A. K. Sharma, A. Nandal, A. Dhaka, D. Koundal, D. C. Bogatinoska, and H. Alyami, "Enhanced Watershed Segmentation Algorithm-Based Modified ResNet50 Model for Brain Tumor Detection," *Biomed Res Int*, vol. 2022, 2022, doi: 10.1155/2022/7348344.

[10] Y. Chu, X. Yue, L. Yu, M. Sergei, and Z. Wang, "Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention," *Wirel Commun Mob Comput*, vol. 2020, 2020, doi: 10.1155/2020/8909458.

[11] C. Wang *et al.*, "Pulmonary image classification based on inception-v3 transfer learning model," *IEEE Access*, vol. 7, pp. 146533–146541, 2019, doi: 10.1109/ACCESS.2019.2946000.

[12] A. Kumar Yadav, E. Joyal Nadar, K. Chaudhary, M. Pal, and A. Professor, "IMAGE CAPTION GENERATOR USING CNN AND RNN (LSTM)." [Online]. Available: www.irjmets.com

[13] R. Staniute and D. Šešok, "A systematic literature review on image captioning," *Applied Sciences (Switzerland)*, vol. 9, no. 10. MDPI AG, May 01, 2019. doi: 10.3390/app9102024.

[14] J. Wieting, T. Berg-Kirkpatrick, K. Gimpel, and G. Neubig, "Beyond BLEU: Training Neural Machine Translation with Semantic Similarity," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.06694

[15] L. Lei and H. Wang, "Design and Analysis of English Intelligent Translation System Based on Internet of Things and Big Data Model," *Comput Intell Neurosci*, vol. 2022, pp. 1–9, May 2022, doi: 10.1155/2022/6788813.

[16] Y. Fauziyah *et al.*, "MESIN PENTERJEMAH BAHASA INDONESIA-BAHASA SUNDA MENGGUNAKAN RECURRENT NEURAL NETWORKS," 2022. [Online]. Available: https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/index

[17] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, "Automatic Image and Video Caption Generation with Deep Learning: A Concise Review and Algorithmic Overlap," *IEEE Access*, vol. 8. Institute of Electrical and Electronics Engineers Inc., pp. 218386–218400, 2020. doi: 10.1109/ACCESS.2020.3042484.

[18] D. Setiawan, M. A. Coenradina Saffachrissa, S. Tamara, and D. Suhartono, "INTERNATIONAL JOURNAL ON INFORMATICS

VISUALIZATION journal homepage : www.joiv.org/index.php/joiv INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION Image Captioning with Style Using Generative Adversarial Networks." [Online]. Available: www.joiv.org/index.php/joiv

[19] A. Hanifa, S. A. Fauzan, M. Hikal, and M. B. Ashfiya, "PERBANDINGAN METODE LSTM DAN GRU (RNN) UNTUK KLASIFIKASI BERITA PALSU BERBAHASA INDONESIA COMPARISON OF LSTM AND GRU (RNN) METHODS FOR FAKE NEWS CLASSIFICATION IN INDONESIAN." [Online]. Available: https://covid19.go.id/p/hoax-buster.

[20] Y. Pan, L. Wang, S. Duan, X. Gan, and L. Hong, "Chinese image caption of Inceptionv4 and double-layer GRUs based on attention mechanism," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Apr. 2021. doi: 10.1088/1742-6596/1861/1/012044.