

# Keterangan Gambar Otomatis Berbahasa Indonesia dengan CNN dan LSTM

*by* Amiin Majiid Nugroho

---

**Submission date:** 24-Nov-2020 10:16PM (UTC+0700)

**Submission ID:** 1455870461

**File name:** kg\_otomatis.pdf (535.3K)

**Word count:** 1655

**Character count:** 10185

# Keterangan Gambar Otomatis Berbahasa Indonesia dengan CNN dan LSTM

**Abstract**—*Image captioning* adalah proses untuk menghasilkan suatu kalimat atau lebih untuk menjelaskan konten visual dari suatu gambar. *Image Captioning* bermanfaat untuk kebutuhan dimasa mendatang untuk membantu kegiatan manusia memahami konten visual seperti keterangan pada citra medis, interaksi manusia dengan robot dan membantu mendeskripsikan gambar kepada tunanetra. Untuk menghasilkan suatu deskripsi gambar dibutuhkan gabungan antara *Computr Vision* dan *Natural Language Proccsing*. Penelitian ini bertujuan untuk menghasilkan deskripsi gambar (*caption*) berbahasa Indonesia sekaligus menganalisis seberapa baik metode yang diterapkan dalam menghasilkan *caption* tersebut. Penelitian ini menggunakan *dataset* gambar dari MSCOCO. Pada penelitian ini menggunakan dua buah metode yaitu CNN dan LSTM untuk mengasilkan *caption* dari gambar. Beberapa *caption* sesuai dengan gambar yang ditampilkan dan hasil *caption* yang didapat pada penelitian ini memperoleh skor BLEU {1,2,3,4} sebesar {0.29, 0.47, 0.56, 0.60}.

**Keywords**—*image captioning, CNN, LSTM*

## I. PENDAHULUAN

Dalam beberapa tahun terakhir, visi komputer di bidang pemrosesan gambar telah membuat kemajuan yang signifikan, seperti klasifikasi gambar dan deteksi objek. Manfaat dari kemajuan pada bidang klasifikasi gambar dan deteksi objek menjadi memungkinkan untuk secara otomatis menghasilkan satu kalimat atau lebih untuk menjelaskan konten visual dari suatu gambar, yang dikenal sebagai *Image Captioning*. Membuat deskripsi gambar yang lengkap dan alami secara otomatis memiliki manfaat yang besar, seperti pada pembuatan judul yang dilampirkan pada gambar berita, deskripsi yang terkait dengan gambar medis, pengambilan gambar berbasis teks, informasi yang diakses oleh pengguna tunanetra, interaksi manusia-robot[1]. Dengan mendapat informasi berupa deskripsi teks dari konten visual seperti gambar, maka informasi yang diperoleh akan lebih mudah diolah.

Menghasilkan deskripsi dengan bahasa alami yang memiliki makna dari sebuah gambar membutuhkan tingkat pemahaman yang lebih tinggi dari klasifikasi dan deteksi gambar. Permasalahan tersebut sangat menarik karena menghubungkan antara *Computer Vision* dan *Natural Language Proccsiong* yang merupakan dua bidang utama dalam *Artificial Intelligence*[2].

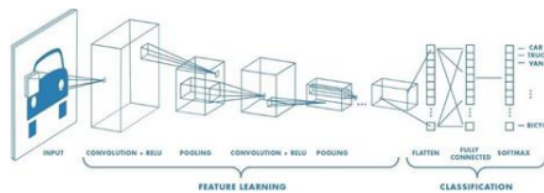
Pada penelitian kali ini, akan menggunakan dua metode yaitu *Convolutional Neural Network (CNN)* dan *Long Short Term Memory (LSTM)*. CNN digunakan untuk melakukan *encoding* pada gambar sedang LSTM yang merupakan sebuah

jaringan syaraf berulang akan digunakan untuk *gerete caption* [3].

Pada penelitian ini, diharapkan hasil dari metode CNN dan LSTM mampu menghasilkan deskripsi dari gambar dengan deskripsi berbahasa Indonesia. Selain mampu menghasilkan deskripsi Berbahasa Indonesia metode yang digunakan diharapkan mampu menghasilkan deskripsi dengan tata bahasa senatural mungkin seperti bahasa manusia. Untuk menguji seberapa baik deskripsi yang dihasilkan pada penelitian ini, hasil *caption* akan di evaluasi dengan skor BLEU.

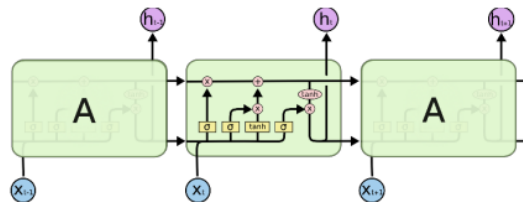
## II. TINJAUAN PUSTAKA

*Convolutional Neural Network (CNN)* pada penelitian ini digunakan untuk ekstraksi fitur pada citra. *Convolutional Neural Network (CNN)* merupakan jaringan saraf tiruan *feed-forward* dimana jaringan saraf tersebut mempelajari struktur hierarki dengan mempelajari representasi fitur internal dan menggeneralisasi fitur dalam sebuah gambar secara umum seperti pengenalan objek dan permasalahan *computer vision* lainnya [4]



Gambar 1 Arsitektur CNN

*Long Short Term Memory (LSTM)* adalah arsitektur dari RNN yang telah dimodifikasi sedemikian rupa untuk mengatasi penyimpanan memori dalam jangka waktu yang lama karena telah menambahkan *memory cell* [4].



Gambar 2 Arsitektur LSTM

Pada penelitian yang berkaitan dengan *image captioning* menggunakan Bahasa Arab, penggunaan metode gabungan CNN-LSTM mampu menghasilkan keluaran berupa *caption*

berbahasa Arab sekaligus memiliki akurasi yang lebih tinggi sejumlah 10% dari akurasi pada penelitian yang pernah dilakukan dengan bahasa arab sebelumnya. Penelitian ini menggunakan *dataset* Flickr8K dan MS COCO dan untuk ekstraksi fitur menggunakan VGG16[5].

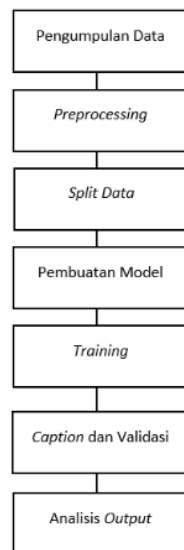
Pada penelitian gabungan antara CNN dan LSTM, pembuatan deskripsi gambar (*image captioning*) dengan menggunakan MS COCO *dataset* menghasilkan keterangan yang masuk akal untuk beberapa kasus[6].

Pada penelitian yang membahas tentang informasi pelengkap ini, diusulkan penggunaan jaringan fusi baru (RFnet) dengan tugas pembuatan teks pada gambar. Eksperimen dengan *dataset* MSCOCO mendemonstrasikan efektivitas RFnet[7].

Pada penelitian yang membandingkan CNN dan LSTM, pembuatan deskripsi gambar menggunakan CNN lebih efektif daripada penggunaan LSTM pada *dataset* MSCOCO ditambah dengan penggunaan waktu pelatihan yang lebih singkat[8].

### III. METODOLOGI PENELITIAN

Metodologi menjelaskan mengenai tahapan yang dilakukan dalam penelitian ini.



Gambar 3 Alur Penelitian.

#### A. Pengumpulan Data

Data yang digunakan adalah data berupa gambar yang diambil dari MSCOCO *dataset*. MSCOCO dipilih karena *dataset* ini berisi 91 objek dengan total 2,5 juta label dalam 328.000 gambar[9]. Data kemudian dipilih sejumlah 5000 gambar. Pada *dataset* tersebut setiap gambar sudah memiliki 5 deskripsi gambar (*caption*) berbahasa Inggris yang merepresentasikan isi gambar. Selanjutnya setiap *caption* tersebut diubah menjadi Bahasa Indonesia dengan menggunakan fungsi dari google translate.

#### B. Preprocessing

Langkah selanjutnya adalah melakukan tahap *preprocessing* terhadap data gambar dan data *caption*.

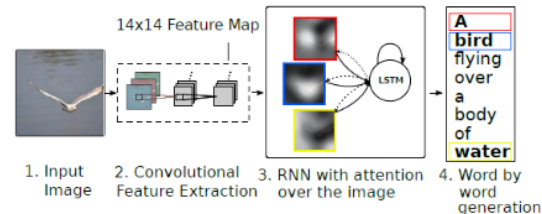
- *Preprocessing* gambar menggunakan NesNetMobile untuk melakukan ekstraksi *feature* setiap gambar dengan mengubah ukuran menjadi 224px \* 224px dan menormalkan gambar sehingga berisi piksel dalam kisaran -1 hingga 1, yang sesuai dengan format gambar yang digunakan untuk melatih NASNet.
- Inisialisasi NASNetMobile dan memuat bobot dari *pretrained* imagenet.
- Caching fitur yang diekstrak dari NASNetMobile.
- Setelah melakukan ekstraksi selanjutnya adalah *preprocessing caption* dan melakukan tokenisasi. Tujuannya untuk mendapatkan setiap kosa kata, selanjutnya setiap kosa kata dipilih 5000 kosa kata tertinggi.

#### C. Split Data

Selanjutnya adalah membagi data menjadi data *training* dan *validation* yang masing-masing dengan perbandingan 80:20 sehingga sejumlah 4000 adalah *training* dan 1000 adalah *validation*, kemudian untuk data dipilih secara acak.

#### D. Pembuatan Model

Pada penelitian kali ini menggunakan dua metode yaitu CNN dan LSTM. Metode CNN digunakan untuk *generate feature* pada gambar sedangkan LSTM digunakan untuk melakukan *generate caption*.



Gambar 4 Arsitektur Model

Dalam hal ini, ekstraksi fitur dari *convolutional layer* bawah NASNetMobile yang memberi bentuk vektor (7,7,1056) kemudian ditekan menjadi bentuk vektor (49,1056) kemudian vektor dilewatkan melalui Encoder CNN (yang terdiri dari *Fully connected layer*). Kemudian LSTM untuk memprediksi kata berikutnya.

#### E. Training

Setelah mendapatkan model, langkah selanjutnya adalah melakukan pelatihan terhadap gambar dan *caption* yang telah dipersiapkan. Pada encoder output, *hidden state* diinisialisasi menjadi 0.

#### F. Caption dan Validasi

Tahapan ini digunakan untuk *generate caption* yang menghasilkan prediksi dari *caption* berdasar model yang telah dibuat. Tahapan ini dilakukan untuk membandingkan hasil prediksi *caption* dibandingkan dengan *caption* yang sebenarnya dari gambar.

### G. Analisis Output

Untuk menganalisis ketepatan dari prediksi *caption* maka dilakukan analisa dengan menggunakan skor BLEU. BLEU (*Bilingual Evaluation Understudy*) merupakan sebuah metode yang digunakan untuk mengevaluasi otomatis terjemahan mesin [10]. Cara perhitungan skor BLEU adalah sebagai berikut:

$$\text{Unigram precision } P = \frac{m}{wt_t} \quad (1)$$

$$\text{Brievy penalty } P = f(x) = \begin{cases} 1, & \text{if } c > r \\ 3^{(1-\frac{r}{c})}, & \text{if } c \leq r \end{cases} \quad (2)$$

$$\text{BLEU} = p \cdot e^{\frac{N}{n-1} \left( \frac{1}{N} * \log P_n \right)} \quad (3)$$

## IV. HASIL DAN PEMBAHASAN

Bagian ini membahas mengenai penelitian yang telah dilakukan berdasarkan langkah-langkah yang telah dilalui.

### A. Caption yang Dihasilkan




Gambar 5 Validasi Caption

	Perbandingan Hasil	
	Real Caption	Prediction Caption
Gambar 3	seorang pemain bisbol di kemeja merah siap untuk memukul bola	seorang pria di bisbol memegang tongkat di tangannya

Berdasarkan prediksi *caption* yang dihasilkan dari model tersebut, hasil *caption* memiliki makna yang sama dan dapat mendeskripsikan isi gambar yang dihasilkan. Prediksi *caption* jika dilihat memang memiliki perbedaan kata dari *caption* yang sebenarnya. Selanjutnya adalah melakukan perhitungan dengan BLEU skor. Hasil yang didapatkan dari BLEU {1,2,3,4} adalah sebagai berikut {0.29, 0.47, 0.56, 0.60}.

### B. Eksperimen

Pada tahapan ini dilakukan percobaan dengan menerapkan metode untuk melakukan *generate caption* pada gambar acak yang ada di internet, didapatkan hasil sebagai berikut:

		Eksperimen	
		Gambar	Prediction Caption
1		seorang pria berdiri di samping bangku di jalan	
2		sebuah kursi dekat jendela di samping meja	
3		seorang wanita memegang raket tenis berusaha untuk memukul bola	
4		sebuah truk besar diparkir di samping garasi	
5		seekor kucing hitam dan putih di tempat tidur dengan	
6		sebuah kereta hitam dan hitam di trek dalam stasiun	

Dilihat dari hasil *generate caption* beberapa *caption* yang dihasilkan sesuai dengan gambar tapi masih terdapat *caption* yang kurang sesuai. Deskripsi yang dihasilkan pada gambar masih terdapat kata yang kurang tepat peletakannya sehingga kalimat terlihat aneh.

## V. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan maka dapat disimpulkan bahwa metode yang digunakan yaitu CNN dan LSTM mampu menghasilkan keterangan gambar (*caption*) Berbahasa Indonesia dengan cukup baik walaupun masih mengalami kekurangan seperti ketepatan dengan gambar. Kekurangan lain pada metode tersebut adalah *caption* yang dihasilkan terkadang kurang lengkap secara kosa kata dan

terlihat ambigu. Skor BLEU {1,2,3,4} yang dihasilkan adalah {0.29, 0.47, 0.56, 0.60}. Dilihat dari hasil yang telah didapatkan sangat disarankan untuk menggunakan data yang lebih besar lagi untuk menghasilkan keterangan gambar yang lebih akurat. Selain itu disarankan menggunakan *caption* dengan tata bahasa yang sudah benar dan bukan hasil dari *google translate*.

#### VI. REFERENCES

- [1] S. Liu, L. Bai, Y. Hu, and H. Wang, "Image Captioning Based on Deep Neural Networks," *MATEC Web Conf.*, vol. 232, pp. 1–7, 2018.
- [2] Q. You *et al.*, "Image Captioning with Semantic Attention," *Rbi*, 2016.
- [3] Y. Pu *et al.*, "Variational autoencoder for deep learning of images, labels and captions," *Adv. Neural Inf. Process. Syst.*, no. Nips, pp. 2360–2368, 2016.
- [4] N. K. Manaswi, *Deep Learning with Applications Using Python*. 2018.
- [5] H. A. Al-muzaini, T. N. Al-yahya, and H. Benhidour, "Automatic Arabic image captioning using RNN-LSTM-based language model and CNN," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, pp. 67–73, 2018.
- [6] J. Gu, G. Wang, J. Cai, and T. Chen, "An Empirical Study of Language CNN for Image Captioning," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-October, pp. 1231–1240, 2017.
- [7] W. Jiang, L. Ma, Y. G. Jiang, W. Liu, and T. Zhang, "Recurrent Fusion Network for Image Captioning," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11206 LNCS, pp. 510–526, 2018.
- [8] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 5561–5570, 2018.
- [9] T. Y. Lin *et al.*, "Microsoft COCO: Common objects in context," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8693 LNCS, no. PART 5, pp. 740–755, 2014.
- [10] K. Papineni, S. Roukos, T. Ward, W. Zhu, and Y. J. Huang, "IBM Research Report Bleu : a Method for Automatic Evaluation of Machine Translation," *Science (80-. )*, vol. 22176, no. February, pp. 1–10, 2001.

# Keterangan Gambar Otomatis Berbahasa Indonesia dengan CNN dan LSTM

## ORIGINALITY REPORT

12%

SIMILARITY INDEX

10%

INTERNET SOURCES

11%

PUBLICATIONS

0%

STUDENT PAPERS

## PRIMARY SOURCES

- 1 [www.cs.stir.ac.uk](http://www.cs.stir.ac.uk) 2%  
Internet Source
- 2 Abdulkareem Alsudais. "Image Classification in Arabic: Exploring Direct English to Arabic Translations", IEEE Access, 2019 1%  
Publication
- 3 Ariyo Oluwasanmi, Enoch Frimpong, Muhammad Umar Aftab, Edward Y. Baagyere, Zhiguang Qin, Kifayat Ullah. "Fully Convolutional CaptionNet: Siamese Difference Captioning Attention Model", IEEE Access, 2019 1%  
Publication
- 4 [fandango-project.eu](http://fandango-project.eu) 1%  
Internet Source
- 5 Min Yang, Junhao Liu, Ying Shen, Zhou Zhao, Xiaojun Chen, Qingyao Wu, Chengming Li. "An Ensemble of Generation- and Retrieval-Based Image Captioning With Dual Generator Generative Adversarial Network", IEEE 1%

# Transactions on Image Processing, 2020

Publication

- 
- |    |  |    |
|----|--|----|
| 6  | M Abu, N A H Zahri, A Amir, I Ismail. "Machine Learning Algorithms for Optic Pathway Disease Diagnostics: A Review", IOP Conference Series: Materials Science and Engineering, 2020<br>Publication | 1% |
| 7  | Jian Wang, Jie Feng. "Hybrid Attention Distribution and Factorized Embedding Matrix in Image Captioning", IEEE Access, 2020<br>Publication   | 1% |
| 8  | <a href="http://www.coursehero.com">www.coursehero.com</a><br>Internet Source  | 1% |
| 9  | <a href="http://library.binus.ac.id">library.binus.ac.id</a><br>Internet Source  | 1% |
| 10 | Preethi Kurian, Vijay Jeyakumar. "Multimodality medical image retrieval using convolutional neural network", Elsevier BV, 2020<br>Publication  | 1% |
| 11 | <a href="http://rita.det.uvigo.es">rita.det.uvigo.es</a><br>Internet Source  | 1% |
| 12 | <a href="http://mafiadoc.com">mafiadoc.com</a><br>Internet Source  | 1% |
| 13 | <a href="http://www.groundai.com">www.groundai.com</a><br>Internet Source  | 1% |
-

14 eprints.umm.ac.id  
Internet Source

<1%

---

15 zombiedoc.com  
Internet Source

<1%

---

Exclude quotes On

Exclude matches Off

Exclude bibliography On