# Aplikasi *Image Captioning* Menggunakan *Convolutional Neural Network* dan *Long Short Term Memory*

**Javier Jean Vito Sengka*[1], Oktavian Abraham Lantang[2], Muhamad Dwisnanto Putro[3]**
[1,2,3]Universitas Sam Ratulangi Manado, Jalan Kampus Bahu, 95115
e-mail: *[1]sengkavito@gmail.com, [2]oktavian_lantang@unsrat.ac.id, [3]dwisnantoputro@unsrat.ac.id

***Abstrak***

*Image captioning merupakan teknologi yang menghasilkan deskripsi otomatis dari gambar dengan menggabungkan computer vision dan pemrosesan bahasa alami. Penelitian ini bertujuan membangun aplikasi image captioning berbahasa Indonesia dengan menerapkan arsitektur Convolutional Neural Network (CNN) dan Long Short-Term Memory (LSTM). Proses dimulai dengan ekstraksi fitur visual menggunakan model VGG16, kemudian dilanjutkan dengan pelatihan model captioning berbasis LSTM. Dataset yang digunakan adalah Flickr8k yang telah diterjemahkan ke dalam bahasa Indonesia. Evaluasi performa dilakukan menggunakan metrik BLEU, METEOR, CIDEr, dan ROUGE. Hasil menunjukkan bahwa model mampu menghasilkan deskripsi gambar dengan nilai BLEU-1 sebesar 0,6216, yang menandakan tingkat kesesuaian cukup baik antara caption hasil prediksi dan referensi. Selain itu, sistem telah diimplementasikan ke dalam aplikasi web berbasis Streamlit agar dapat diakses secara praktis. Model yang dirancang menunjukkan kemampuan menghasilkan caption yang informatif dan sesuai konteks visual gambar dalam bahasa Indonesia.*

***Kata kunci****— Aplikasi, Convolutional Neural Network, Image Captioning, Long Short Term Memory, Visi Komputer*

***Abstract***

*Image captioning is a technology that automatically generates descriptions of images by combining computer vision and natural language processing. This study aims to develop an Indonesian-language image captioning application by implementing a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture. The process begins with visual feature extraction using the VGG16 model, followed by training an LSTM-based captioning model. The dataset used is Flickr8k, which has been translated into Indonesian. Performance evaluation is conducted using BLEU, METEOR, CIDEr, and ROUGE metrics. The results show that the model is able to generate image descriptions with a BLEU-1 score of 0.6216, indicating a fairly good level of agreement between the predicted captions and the reference captions. In addition, the system has been implemented into a Streamlit-based web application to enable practical accessibility. The proposed model demonstrates the ability to generate informative captions that are consistent with the visual context of the images in the Indonesian language.*

***Keywords****—Application, Convolutional Neural Network, Image Captioning, Long Short Term Memory, Computer Vision*

## 1. PENDAHULUAN

*Image captioning* adalah proses menghasilkan deskripsi tekstual yang alami dan informatif dari sebuah gambar. Teknologi ini menggabungkan bidang *computer vision* dan *natural language processing* (NLP) untuk menjembatani informasi visual dan tekstual [1], [2]. *Computer vision* berperan dalam "memahami" konten visual gambar, sementara NLP berperan dalam "menerjemahkan" pemahaman tersebut ke dalam bahasa manusia yang natural dan mudah dipahami. *Image captioning* memiliki potensi besar untuk diterapkan di berbagai bidang, mulai dari membantu tunanetra "melihat" dunia melalui deskripsi teks, meningkatkan akurasi pencarian gambar, menganalisis konten dan sentimen di media sosial, hingga memungkinkan komputer "memahami" dan "menjelaskan" gambar kepada manusia [3], [4].

Meskipun *image captioning* telah banyak diteliti dan dikembangkan, penerapannya dalam bahasa Indonesia masih terbatas [5], [6]. Padahal, *image captioning* berbahasa Indonesia memiliki peran krusial, misalnya dalam meningkatkan aksesibilitas informasi bagi penyandang tunanetra di Indonesia. Dengan *image captioning*, tunanetra dapat mengakses berbagai jenis konten visual, seperti gambar di website, media sosial, atau buku elektronik, yang sebelumnya sulit diakses [7], [8]. *Image captioning* akan menghasilkan deskripsi teks dari gambar yang kemudian dapat diakses oleh tunanetra melalui berbagai cara, seperti *screen reader* yang membacakan deskripsi tersebut,

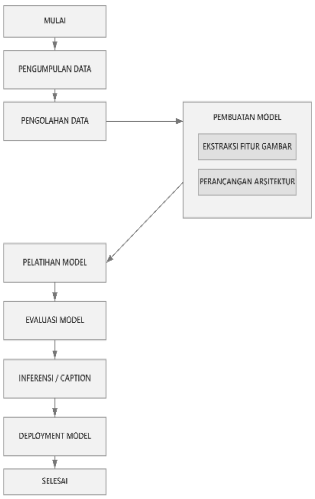
*Braille display* yang menampilkan deskripsi dalam huruf Braille, atau dikombinasikan dengan *audio description* untuk memberikan informasi visual yang lebih lengkap [9], [10]. Selain itu, *image captioning* berbahasa Indonesia dapat mendukung personalisasi konten lokal, menghasilkan deskripsi yang lebih relevan dan mudah dipahami oleh pengguna Indonesia [11], [12]. *Image captioning* juga dapat meningkatkan akurasi pencarian gambar berbahasa Indonesia dengan memberikan informasi teks yang lebih akurat pada mesin pencari, serta membantu memahami konten visual di media sosial berbahasa Indonesia, misalnya untuk analisis sentimen atau pemahaman opini publik [13], [14].

Namun, pengembangan sistem *image captioning* yang akurat dan natural dalam bahasa Indonesia menghadapi beberapa tantangan. Salah satu tantangan utama adalah keterbatasan dataset *image captioning* berbahasa Indonesia [15], [16]. Hal ini menyulitkan pelatihan model yang efektif, karena model membutuhkan data yang cukup dan representatif untuk mempelajari pola-pola bahasa dan visual, sejalan dengan temuan penelitian image captioning dalam bahasa Indonesia masih minim dan kekurangan sumber data [17], [18]. Selain itu, bahasa Indonesia memiliki struktur gramatikal dan variasi kosakata yang kompleks, sehingga model perlu dilatih dengan data yang cukup dan strategi yang tepat untuk menghasilkan *caption* yang natural dan gramatikal. Tantangan lainnya adalah variasi gambar yang sangat besar sehingga model *image captioning* perlu mampu menangani variasi tersebut dan menghasilkan *caption* yang akurat dan relevan [19], [20].

Penelitian ini berfokus pada pembangunan model *image captioning* berbasis CNN-LSTM yang mampu menghasilkan *caption* gambar berbahasa Indonesia secara akurat dan natural. Pendekatan CNN-LSTM dipilih karena telah terbukti unggul dalam menghasilkan *caption* yang akurat dan sekuensial CNN akan mengekstrak fitur-fitur penting dari gambar, sedangkan LSTM akan membentuk *caption* yang terstruktur dan informatif dari fitur-fitur tersebut. Model *image captioning* yang dihasilkan diharapkan memiliki kemampuan seperti akurasi yang tinggi, *natural language generation* yang baik, dan generalisasi yang baik untuk berbagai jenis gambar. Model *image captioning* ini diharapkan dapat diterapkan pada aplikasi *mobile*, sehingga dapat meningkatkan aksesibilitas informasi, mendukung personalisasi konten lokal, dan memperluas penerapan *image captioning* di berbagai bidang.

## 2. METODE PENELITIAN

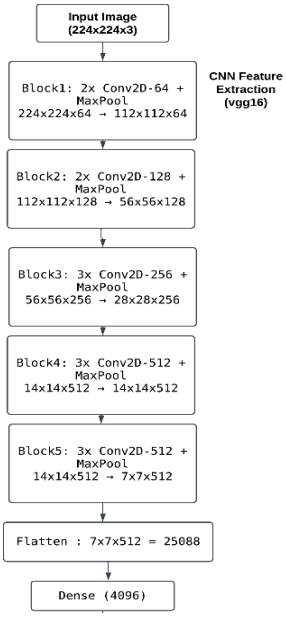
Metode yang digunakan dalam penelitian ini mengikuti alur kerja yang ditunjukkan pada Gambar 1, di mana seluruh proses dimulai dari pengumpulan data, pengolahan data, pengembangan model, pelatihan, hingga tahap evaluasi dan deployment. Dataset yang digunakan adalah Flickr8k yang telah diterjemahkan ke dalam bahasa Indonesia, berisi 8.091 gambar yang masing-masing memiliki lima caption deskriptif. Dataset ini dipilih karena telah banyak digunakan dalam penelitian image captioning dan mendukung pengembangan sistem berbasis bahasa Indonesia.




Gambar 1. Alur Penelitian

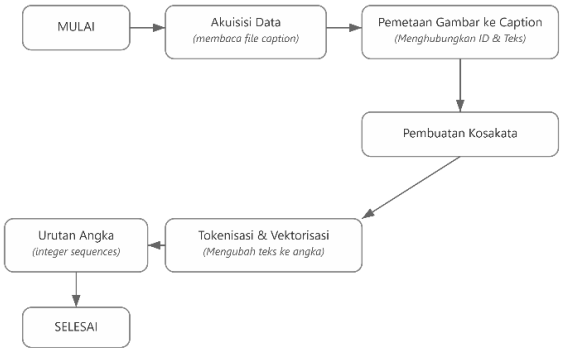
Pengolahan data dimulai dari tahap preprocessing gambar, di mana seluruh citra diubah menjadi ukuran 224×224 piksel mengikuti standar input VGG16 [19], [21]. Gambar kemudian dikonversi menjadi array NumPy, disesuaikan dimensinya menjadi (1,224,224,3), serta dinormalisasi menggunakan fungsi preprocess_input() sebelum diekstraksi fiturnya. Proses ekstraksi fitur menggunakan VGG16 yang telah dimodifikasi untuk hanya

mengambil keluaran dari dense layer berukuran 4096 neuron. Representasi fitur ini digunakan sebagai input visual untuk model captioning. Alur ekstraksi fitur tersebut ditunjukkan pada Gambar 2.



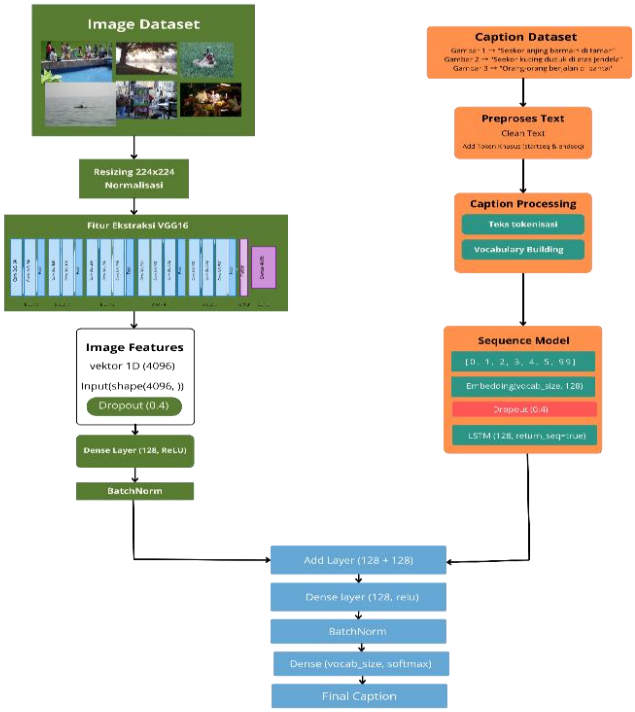

Gambar 2. Proses Ekstraksi Fitur gambar

Pemrosesan terhadap caption dilakukan dengan mengakuisisi data deskripsi dari metadata.csv dan memetakan setiap gambar ke caption yang relevan. Proses pembersihan teks meliputi normalisasi huruf kecil, penghapusan karakter non-alfabet, penghilang spasi berlebih, serta penambahan token khusus 'startseq' dan 'endseq'. Kosakata dibentuk dari seluruh kata unik dalam dataset, kemudian diterapkan tokenisasi untuk mengubah setiap caption menjadi urutan indeks numerik. Panjang caption diseragamkan berdasarkan panjang maksimum untuk memastikan konsistensi input model. Alur pemrosesan caption dapat dilihat pada Gambar 3.




Gambar 3. Alur Pemrosesan Caption

Pengembangan model dilakukan dengan menggabungkan representasi gambar dan teks dalam arsitektur CNN–LSTM. Representasi visual berukuran 4096 dari VGG16 diproses melalui dense layer dengan 128 neuron dan aktivasi ReLU, disertai batch normalization. Sementara itu, caption dalam bentuk urutan indeks melewati embedding layer berdimensi 128, kemudian diproses menggunakan LSTM dengan 128 unit untuk menangkap konteks sekuens kata. Kedua keluaran ini digabungkan melalui operasi penjumlahan (Add layer), lalu diproses melalui dense layer tambahan sebelum akhirnya diteruskan ke dense layer akhir dengan aktivasi softmax untuk memprediksi kata berikutnya. Arsitektur lengkap model ditunjukkan pada Gambar 4.




Gambar 4. Arsitektur yang Digunakan pada penelitian

Pelatihan model dilakukan dengan membagi data ke dalam set pelatihan, validasi, dan pengujian guna memastikan kemampuan generalisasi model terhadap data baru. Fungsi loss yang digunakan adalah Categorical Cross-Entropy untuk membandingkan distribusi probabilitas kata hasil prediksi dengan target referensi [22], [23]. Optimisasi dilakukan menggunakan Adam Optimizer yang menyesuaikan learning rate secara adaptif. Regularisasi ditambahkan melalui Dropout pada LSTM serta Early Stopping yang menghentikan pelatihan apabila validation loss tidak mengalami peningkatan selama lima epoch. Pengaturan hyperparameter mencakup ReduceLROnPlateau dengan penurunan learning rate sebesar 20%, penggunaan batch size 32, pelatihan hingga 100 epoch dengan Early Stopping, panjang maksimum caption 25 kata, dan ukuran kosakata hasil tokenisasi.

Tahap evaluasi model dilakukan dengan menggunakan beberapa metrik yang lazim digunakan dalam penelitian image captioning, yaitu BLEU, METEOR, ROUGE, dan CIDEr. Keempat metrik ini digunakan untuk menilai kualitas caption yang dihasilkan berdasarkan kesesuaian dengan referensi dan aspek linguistik lainnya [24], [25]. Tahap akhir berupa deployment dilakukan menggunakan Streamlit untuk mengimplementasikan model ke dalam aplikasi web yang interaktif sehingga pengguna dapat mengunggah gambar dan langsung memperoleh hasil caption secara real time. Pendekatan ini memastikan bahwa model tidak hanya diuji secara akademik, tetapi juga siap digunakan sebagai sistem operasional yang dapat diakses oleh pengguna akhir.

## 3. HASIL DAN PEMBAHASAN

### 3.1 Data Set

Dataset yang digunakan dalam penelitian ini adalah Flickr 8k, yang terdiri dari 8.091 gambar dengan 5 caption deskriptif untuk setiap gambarnya. Dataset ini merupakan dataset standar yang sering digunakan dalam penelitian image captioning. Gambar 5 adalah salah satu contoh gambar dataset Flickr 8k

Gambar perlu dibagi menjadi tiga subset: data latih (training), data validasi (validation), dan data uji (test). Pembagian ini bertujuan untuk melatih model, memantau performanya selama pelatihan, dan menguji kemampuan generalisasi model pada data yang belum pernah dilihat sebelumnya.

Pada penelitian ini, pembagian data dilakukan dengan proporsi seperti pada Tabel 1.

Tabel 1. Pemisahan Data

| **Set** | **Gambar** | **Caption** |
|---|---|---|
| Train | 6.473 | 32.365 |
| Validation (Val) | 809 | 4.045 |
| Test | 809 | 4.045 |
| **Total** | 8.091 | 40.455 |

### 3.2 Hasil Pengolahan Data

#### 3.2.1. Hasil Pengolahan Gambar

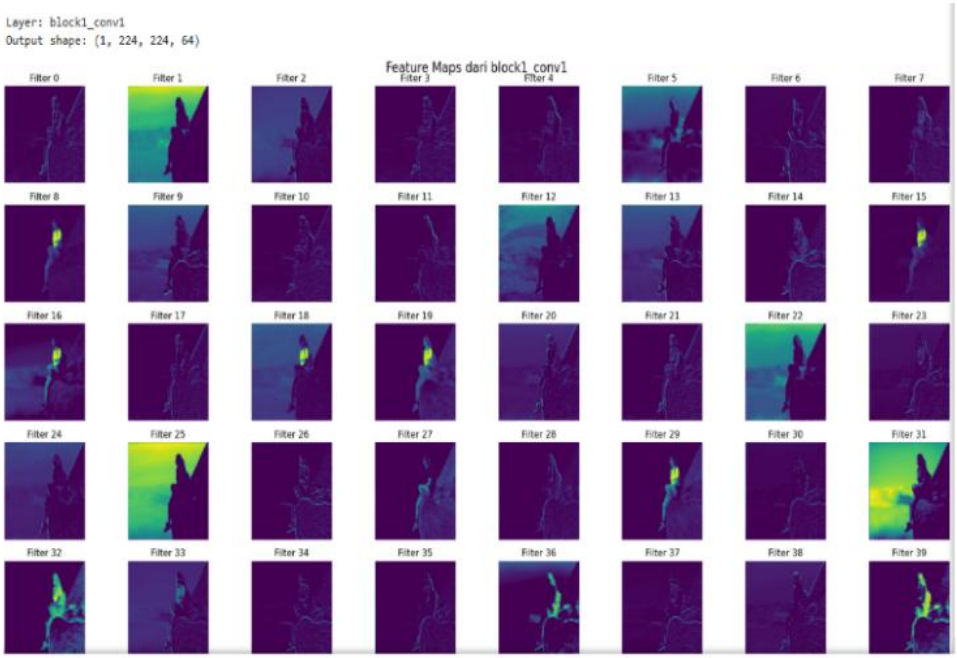
Seluruh data gambar telah diproses menggunakan model VGG16 untuk mengekstraksi fitur visual. Hasil ekstraksi berupa vektor berdimensi 4096. Contoh visualisasi sebagian hasil ekstraksi dapat dilihat pada Gambar 5. feature maps di layer awal menggambarkan bagaimana model mulai mengenali elemen dasar dalam gambar, seperti tepi, garis, dan pola sederhana. Hal ini menegaskan bahwa model CNN VGG16 mampu menangkap pola visual awal melalui filter konvolusi, yang memberikan pemahaman mendalam tentang proses pembelajaran fitur dasar sebelum model mendeteksi pola yang lebih kompleks di lapisan berikutnya.

#### 3.2.2. Hasil Pengolahan Caption

Kemudian proses data caption telah melalui proses pembersihan dan tokenisasi. Setiap caption ditambahkan token khusus startseq dan endseq, lalu dikonversi menjadi urutan angka berdasarkan kamus kata vocabulary yang terbentuk. Tabel II menyajikan gambaran lengkap proses tokenisasi dan text-to-sequence pada kalimat asli yang digunakan dalam dataset penelitian.

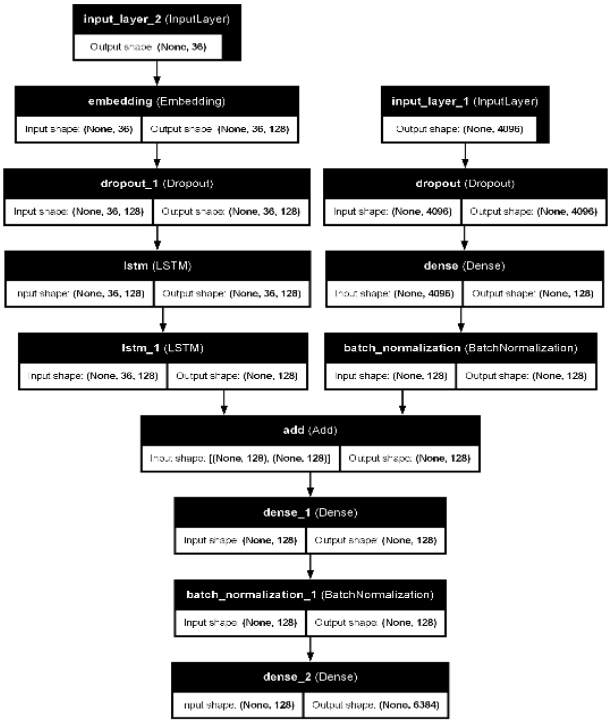
Tabel 2. Gambaran Hasil Tokenisasi dan Text-to-Sequence

| Kalimat Asli | Kalimat Setelah Tokenisasi | Hasil Text-to-Sequence |
|---|---|---|
| Seorang anak berpakaian merah muda sedang menaiki tangga di jalan masuk. | 'startseq', 'seorang', 'anak', 'berpakaian', 'merah', 'muda', 'sedang', 'menaiki', 'tangga', 'di', 'jalan', 'masuk', 'endseq' | [1, 4, 10, 75, 21, 36, 7, 194, 165, 3, 38, 545, 2] |
| Seorang gadis memasuki sebuah bangunan kayu. | 'startseq', 'seorang', 'gadis', 'memasuki', 'sebuah', 'bangunan', 'kayu', 'endseq' | [1, 4, 16, 1353, 42, 207, 135, 2] |
| Seorang gadis kecil naik ke rumah bermain kayu. | 'startseq', 'seorang', 'gadis', 'kecil', 'naik', 'ke', 'rumah', 'bermain', 'kayu', 'endseq' | [1, 4, 16, 25, 432, 20, 263, 19, 135, 2] |
| Seorang gadis kecil menaiki tangga menuju rumah bermainnya. | 'startseq', 'seorang', 'gadis', 'kecil', 'menaiki', 'tangga', 'menuju', 'rumah', 'bermainnya', 'endseq' | [1, 4, 16, 25, 194, 165, 195, 263, 2595, 2] |
| Seorang gadis kecil berpakaian merah muda masuk ke kabin kayu. | 'startseq', 'seorang', 'gadis', 'kecil', 'berpakaian', 'merah', 'muda', 'masuk', 'ke', 'kabin', 'kayu', 'endseq' | [1, 4, 16, 25, 75, 21, 36, 545, 20, 2596, 135, 2] |




Gambar 5. Visualisasi proses ekstraksi fitur gambar

### 3.3 Hasil Pemodelan

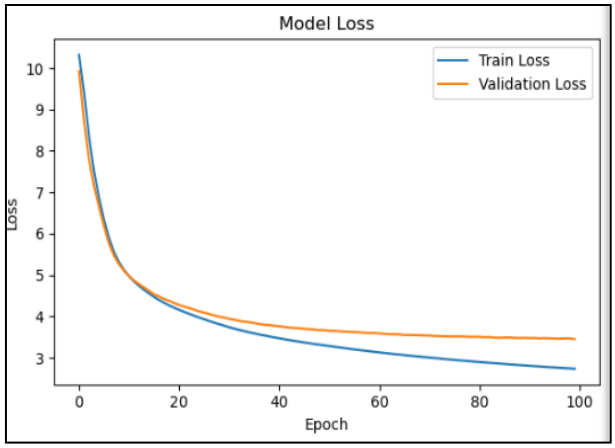



Gambar 6. Plot Model yang dihasilkan

### 3.4 Hasil Training Model

Tabel 3. Hasil Training Model

| Metrik | Nilai |
|---|---|
| Training Loss | 2.7452 |
| Validation Loss | 3.4510 |




Gambar 7. Nilai Loss Model

Model dilatih menggunakan algoritma optimasi Adam dengan learning rate 0.00005 dan categorical cross-entropy sebagai fungsi loss. Batch size ditetapkan sebesar 32, dan pelatihan dilakukan hingga maksimum 100 epoch dengan callback EarlyStopping untuk menghentikan pelatihan jika val_loss tidak meningkat selama 5 epoch berturut-turut, serta ReduceLROnPlateau untuk menurunkan learning rate jika val_loss stagnan. Tabel 3 dan Gambar 8 menunjukkan hasil pelatihan dengan nilai training loss mencapai 2.7452 dan validation loss 3.4510. Grafik model loss memperlihatkan penurunan yang signifikan pada kedua nilai loss selama proses pelatihan, dimana training loss terus menurun hingga sekitar 2.7, sementara validation loss stabil pada nilai sekitar 3.4, mengindikasikan model yang telah konvergen dengan selisih yang wajar antara performa pada data training dan validasi.

### 3.5 Evaluasi Performa Model

Evaluasi komprehensif terhadap model dilakukan menggunakan berbagai metrik standar dalam bidang

natural language generation dan image captioning. Hasil evaluasi ditampilkan dalam beberapa tabel berikut.

Tabel 4. Hasil Evaluasi Blue Score

| **Metrik** | **Nilai** |
|---|---|
| BLEU-1 | 0.6216 |
| BLEU-2 | 0.4352 |
| BLEU-3 | 0.3067 |
| BLEU-4 | 0.2030 |
| METEOR | 0.4617 |
| CIDEr | 0.5166 |
| ROUGE-1 | 0.2085 |
| ROUGE-2 | 0.0941 |
| ROUGE-L | 0.1958 |

BLEU score pada Tabel 4 menunjukkan kemampuan model untuk menghasilkan n-grams yang relevan dibandingkan dengan referensi. Nilai BLEU-1 sebesar 0.621579 menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam menghasilkan kata-kata individu (1-gram) yang sesuai. Namun, nilai ini menurun secara bertahap untuk n-grams yang lebih tinggi (BLEU-2 hingga BLEU-4), menunjukkan bahwa model menghadapi tantangan dalam menangkap hubungan antar kata atau frasa yang lebih kompleks. Hal ini dapat disebabkan oleh keterbatasan dalam dataset atau model yang belum optimal untuk struktur bahasa tertentu.

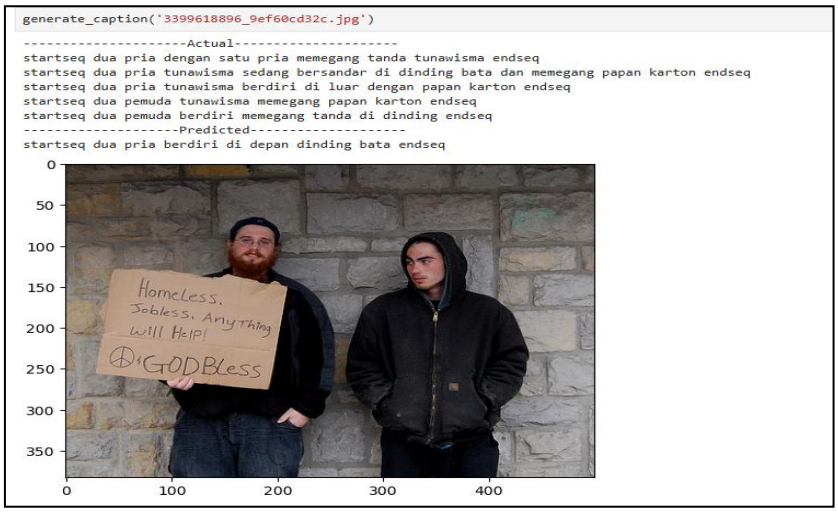
METEOR score yang diperoleh pada Tabel IV adalah 0.4617, yang mengindikasikan bahwa model cukup baik dalam mencocokkan kata-kata dengan referensi, termasuk sinonim, stemmed words, dan urutan kata. Dibandingkan dengan BLEU, METEOR lebih memperhatikan aspek semantik sehingga hasil ini menunjukkan kemampuan model untuk memahami makna kata-kata dalam konteks. Meskipun nilainya cukup baik, masih ada ruang untuk peningkatan, khususnya dalam menangani variasi bahasa yang lebih kompleks.

Nilai ini menunjukkan bahwa model dapat menghasilkan caption yang cukup serupa dengan referensi, meskipun masih terdapat beberapa ketidaksesuaian. Penggunaan CIDEr sangat relevan untuk evaluasi image captioning karena mempertimbangkan frekuensi kemunculan kata unik dalam referensi.

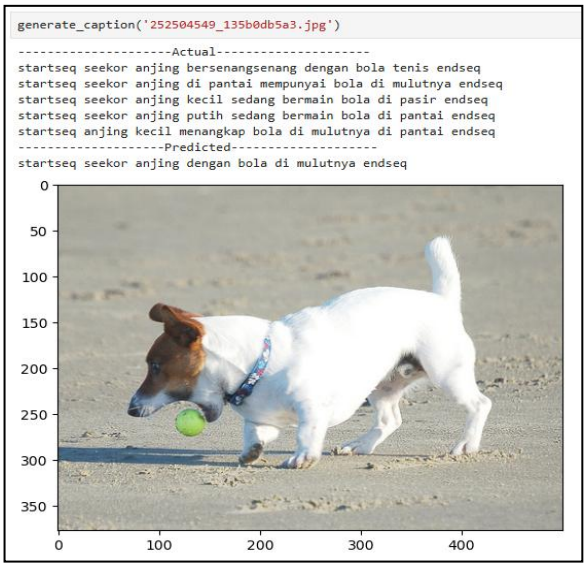
ROUGE score pada tabel IV mengukur kesamaan antara caption yang dihasilkan dengan referensi berdasarkan n-grams dan subsekuensi kata. Nilai ROUGE-1 sebesar 0.2085 dan ROUGE-2 sebesar 0.0941 menunjukkan bahwa model memiliki kemampuan terbatas dalam mencocokkan n-grams spesifik, sedangkan ROUGE-L sebesar 0.1958 mengindikasikan bahwa model cukup baik dalam menangkap urutan kata terpanjang yang sesuai. Nilai yang rendah pada ROUGE-2 menunjukkan bahwa model belum sepenuhnya optimal dalam mencocokkan frasa yang lebih kompleks.

### 3.6 Analisis Hasil Prediksi Model

Untuk mengevaluasi performa model secara kualitatif, dilakukan analisis terhadap beberapa sampel hasil prediksi dengan menggunakan data testing.




Gambar 8. Hasil Prediksi Model



Gambar 9. Hasil Prediksi Model

Dari analisis terhadap sampel terhadap Gambar 8 dan Gambar 9, dapat disimpulkan bahwa model memiliki kemampuan untuk menghasilkan deskripsi yang relevan secara visual dengan gambar. Namun, terdapat keterbatasan dalam menangkap detail kecil, konteks spesifik, dan elemen penting dalam gambar. Performa model dapat ditingkatkan dengan memperluas dataset, menambahkan anotasi lebih mendetail, atau menerapkan teknik fine-tuning untuk pengenalan objek dan konteks yang lebih baik.

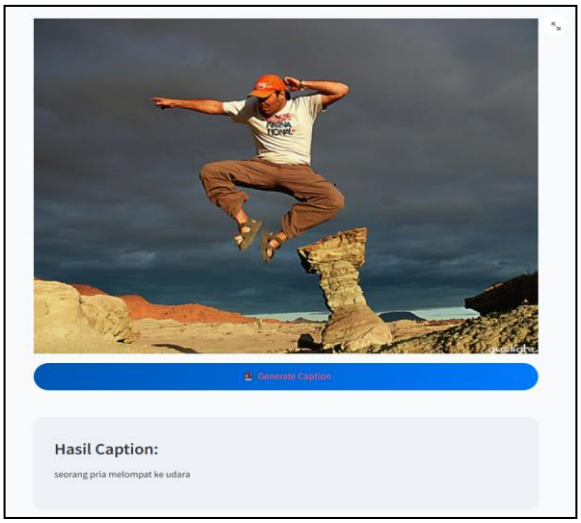
### 3.7 Hasil Deployment

Model yang telah dilatih kemudian disimpan dalam format .h5 menggunakan TensorFlow atau Keras untuk memudahkan integrasi. Selain itu, tokenizer yang berfungsi memproses teks pada model juga disimpan dalam format ".pkl" (pickle). Proses ini memastikan bahwa seluruh elemen yang dibutuhkan untuk prediksi berjalan dengan lancar pada aplikasi. Model diimplementasikan dalam sebuah aplikasi berbasis Streamlit yang dirancang untuk memberikan pengalaman pengguna yang interaktif dan mudah digunakan. Berikut merupakan tampilan website seperti pada Gambar 10 dan Gambar 11 dan Gambar 12.

Gambar 10 Tampilan Halaman Beranda Aplikasi

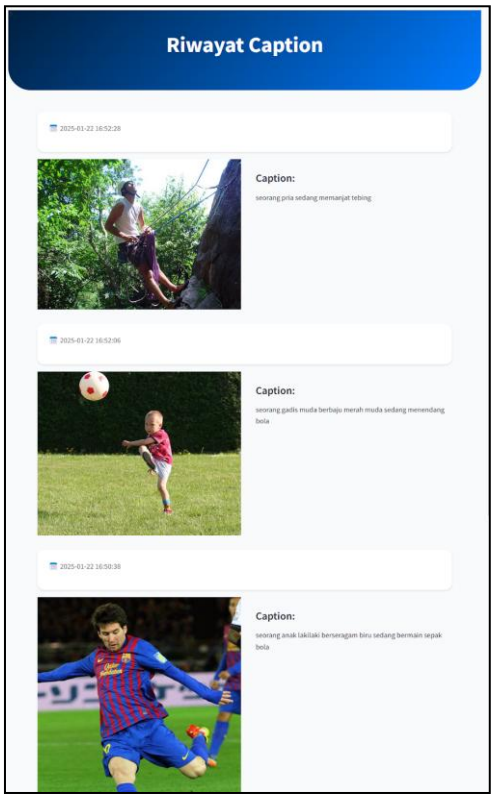
Gambar 11. Tampilan Halaman Upload Gambar

Pada Gambar 11 menampilkan bagian utama aplikasi dengan judul Generate Caption yang menjelaskan fungsi utama aplikasi. Pada halaman ini, pengguna dapat mengunggah gambar melalui fitur drag-and-drop atau tombol Browse Files yang disediakan dan jika sudah mengupload gambar dan mengklik tombol Generate Caption, Caption yang dihasilkan akan langsung ditampilkan di bawah gambar seperti pada Gambar 12.




Gambar 12. Tampilan Setelah Generate Caption

Setelah melakukan Generate Caption, pengguna dapat melihat riwayat captioning dari gambar-gambar yang sebelumnya telah diunggah melalui menu Riwayat Captioning yang dapat diakses melalui Menu Navigasi. Fitur ini dirancang untuk mempermudah pengguna dalam mengakses kembali hasil prediksi tanpa harus mengunggah ulang gambar yang sama, seperti ditampilkan pada Gambar 13 di bawah ini.




Gambar 13. Tampilan Menu Riwayat Captioning

## 4. KESIMPULAN

Penelitian ini berhasil mengimplementasikan model image captioning dengan arsitektur VGG16 sebagai encoder dan LSTM sebagai decoder untuk menghasilkan deskripsi otomatis dari gambar, dengan pelatihan dan evaluasi menggunakan dataset Flickr8k. Evaluasi performa model dalam menghasilkan deskripsi gambar berbahasa Indonesia menunjukkan hasil yang cukup baik dengan nilai BLEU-1 sebesar 0.6216, meskipun skor BLEU-4 yang lebih rendah (0.2030) mengindikasikan keterbatasan dalam kohesi kalimat secara keseluruhan, didukung oleh metrik METEOR (0.4617), CIDEr (0.5166), dan ROUGE-L (0.1958) yang mengkonfirmasi kemampuan dasar model yang memadai namun masih memerlukan pengembangan lebih lanjut untuk hasil yang lebih natural. Model

juga berhasil diimplementasikan dalam bentuk aplikasi web interaktif yang memungkinkan pengguna mengunggah gambar dan mendapatkan deskripsi otomatis, mendemonstrasikan potensi praktis dari sistem yang dikembangkan.

Untuk meningkatkan kualitas sistem, disarankan penelitian selanjutnya dapat mempertimbangkan penyempurnaan terjemahan pada dataset melalui preprocessing tambahan atau penggunaan model NLP untuk koreksi tata bahasa, mengingat masih terdapat beberapa caption yang kurang sesuai dengan kaidah bahasa Indonesia; selain VGG16, dapat dicoba arsitektur ekstraksi fitur lain seperti ResNet, EfficientNet, atau InceptionV3 untuk meningkatkan representasi visual sebelum proses captioning; serta implementasi berbagai transformasi atau parameter augmentasi data seperti perubahan pencahayaan, rotasi, skala, dan pemangkasan gambar guna meningkatkan keberagaman data latih dan mengurangi overfitting

## DAFTAR PUSTAKA


[1] M. D. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, 2019, doi: 10.1145/3295748.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3156–3164. doi: 10.1109/CVPR.2015.7298935.

[3] A. M. Nugroho and A. F. Hidayatullah, "Keterangan gambar otomatis berbahasa Indonesia menggunakan CNN dan LSTM," Sleman, Yogyakarta, Indonesia, 2018.

[4] S. Liu, L. Bai, Y. Hu, and H. Wang, "Image captioning based on deep neural networks," in *MATEC Web of Conferences*, 2018, p. 01052. doi: 10.1051/matecconf/201823201052.

[5] M. Stefanini, L. Baraldi, M. Cornia, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," 2021. doi: 10.48550/arXiv.2107.06912.

[6] Y. Azhar, M. R. Anugerah, M. A. R. Fahlopy, and A. Yusriansyah, "Image captioning using hybrid of VGG16 and bidirectional LSTM model," *KINETIK*, vol. 7, no. 4, pp. 391–398, 2019, doi: 10.22219/kinetik.v7i4.1568.

[7] K. Suzuki, "AI: A new open access journal for artificial intelligence," *AI*, vol. 1, no. 1, pp. 1–3, 2020.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the ACL*, 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

[9] S. Haykin, *Neural networks and learning machines*. Pearson, 2009.

[10] N. K. Manaswi, *Deep learning with applications using Python*. Apress, 2018. doi: 10.1007/978-1-4842-3516-4.

[11] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[12] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *ICET*, 2017. doi: 10.1109/ICET.2017.8308186.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[14] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," 2005, pp. 65–72.

[15] R. Santoso, "Sistem pendeskripsian gambar pemandangan sekitar bagi penyandang tunanetra berbahasa Indonesia," *J. Ilm. Inform. dan Sist. Inf.*, vol. 8, no. 1, pp. 45–54, 2024.

[16] M. Rifki, "Pengembangan CNN-LSTM-based image captioning dataset Indonesia untuk mendukung kemandirian penyandang tunanetra di ruang publik," *J. Teknol. Inf. dan Komun.*, vol. 9, no. 1, pp. 22–30, 2025.

[17] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *ACL Workshop*, doi: 10.3115/1073445.1073465.

[18] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of CVPR*, pp. 4566–4575. doi: 10.1109/CVPR.2015.7299087.

[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv. doi: 10.48550/arXiv.1409.1556.

[20] E. Mulyanto, E. I. Setiawan, E. M. Yuniarno, and M. H. Purnomo, "Automatic Indonesian image caption generation using CNN-LSTM model and FEEH-ID dataset," 2019. doi: 10.1109/CIVEMSA45640.2019.9071632.

[21] A. Apandi, A. B. Mutiara, and D. Dharmayanti, "Generating image captions in Indonesian using a deep learning approach based on vision transformer and IndoBERT," *J. Appl. Data Sci.*, vol. 6, no. 2, pp. 1191–1202, 2025, doi: 10.47738/jads.v6i2.672.

[22] I. Huda, "Implementasi natural language processing untuk aplikasi pencarian lokasi," *J. Nas. Teknol. Terap.*, 2024.

[23] U. A. Al Faruq and D. H. Fudholi, "Implementasi arsitektur transformer pada image captioning berbahasa

Indonesia," in *AUTOMATA*, 2023.
[24] D. Sudrajat, R. D. Permatasari, I. M. S. Wijaya, A. E. Setyawan, and N. Rahayu, "Pemanfaatan kecerdasan buatan sebagai upaya pengembangan media pembelajaran berbasis multimedia," *J. Kridatama Sains dan Teknol.*, vol. 5, no. 2, pp. 590–598, doi: 10.53863/kst.v5i02.999.
[25] B. Setiawan, "Kecerdasan buatan manusia: Artificial intelligence sebagai teknologi masa depan," *Ulul Albab*, vol. 9, no. 2, pp. 269–281, 2008.