

**PENERAPAN MODEL *CONVOLUTIONAL NEURAL NETWORK* DAN *LONG SHORT-TERM MEMORY* PADA
SISTEM DESKRIPSI CITRA OTOMATIS MENGGUNAKAN
DATASET FLICKR8K**

TUGAS AKHIR

Diajukan oleh:

Muhammad Bilal Akbar

NIM. 220705070

Mahasiswa Fakultas Sains dan Teknologi
Program Studi Teknologi Informasi



**FAKULTAS SAINS DAN TEKNOLOGI
UNIVERSITAS ISLAM NEGERI AR-RANIRY
BANDA ACEH
2026 M/1448 H**

LEMBAR PERSETUJUAN

PENERAPAN *MODEL CONVOLUTIONAL NEURAL NETWORK* DAN *LONG SHORT-TERM MEMORY* PADA SISTEM DESKRIPSI CITRA OTOMATIS MENGGUNAKAN DATASET FLICKR8K

TUGAS AKHIR

Diajukan Kepada Fakultas Sains dan Teknologi
Universitas Islam Negeri (UIN) Ar-Raniry Banda Aceh
Sebagai Salah Satu Beban Studi Memperoleh Gelar Sarjana (S1)
dalam Prodi Teknologi Informasi

Oleh:

Muhammad Bilal Akbar

NIM. 220705070

Mahasiswa Fakultas Sains dan Teknologi
Program Studi Teknologi Informasi

Disetujui untuk dimunaqasyahkan Oleh:

Pembimbing I,

Ghufran Ibnu Yasa, M.T.

NIP. 198409262014031005

Pembimbing II,

Raihan Islamadina, M.T.

NIP. 198901312020122011

Mengetahui,

Ketua Program Studi Teknologi Informasi

Malahayati, M.T.

NIP. 198301272015032003

LEMBAR PENGESAHAN

PENERAPAN MODEL CONVOLUTIONAL NEURAL NETWORK DAN LONG SHORT-TERM MEMORY PADA SISTEM DESKRIPSI CITRA OTOMATIS MENGGUNAKAN DATASET FLICKR8K

TUGAS AKHIR

Telah Diuji oleh Panitia Ujian Munaqasyah Tugas Akhir
Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh dan Dinyatakan Lulus
Serta Diterima Sebagai Salah Satu Beban Studi Program Sarjana (S1)
Dalam Program Studi Teknologi Informasi

Pada Hari/Tanggal: Rabu, 12 Agustus 2026 M
28 Shafar 1448 H

Di Darussalam, Banda Aceh

Panitia Ujian Munaqasyah Tugas Akhir:

Ketua,


Ghufran Ibnu Nasa, M.T.
NIP. 198409262014031005

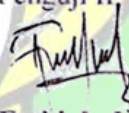
Sekretaris,


Raihan Islamadina, M.T.
NIP. 198901312020122011

Penguji I,


Dr. Hendri Ahmadian, M.L.M.
NIP. 198301042014031002

Penguji II,


Fathiah, M.Eng.
NIP. 198606152019032010

Mengetahui



Dekan Fakultas Sains dan Teknologi
UIN Ar-Raniry Banda Aceh


Prof. Dr. J. Muhammad Dirhamsyah, M.T., IPU
NIP. 196210021988111001

LEMBAR PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan di bawah ini:

Nama : Muhammad Bilal Akbar
NIM : 220705070
Program Studi : Teknologi Informasi
Fakultas : Sains dan Teknologi
Judul : Penerapan model *convolutional neural network* dan *long short-term memory* pada sistem deskripsi citra otomatis menggunakan dataset Flickr8k

Dengan ini menyatakan bahwa dalam penulisan tugas akhir ini, saya:

1. Tidak menggunakan ide orang lain tanpa mampu mengembangkan dan mempertanggung jawabkan.
2. Tidak melakukan plagiasi terhadap naskah karya orang lain.
3. Tidak menggunakan karya orang lain tanpa menyebutkan sumber asli atau tanpa izin pemilik karya.
4. Tidak memanipulasi dan memalsukan data.
5. Mengerjakan sendiri karya ini dan mampu bertanggung jawab atas karya ini.

Bila di kemudian hari ada tuntutan dari pihak lain atas karya saya, dan telah melalui pembuktian yang dapat dipertanggungjawabkan serta ditemukan bukti bahwa saya telah melanggar pernyataan ini, maka saya siap dikenai sanksi berdasarkan aturan yang berlaku di Fakultas Sains dan Teknologi UIN Ar-Raniry Banda Aceh.

Demikian pernyataan ini saya buat dengan sesungguhnya dan tanpa paksaan dari pihak manapun.

Banda Aceh, 28 April 2026

Yang Menyatakan,



Muhammad Bilal Akbar

ABSTRAK

Nama: Muhammad Bilal Akbar

NIM: 220705070

Program Studi: Teknologi Informasi

Judul: Penerapan Model *Convolutional Neural Network* dan *Long Short-Term Memory* pada Sistem Deskripsi Citra Otomatis Menggunakan Dataset Flickr8k

Tanggal Sidang: 12 Agustus 2026

Jumlah Halaman: 70 Halaman

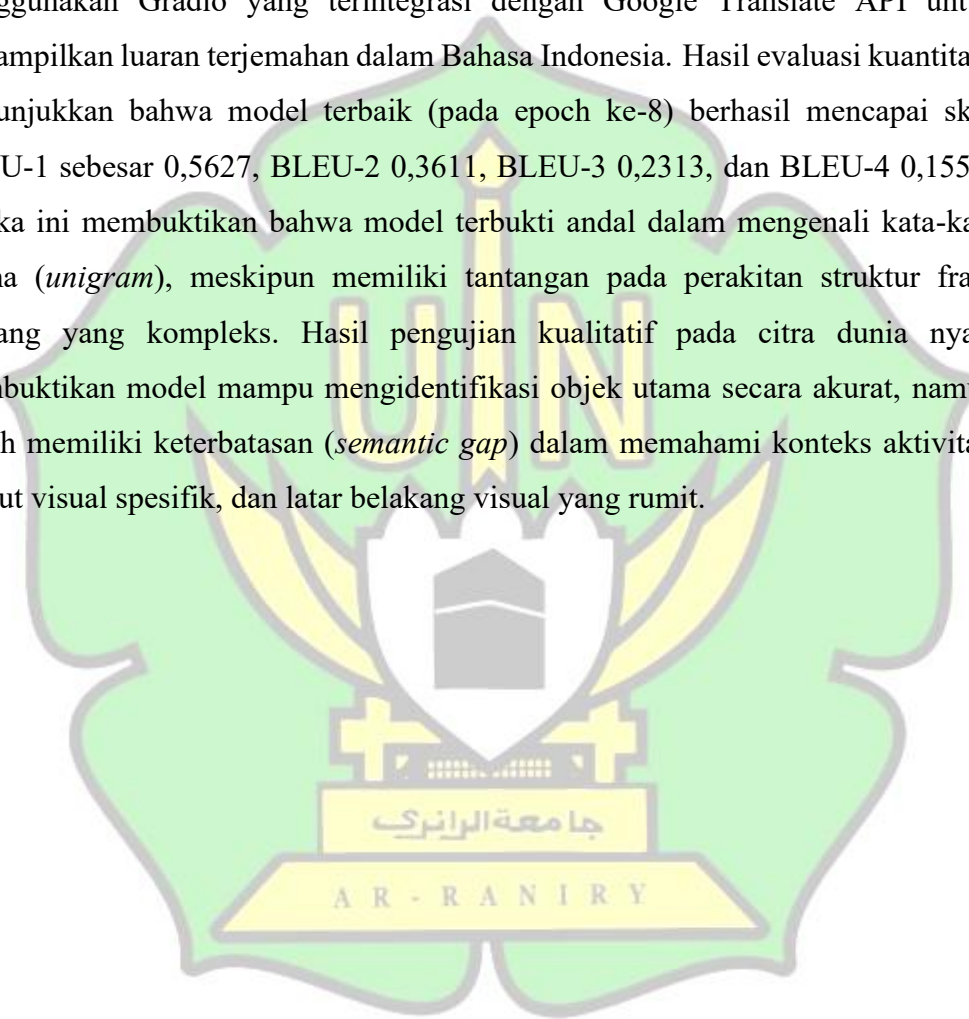
Pembimbing I: Ghufraan Ibnu Yasa, M.T.

Pembimbing II: Raihan Islamadina, M.T.

Kata Kunci: *Image Captioning*, *Convolutional Neural Network*, *Long Short-Term Memory*, VGG16, Dataset Flickr8k, *BLEU Score*.

Pertumbuhan masif data citra digital di internet menimbulkan tantangan dalam pengelolaan dan aksesibilitas konten visual, khususnya bagi penyandang disabilitas netra yang sangat bergantung pada teknologi pembaca layar. Pelabelan gambar secara manual menjadi proses yang tidak efisien dan memakan waktu. Oleh karena itu, diperlukan sistem *Image Captioning* otomatis yang mampu mengintegrasikan kemampuan *Computer Vision* dan *Natural Language Processing* (NLP) untuk menerjemahkan informasi visual menjadi representasi bahasa alami manusia. Penelitian ini bertujuan untuk merancang dan mengimplementasikan arsitektur *Encoder-Decoder* guna membangun sistem deskripsi citra otomatis berbahasa Inggris menggunakan himpunan data publik Flickr8k. Sistem ini memanfaatkan arsitektur *Convolutional Neural Network* (CNN) model VGG16 yang telah dilatih sebelumnya pada ImageNet sebagai *encoder* untuk mengekstraksi fitur visual. Sementara itu, *Long Short-Term Memory* (LSTM) digunakan sebagai *decoder* untuk memodelkan struktur kalimat secara sekuensial. Proses pembangkitan kalimat dioptimalkan menggunakan metode *Beam Search* dengan beam width 5 dan alpha 0,7 untuk menghasilkan kandidat deskripsi yang paling koheren secara

struktural. Pelatihan model dilakukan menggunakan fungsi kerugian *Sparse Categorical Crossentropy* dan algoritma optimasi Adam. Kinerja sistem dievaluasi melalui dua pendekatan, yaitu pengujian kuantitatif menggunakan metrik *Bilingual Evaluation Understudy* (BLEU) Score pada data uji standar, serta pengujian kualitatif untuk menilai akurasi pengenalan objek, atribut, dan aktivitas pada citra dunia nyata di luar dataset latih. Untuk mendukung antarmuka pengguna, sistem juga diimplementasikan ke dalam bentuk aplikasi prototipe berbasis web menggunakan Gradio yang terintegrasi dengan Google Translate API untuk menampilkan luaran terjemahan dalam Bahasa Indonesia. Hasil evaluasi kuantitatif menunjukkan bahwa model terbaik (pada epoch ke-8) berhasil mencapai skor BLEU-1 sebesar 0,5627, BLEU-2 0,3611, BLEU-3 0,2313, dan BLEU-4 0,1556. Angka ini membuktikan bahwa model terbukti andal dalam mengenali kata-kata utama (*unigram*), meskipun memiliki tantangan pada perakitan struktur frasa panjang yang kompleks. Hasil pengujian kualitatif pada citra dunia nyata membuktikan model mampu mengidentifikasi objek utama secara akurat, namun masih memiliki keterbatasan (*semantic gap*) dalam memahami konteks aktivitas, atribut visual spesifik, dan latar belakang visual yang rumit.



ABSTRACT

Name: Muhammad Bilal Akbar

Student Number: 220705070

Department: Information Technology

Title: Application of *Convolutional Neural Network* and *Long Short-Term Memory* Models for Automatic *Image Captioning* Systems Using the Flickr8k Dataset

Date: 12 August 2026

Number of Pages: 70 Pages

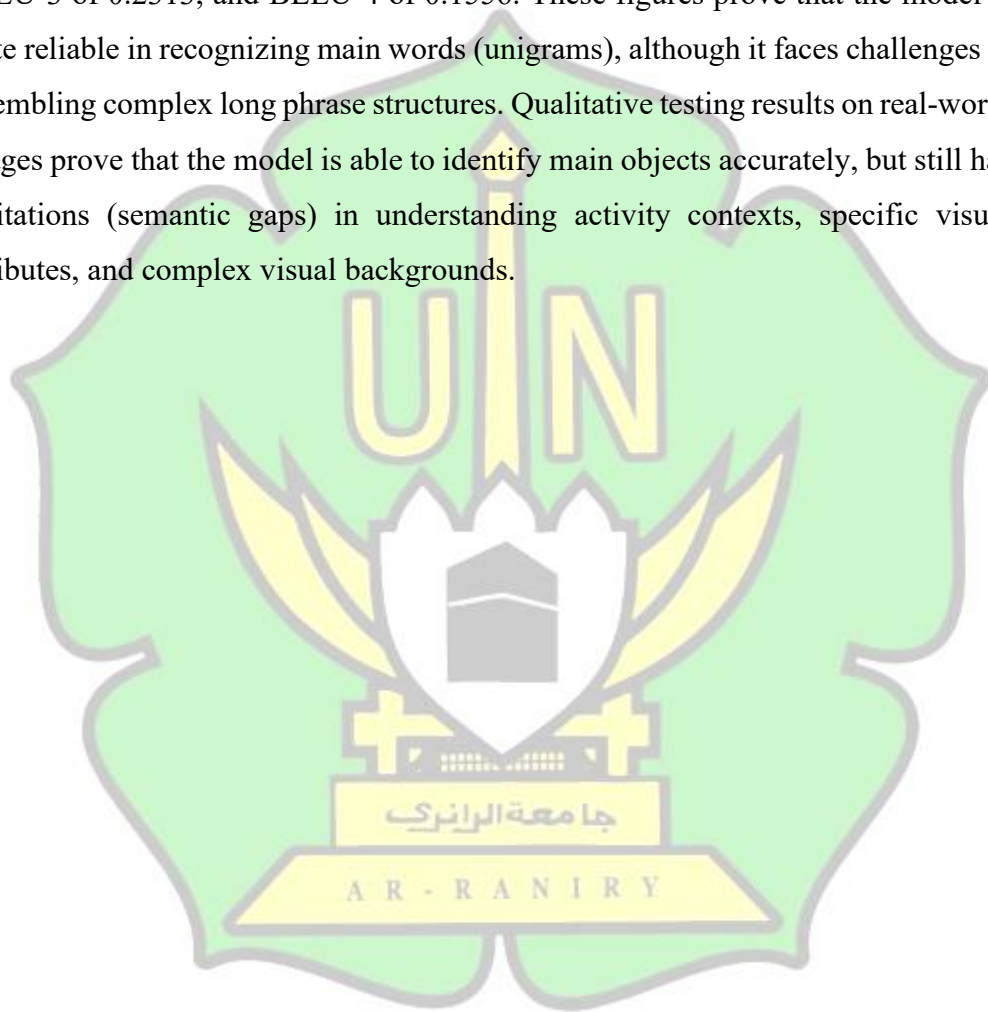
Supervisor I: Ghufran Ibnu Yasa, M.T.

Supervisor II: Raihan Islamadina, M.T.

Keywords: *Image Captioning*, *Convolutional Neural Network*, *Long Short-Term Memory*, VGG16, Flickr8k Dataset, *BLEU Score*.

The massive growth of digital image data on the internet poses challenges in managing and accessing visual content, especially for visually impaired individuals who heavily rely on screen reader technology. Manual image labeling is an inefficient and time-consuming process. Therefore, an automatic *Image Captioning* system capable of integrating *Computer Vision* and *Natural Language Processing* (NLP) capabilities is required to translate visual information into human natural language representation. This research aims to design and implement an Encoder-Decoder architecture to build an automatic English image captioning system using the Flickr8k public dataset. This system utilizes a *Convolutional Neural Network* (CNN) architecture, specifically the VGG16 model *pre-trained* on *ImageNet*, as an encoder to extract visual features. Meanwhile, *Long Short-Term Memory* (LSTM) is used as a decoder to model sentence structure sequentially. The sentence generation process is optimized using the *Beam Search* method with a beam width of 5 and an alpha of 0.7 to generate the most structurally coherent description candidates. Model training was conducted using the *Sparse Categorical Crossentropy* loss function and the Adam optimization algorithm. System performance was evaluated through two approaches: quantitative testing using the

Bilingual Evaluation Understudy (BLEU) Score metric on standard test data, and qualitative testing to assess the accuracy of object, attribute, and activity recognition on real-world images outside the training dataset. To support the user interface, the system was also implemented into a web-based prototype application using Gradio, integrated with the Google Translate API to display the translated output in Indonesian. The quantitative evaluation results show that the best model (at the 8th epoch) successfully achieved a BLEU-1 score of 0.5627, BLEU-2 of 0.3611, BLEU-3 of 0.2313, and BLEU-4 of 0.1556. These figures prove that the model is quite reliable in recognizing main words (unigrams), although it faces challenges in assembling complex long phrase structures. Qualitative testing results on real-world images prove that the model is able to identify main objects accurately, but still has limitations (semantic gaps) in understanding activity contexts, specific visual attributes, and complex visual backgrounds.



KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Allah SWT atas segala rahmat, hidayah, dan karunia-Nya sehingga penulis dapat menyelesaikan Skripsi yang berjudul “Penerapan Model *Convolutional Neural Network* dan *Long Short-Term Memory* pada Sistem Deskripsi Citra Otomatis Menggunakan Dataset Flickr8k” dengan baik dan tepat waktu.

Skripsi ini disusun sebagai salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri Ar-Raniry Banda Aceh. Penelitian ini bertujuan untuk mengimplementasikan model berbasis *Deep Learning* dalam menghasilkan deskripsi citra secara otomatis, serta mengevaluasi performa model menggunakan pendekatan kuantitatif dan kualitatif.

Dalam proses penyusunan Skripsi ini, penulis menghadapi berbagai kendala dan tantangan, baik dalam proses pengumpulan data, implementasi model, maupun analisis hasil penelitian. Keterbatasan pengetahuan, pengalaman, serta kompleksitas permasalahan yang diangkat menjadi bagian dari proses pembelajaran yang sangat berharga bagi penulis. Namun demikian, berkat bimbingan, arahan, serta dukungan dari berbagai pihak, seluruh tahapan penelitian dapat diselesaikan dengan baik.

Pada kesempatan ini, penulis dengan penuh rasa hormat dan terima kasih yang sebesar-besarnya menyampaikan apresiasi kepada:

1. Bapak Ghufran Ibnu Yasa, M.T. selaku Pembimbing I yang telah memberikan bimbingan, arahan, serta masukan yang konstruktif selama proses penyusunan Skripsi ini.
2. Ibu Raihan Islamadina, M.T. selaku Pembimbing II yang telah memberikan saran, kritik, serta motivasi dalam penyempurnaan penelitian ini.
3. Ibu Malahayati, M.T. selaku Ketua Program Studi Teknologi Informasi yang telah memberikan dukungan dalam proses akademik penulis.

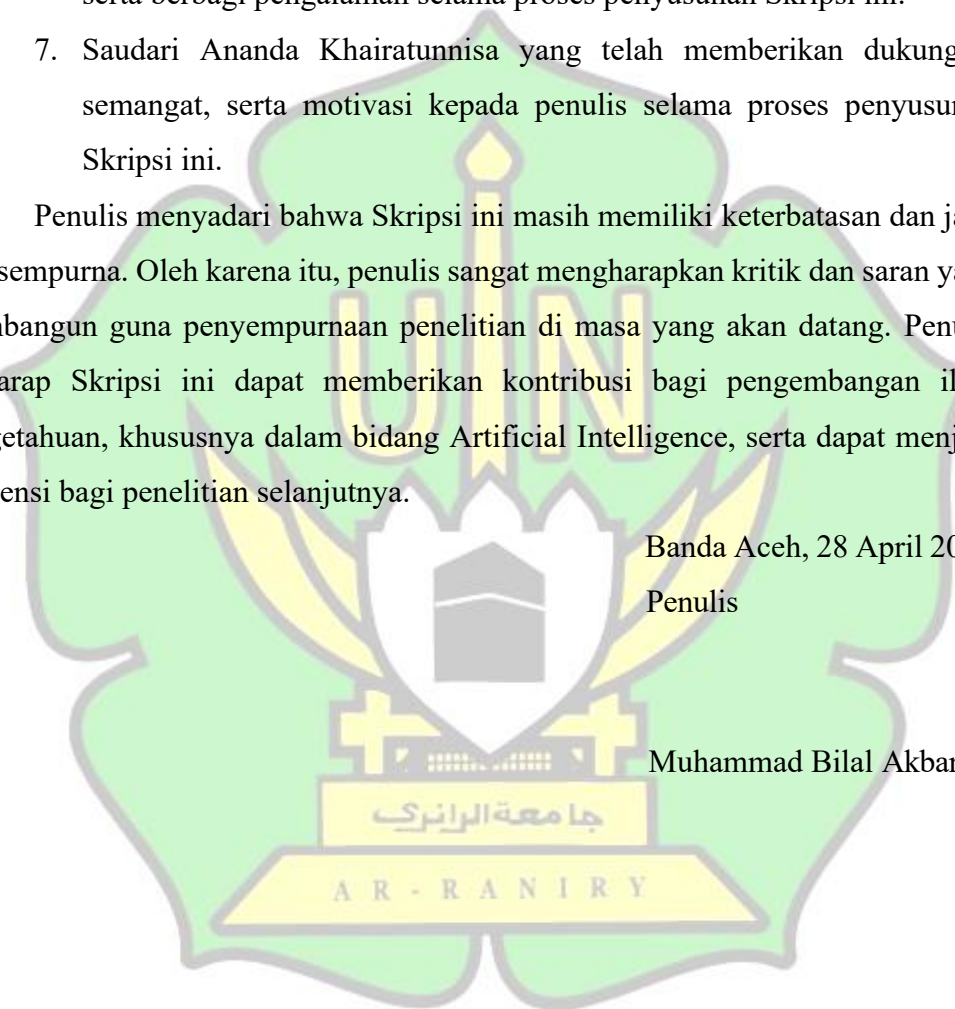
4. Ibu Cut Ida Rahmadiana, S.Si. selaku staf administrasi yang telah banyak membantu penulis dalam berbagai urusan administrasi selama proses penyusunan Tugas Akhir.
5. Kedua orang tua penulis yang senantiasa memberikan doa, dukungan, serta motivasi yang tidak ternilai, baik secara moral maupun material, sehingga penulis dapat menyelesaikan studi ini dengan baik.
6. Rekan-rekan seperjuangan yang telah memberikan semangat, dukungan, serta berbagi pengalaman selama proses penyusunan Skripsi ini.
7. Saudari Ananda Khairatunnisa yang telah memberikan dukungan, semangat, serta motivasi kepada penulis selama proses penyusunan Skripsi ini.

Penulis menyadari bahwa Skripsi ini masih memiliki keterbatasan dan jauh dari sempurna. Oleh karena itu, penulis sangat mengharapkan kritik dan saran yang membangun guna penyempurnaan penelitian di masa yang akan datang. Penulis berharap Skripsi ini dapat memberikan kontribusi bagi pengembangan ilmu pengetahuan, khususnya dalam bidang Artificial Intelligence, serta dapat menjadi referensi bagi penelitian selanjutnya.

Banda Aceh, 28 April 2026

Penulis

Muhammad Bilal Akbar

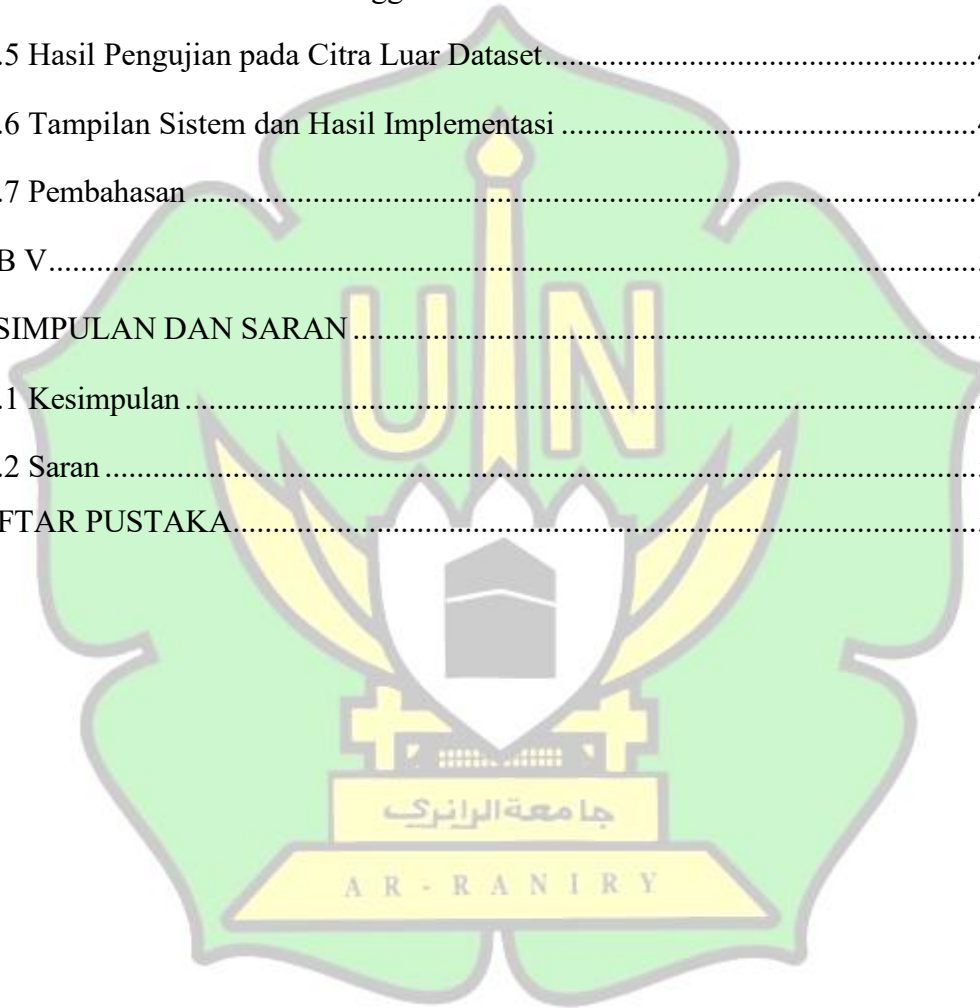


DAFTAR ISI

LEMBAR PERSETUJUAN	i
LEMBAR PENGESAHAN	ii
LEMBAR PERNYATAAN KEASLIAN TUGAS AKHIR	iii
ABSTRAK	iv
ABSTRACT	vi
KATA PENGANTAR.....	viii
DAFTAR ISI	x
DAFTAR TABEL	xiii
DAFTAR GAMBAR	xiv
BAB I.....	1
PENDAHULUAN.....	1
1.1 Latar Belakang.....	1
1.2 Rumusan Masalah.....	2
1.3 Tujuan Penelitian	2
1.4 Batasan Masalah	3
1.5 Manfaat Penelitian	4
BAB II.....	6
TINJAUAN PUSTAKA.....	6
2.1 Penelitian Terdahulu.....	6
2.2 Landasan Teori	7
2.2.1 <i>Image Captioning</i>	7
2.2.2 <i>Convolutional Neural Network (CNN)</i>	9
2.2.3 <i>Long Short-Term Memory (LSTM)</i>	10
2.2.4 Dataset Flickr8k	13

2.2.5 BLEU Score (<i>Bilingual Evaluation Understudy</i>).....	14
2.2.6 <i>Beam Search</i>	17
2.2.7 Pengolahan Data Awal (Preprocessing)	17
2.2.8 Optimasi Model (<i>Model Optimization</i>)	19
2.2.9 Lingkungan Pengembangan (<i>Development Environment</i>).....	20
BAB III.....	21
METODE PENELITIAN	21
3.1 Alur Penelitian	21
3.1.1 Diagram Alir Penelitian (<i>Flowchart</i>)	22
3.2 Metode Pengumpulan Data.....	25
3.3 Tahapan Pengolahan Data Awal (Preprocessing).....	27
3.3.1 Pengolahan Citra (Image Preprocessing)	27
3.3.2 Pengolahan <i>Caption</i> (Text Preprocessing)	28
3.4 Perancangan Arsitektur Model	29
3.4.1 Encoder (CNN VGG16).....	29
3.4.2 <i>Decoder</i> (LSTM)	29
3.4.3 Fitur Penerjemahan (Translation).....	31
3.5 Skenario Pengujian	31
3.5.1 Pengujian Kuantitatif (Quantitative Testing).....	32
3.5.2 Pengujian Kualitatif (Qualitative Testing)	33
3.6 Lingkungan Pengembangan.....	35
3.6.1 Perangkat Keras (<i>Hardware</i>)	35
3.6.2 Perangkat Lunak (<i>Software</i>).....	35
3.7 Waktu dan Tempat Penelitian.....	36
3.7.1 Tempat Penelitian.....	36
3.7.2 Waktu Penelitian	37

BAB IV	38
HASIL DAN PEMBAHASAN	38
4.1 Gambaran Umum Hasil Penelitian	38
4.2 Hasil Pengolahan Dataset	39
4.3 Hasil Pelatihan Model Berdasarkan Epoch.....	41
4.4 Hasil Evaluasi Model Menggunakan <i>BLEU Score</i>	43
4.5 Hasil Pengujian pada Citra Luar Dataset.....	44
4.6 Tampilan Sistem dan Hasil Implementasi	47
4.7 Pembahasan	49
BAB V.....	52
KESIMPULAN DAN SARAN.....	52
5.1 Kesimpulan	52
5.2 Saran	53
DAFTAR PUSTAKA.....	54



DAFTAR TABEL

Tabel 2.1 Ilustrasi Konsep Pencocokan N-Gram pada <i>BLEU Score</i>	16
Tabel 3.1 Perangkat Keras.....	35
Tabel 4.1 Pembagian Dataset Flickr8k.....	39
Tabel 4.2 Hasil Pemeriksaan Ukuran Awal Citra.....	40
Tabel 4.3 Hasil Konfigurasi Pelatihan Model CNN VGG16 dan LSTM.....	41
Tabel 4.4 Hasil Evaluasi <i>BLEU Score</i> Model Akhir.....	43
Tabel 4.5 Hasil Pengujian pada Citra Luar Dataset	45



DAFTAR GAMBAR

Gambar 2.1 Arsitektur Generik Sistem <i>Image Captioning</i> Berbasis CNN VGG16 dan LSTM.....	8
Gambar 2.2 Arsitektur <i>Convolutional Neural Network</i> (CNN)	9
Gambar 2.3 Struktur Sel <i>Long Short-Term Memory</i> (LSTM).....	11
Gambar 3.1 Alur Penelitian.....	22
Gambar 3.2 Sampel Citra dan Anotasi Deskripsi pada Dataset Flickr8k... ..	26
Gambar 4.2 Tampilan Sistem <i>Image Captioning</i> CNN VGG16 dan LSTM.....	47



BAB I

PENDAHULUAN

1.1 Latar Belakang

Pesatnya perkembangan *Artificial Intelligence* (AI) melalui metode *Deep Learning* telah membawa perubahan signifikan terhadap sistem komputasi modern dalam menganalisis pola data kompleks seperti teks dan citra (Sarker, 2021a). Tantangan utama dalam bidang ini adalah mengintegrasikan kemampuan *Computer Vision* dan *Natural Language Processing* (NLP) ke dalam satu sistem *multimodal* yang mampu mengaitkan informasi visual dengan representasi bahasa (Stefanini et al., 2021a). Integrasi ini memungkinkan mesin tidak hanya mengenali objek, tetapi juga menghasilkan deskripsi tekstual bermakna yang mendekati cara manusia menafsirkan lingkungannya (Stefanini et al., 2021a).

Pertumbuhan masif data citra digital di internet saat ini menimbulkan tantangan dalam pengelolaan dan aksesibilitas konten visual (Jean et al., 2025). Proses pelabelan gambar secara manual menjadi tidak efisien, terutama bagi penyandang disabilitas netra yang sangat bergantung pada teknologi pembaca layar untuk mengakses informasi digital (Jean et al., 2025). Oleh karena itu, pengembangan sistem *Image Captioning* otomatis menjadi solusi krusial untuk meningkatkan aksesibilitas dan menjembatani kesenjangan informasi bagi kelompok tersebut (Jean et al., 2025).

Berbeda dengan klasifikasi citra konvensional, *Image Captioning* menuntut model untuk memahami keterkaitan antarobjek, konteks spasial, dan struktur gramatikal kalimat secara utuh (Stefanini et al., 2022). Ketidakakuratan dalam mengenali detail visual tertentu dapat berdampak signifikan terhadap makna kalimat yang dihasilkan (Stefanini et al., 2022). Hal ini menuntut model memiliki kemampuan pemetaan fitur visual ke dalam bahasa alami yang presisi (Stefanini et al., 2022).

Penelitian ini menerapkan arsitektur *Encoder-Decoder* berbasis *Convolutional Neural Network* (CNN) dan *Long Short-Term Memory* (LSTM) (Kaur & Kaur, 2025). CNN VGG16 digunakan sebagai *encoder* untuk mengekstraksi fitur spasial citra secara optimal, sementara LSTM bertindak sebagai

decoder untuk memodelkan struktur kalimat yang koheren (Kaur & Kaur, 2025). Model ini dilatih menggunakan dataset Flickr8k yang menyediakan variasi deskripsi representatif dengan beban komputasi yang efisien (Kaur & Kaur, 2025). Selain pengujian pada dataset baku, penelitian ini juga mengkaji keandalan model pada citra dunia nyata (jurnalistik) untuk mengevaluasi kesenjangan semantik antara deskripsi otomatis dan deskripsi alami buatan manusia.

Dalam mengukur kinerja sistem, penelitian ini menggunakan metrik *Bilingual Evaluation Understudy* (BLEU) Score sebagai standar evaluasi utama. Merujuk pada tinjauan literatur terdahulu terhadap arsitektur dasar serupa, performa skor BLEU-1 pada dataset Flickr8k umumnya berada pada rentang 0.50 hingga 0.60 (Stefanini et al., 2022). Angka-angka tersebut ditetapkan sebagai tolok ukur (benchmark) keberhasilan pengujian kuantitatif dalam penelitian ini, mengingat arsitektur CNN VGG16 dan LSTM dasar memiliki tantangan tersendiri dalam menyusun deskripsi citra yang sangat kompleks.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana merancang dan mengimplementasikan arsitektur *Encoder-Decoder* berbasis CNN VGG16 dan LSTM untuk menghasilkan deskripsi citra otomatis berbahasa Inggris pada dataset Flickr8k?
2. Bagaimana kinerja model CNN VGG16 dan LSTM dalam menghasilkan deskripsi citra pada dataset Flickr8k berdasarkan *metrik evaluasi Bilingual Evaluation Understudy* (BLEU Score)?
3. Bagaimana kinerja dan keandalan model dalam menghasilkan deskripsi pada citra dunia nyata (di luar dataset latih), ditinjau dari aspek kesesuaian pengenalan objek visual, kelancaran tata bahasa (grammatical fluency), dan relevansi aktivitas?

1.3 Tujuan Penelitian

Adapun tujuan yang ingin dicapai dari penelitian ini adalah sebagai berikut:

1. Mengembangkan dan mengimplementasikan arsitektur *Encoder-Decoder* berbasis CNN VGG16 dan LSTM untuk menghasilkan deskripsi citra otomatis berbahasa Inggris pada dataset Flickr8k.
2. Mengevaluasi kinerja model CNN VGG16 dan LSTM dalam menghasilkan deskripsi citra pada dataset Flickr8k menggunakan metrik evaluasi *Bilingual Evaluation Understudy* (BLEU Score).
3. Menganalisis serta mengevaluasi kualitas kalimat deskripsi yang dibangkitkan oleh sistem pada saat dihadapkan dengan citra dunia nyata, dengan menitikberatkan pengujian pada relevansi identifikasi objek dan keutuhan struktur tata bahasa buatan mesin.

1.4 Batasan Masalah

Agar penelitian ini dapat dilakukan secara terarah dan sesuai dengan tujuan yang telah ditetapkan, maka batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data latih, data validasi, dan data uji yang digunakan untuk membangun model bersumber dari himpunan data publik Flickr8k, yang terdiri dari 8.000 citra beserta lima variasi kalimat deskripsi referensi berbahasa Inggris per citranya.
2. Arsitektur *Deep learning* yang bertugas sebagai pengekstraksi fitur citra atau *encoder* dibatasi menggunakan model VGG16 yang telah dilatih sebelumnya pada dataset *ImageNet*. Pemilihan VGG16 didasarkan pada kestabilannya sebagai model *feature extractor* serta kemampuannya dalam mengekstraksi fitur visual tingkat tinggi dari citra. Pada penelitian ini, lapisan klasifikasi akhir VGG16 tidak digunakan karena fokus sistem bukan pada klasifikasi citra, melainkan pada pemanfaatan fitur visual sebagai masukan bagi *decoder* LSTM untuk membangkitkan deskripsi citra.
3. Pengujian performa sistem dilakukan melalui dua pendekatan utama. Pengujian kuantitatif diukur menggunakan metrik evaluasi Bilingual Evaluation Understudy (BLEU-1 sampai BLEU-4). Sedangkan pengujian kualitatif difokuskan pada analisis deskriptif standar Machine Learning, yaitu mengevaluasi tingkat akurasi pengenalan objek visual (object

recognition) dan kesesuaian relasi atribut pada citra yang tidak terdistribusi di dalam dataset latih.

4. Proses pembangkitan *caption* pada tahap pengujian menggunakan metode *Beam Search* untuk memilih kandidat urutan kata terbaik. Penelitian ini tidak membandingkan performa *Beam Search* dengan metode decoding lain, seperti *Greedy Search* atau sampling.
5. Terjemahan kalimat deskripsi citra dari bahasa Inggris ke bahasa Indonesia sepenuhnya bergantung pada kualitas *Application Programming Interface* (API) eksternal (*Google Translator* melalui *library Deep Translator*), sehingga struktur linguistik hasil terjemahan bukan merupakan fokus utama optimasi algoritma internal pada penelitian ini.

1.5 Manfaat Penelitian

Adapun manfaat yang diharapkan dari pelaksanaan penelitian ini adalah sebagai berikut:

1. Manfaat Teoretis (Akademis)

a. Kontribusi terhadap Ilmu Pengetahuan

Penelitian ini diharapkan dapat menambah wawasan dan pemahaman dalam bidang *Deep Learning*, khususnya terkait penerapan integrasi metode *Convolutional Neural Network* (CNN) dan *Long Short-Term Memory* (LSTM) untuk menyelesaikan permasalahan yang melibatkan *Computer Vision* dan *Natural Language Processing* (NLP).

b. Referensi Ilmiah

Hasil penelitian ini dapat dijadikan sebagai bahan referensi atau rujukan bagi peneliti selanjutnya, terutama mahasiswa Strata-1 (S1), yang tertarik untuk mengembangkan sistem *image captioning* atau melakukan penelitian lanjutan pada bidang pengolahan citra dan teks.

2. Manfaat Praktis (Masyarakat/Pengguna)

a. Peningkatan Aksesibilitas Informasi

Memberikan kontribusi dalam pengembangan teknologi aksesibilitas dengan menyediakan deskripsi tekstual (*image caption*) yang dapat dimanfaatkan oleh teknologi pembaca layar (*screen reader*), sehingga membantu penyandang tunanetra dalam memahami konten gambar pada media digital.

b. Efisiensi Penyampaian Informasi

Sistem yang dikembangkan dapat membantu pengguna umum maupun pengelola konten digital dalam memahami konteks utama dari sebuah citra secara lebih cepat, tanpa memerlukan proses analisis visual manual yang mendalam.

3. Manfaat Teknis (Pengembangan Sistem)

a. Bukti Empiris Kinerja Model

Penelitian ini memberikan bukti empiris mengenai kinerja arsitektur *Encoder-Decoder* CNN VGG16 dan LSTM dalam menghasilkan deskripsi citra yang bersifat gramatikal dan relevan berdasarkan pengujian pada dataset Flickr8k.

b. Evaluasi pada Citra Dunia Nyata

Selain pengujian kuantitatif, penelitian ini juga membandingkan kualitas *image captioning* otomatis dengan deskripsi natural buatan manusia profesional. Evaluasi pada citra dunia nyata ini secara khusus dilakukan guna mengidentifikasi keterbatasan kontekstual sistem saat ini, serta merumuskan peluang peningkatan kinerja model untuk riset-riset selanjutnya.

BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian mengenai *image captioning* telah banyak dilakukan dengan memanfaatkan berbagai arsitektur *Deep Learning*. Kajian terhadap penelitian terdahulu diperlukan untuk memetakan posisi penelitian ini dalam perkembangan keilmuan (state of the art) serta untuk menunjukkan perbedaan dan kontribusi yang diusulkan.

Stefanini et al. (2022) melakukan studi literatur berjudul *From Show to Tell: A Survey on Deep Learning-based Image Captioning* yang mengkaji perkembangan metode *image captioning* berbasis *Deep Learning* setelah tahun 2020. Penelitian ini membahas evolusi arsitektur *encoder-decoder* modern, penggunaan mekanisme atensi, serta pemanfaatan dataset dan metrik evaluasi terbaru. Meskipun memberikan gambaran komprehensif mengenai tren penelitian, studi ini bersifat non-implementatif. Berbeda dengan penelitian tersebut, penelitian ini berfokus pada implementasi teknis dengan membangun sistem *image captioning* secara langsung menggunakan dataset standar Flickr8k.

Sharma dan Kumar (2021) mengusulkan sistem *image captioning* berbasis CNN VGG16 dan LSTM dengan memanfaatkan VGG16 sebagai pengekstraksi fitur. Hasil penelitian menunjukkan bahwa VGG16 memiliki stabilitas yang lebih baik dalam proses ekstraksi fitur dibandingkan model Inception, khususnya pada penggunaan dataset berukuran relatif kecil. Perbedaan dengan penelitian ini terletak pada ruang lingkup evaluasi, di mana penelitian ini tidak hanya mengandalkan skor *BLEU*, tetapi juga menambahkan analisis kualitatif terhadap citra dunia nyata (*real-world images*) yang bersumber dari media massa.

Penelitian oleh Nugraha dan Santoso (2023) berfokus pada sistem *image captioning* berbahasa Indonesia dengan melatih model CNN VGG16 dan LSTM secara langsung menggunakan dataset lokal. Pendekatan ini memungkinkan sistem menghasilkan deskripsi dalam Bahasa Indonesia tanpa proses penerjemahan tambahan. Berbeda dengan penelitian tersebut, penelitian ini menggunakan pendekatan *post-processing* berupa penerjemahan otomatis untuk memanfaatkan

dataset standar global, yaitu Flickr8k, yang memiliki kualitas anotasi dan validitas evaluasi yang lebih teruji secara internasional.

Yuan et al. (2022) mengembangkan metode *image captioning* dengan mengombinasikan CNN dan LSTM serta teknik *Reinforcement Learning* untuk meningkatkan variasi dan kualitas kata pada deskripsi yang dihasilkan. Pendekatan ini menunjukkan peningkatan performa dibandingkan metode *supervised learning* murni. Namun, penelitian ini membatasi kompleksitas model hanya pada pendekatan CNN VGG16 dan LSTM berbasis *supervised learning* sebagai model dasar (*baseline*), dengan tujuan mengevaluasi performa fundamental sistem sebelum dilakukan optimasi lanjutan.

Bhat et al. (2023) mengusulkan sistem teknologi asistif bagi penyandang disabilitas netra dengan mengintegrasikan *image captioning* berbasis CNN VGG16 dan LSTM dan fitur *Text-to-Speech*. Penelitian tersebut menitikberatkan pada integrasi keluaran suara untuk membantu pengguna dalam memahami konten visual. Perbedaan dengan penelitian ini terletak pada fokus evaluasi, di mana penelitian ini lebih menekankan pada analisis kesesuaian makna dan kualitas semantik deskripsi yang dihasilkan mesin dibandingkan dengan deskripsi manusia (wartawan/profesional), bukan pada pengembangan perangkat keras atau antarmuka suara.

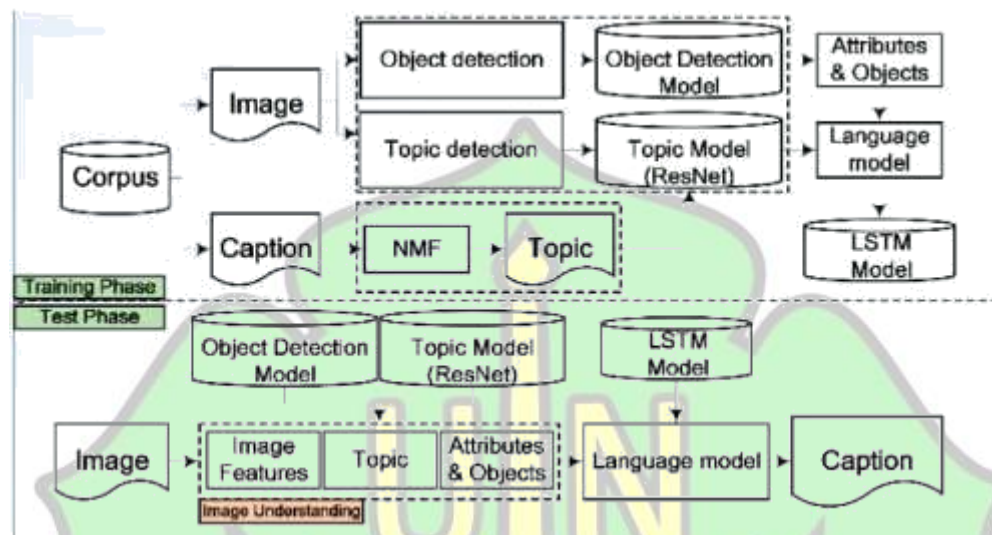
2.2 Landasan Teori

2.2.1 Image Captioning

Image Captioning didefinisikan sebagai proses otomatis dalam menghasilkan deskripsi tekstual berdasarkan informasi visual yang terkandung di dalam sebuah citra digital. Berdasarkan survei komprehensif yang dilakukan oleh (Stefanini et al., 2021b), *image captioning* menempati posisi penting dalam penelitian kecerdasan buatan karena mengharuskan model memiliki kemampuan pemahaman visual (*visual understanding*) sekaligus kemampuan generatif bahasa (*language generation*).

Pendekatan yang saat ini menjadi standar dalam pengembangan sistem *image captioning* adalah arsitektur *encoder-decoder*. Dalam arsitektur ini, jaringan saraf bekerja secara berurutan, di mana bagian *encoder* bertugas mengonversi citra

menjadi representasi fitur abstrak, sedangkan bagian *decoder* menerjemahkan representasi tersebut ke dalam bentuk rangkaian kata yang membentuk kalimat deskriptif. Integrasi kedua komponen ini memungkinkan sistem untuk menjembatani kesenjangan semantik antara representasi numerik citra dan makna bahasa alami manusia (Zakir Hossain et al., 2019).



Gambar 2.1 Arsitektur Generik Sistem Image Captioning Berbasis CNN VGG16 dan LSTM

(Sumber: Stefanini et al., 2022)

Arsitektur ini terdiri atas dua komponen utama, yaitu visual *encoder* dan language *decoder*, yang bekerja secara berurutan.

a. Visual *Encoder* (Bagian Kiri):

Citra masukan terlebih dahulu diproses menggunakan jaringan *Convolutional Neural Network* (CNN). Pada penelitian ini, digunakan model VGG16 sebagai ekstraktor fitur visual utama. Proses ini menghasilkan representasi fitur berupa vektor fitur visual yang mengandung informasi esensial terkait objek, bentuk, serta struktur spasial yang terdapat di dalam citra.

b. Language *Decoder* (Bagian Kanan):

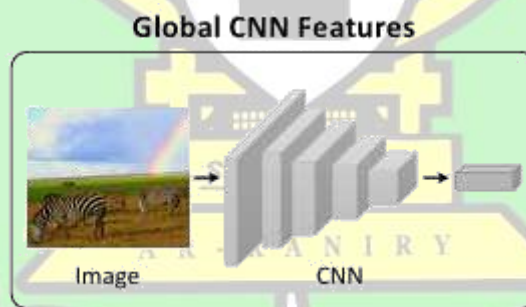
Vektor fitur visual tersebut selanjutnya menjadi masukan bagi jaringan *Long Short-Term Memory* (LSTM) yang bertindak sebagai *decoder* bahasa. Pada setiap langkah waktu, LSTM menghasilkan prediksi

probabilitas kata dengan mempertimbangkan kata yang dihasilkan pada tahap sebelumnya serta konteks visual dari citra tersebut. Proses ini berlangsung secara berulang hingga terbentuk rangkaian kata yang menyusun deskripsi citra secara utuh.

2.2.2 *Convolutional Neural Network (CNN)*

Dalam ranah pengolahan citra digital, arsitektur *Convolutional Neural Network (CNN)* dinilai memiliki efektivitas tinggi dibandingkan metode konvensional. Menurut (Alzubaidi et al., 2021a), CNN memiliki keunggulan dibandingkan jaringan saraf tradisional karena kemampuannya dalam mempertahankan hubungan spasial antar piksel melalui penggunaan lapisan konvolusi (*convolutional layers*).

Dalam sistem *image captioning*, CNN berperan sebagai *feature extractor*, yaitu mengekstraksi informasi visual penting dari citra sebelum diproses oleh model bahasa. Alih-alih menggunakan citra mentah secara langsung, gambar terlebih dahulu diubah menjadi vektor fitur berdimensi tinggi yang kaya informasi. Penelitian ini menggunakan model pra-latih VGG16, yang menurut (Sarker, 2021b) masih banyak digunakan karena struktur arsitekturnya yang relatif sederhana namun mampu dalam untuk menangkap pola visual yang kompleks secara efektif.



Gambar 2.2 Menunjukkan alur kerja arsitektur *Convolutional Neural Network (CNN)* standar (Sumber: Alzubaidi et al., 2021)

Gambar 2.2 menunjukkan alur kerja arsitektur *Convolutional Neural Network (CNN)* standar. Proses diawali dari input citra, kemudian citra diproses melalui lapisan konvolusi untuk mengekstraksi fitur visual, dilanjutkan dengan

pooling untuk mengurangi dimensi fitur, dan diakhiri dengan lapisan fully connected untuk menghasilkan representasi fitur.

Di antara lapisan konvolusi tersebut, diterapkan lapisan pooling (berwarna biru atau abu-abu) yang bertujuan untuk mengurangi dimensi data serta menekan kompleksitas komputasi tanpa menghilangkan informasi utama. Pada tahap akhir (bagian kanan), hasil ekstraksi fitur diratakan (*flatten*) dan diteruskan ke lapisan *fully connected* (berwarna kuning) untuk menghasilkan representasi vektor fitur akhir sebagai keluaran jaringan.

Dalam konteks penelitian ini, vektor fitur keluaran dari CNN tersebut tidak langsung digunakan untuk klasifikasi, melainkan dimanfaatkan sebagai masukan bagi jaringan *Long Short-Term Memory* (LSTM) untuk proses pembangkitan deskripsi kalimat (*image captioning*).

Secara matematis, proses utama pada *Convolutional Neural Network* (CNN) dilakukan melalui operasi konvolusi antara citra masukan dan filter atau kernel. Operasi ini bertujuan untuk mengekstraksi pola visual tertentu, seperti tepi, tekstur, bentuk, warna, dan struktur objek pada citra. Proses konvolusi dapat dituliskan sebagai berikut:

$$h^k = f(W^k * x + b^k) \quad (1)$$

Keterangan:

h^k : keluaran feature map ke-k

W^k : bobot filter atau kernel ke-k

x : citra masukan

b^k : bias pada filter ke-k

$*$: operasi konvolusi

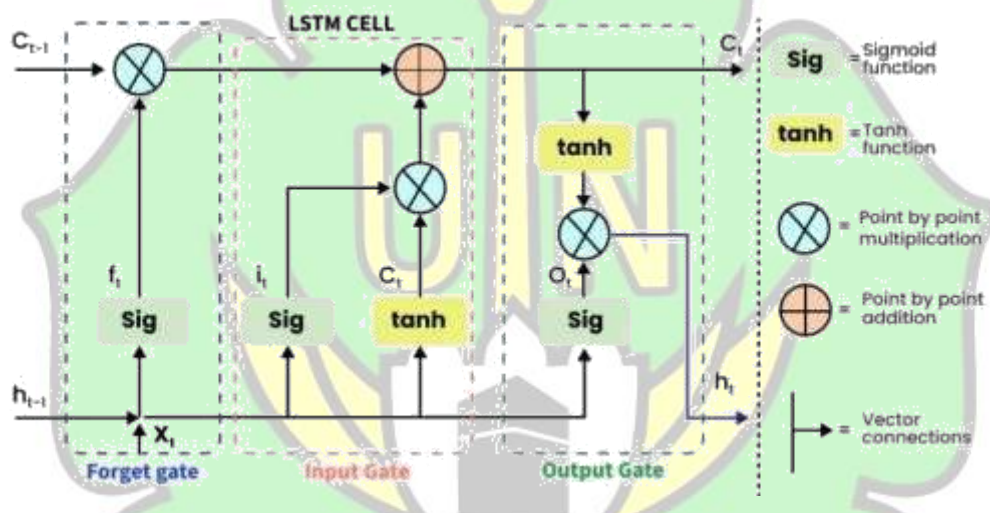
f : fungsi aktivasi, seperti ReLU

2.2.3 Long Short-Term Memory (LSTM)

Untuk menangani data sekuensial berupa teks deskripsi, penelitian ini menggunakan *Long Short-Term Memory* (LSTM). Berdasarkan penelitian oleh Teng et al. (2022), LSTM merupakan pengembangan dari *Recurrent Neural*

Network (RNN) yang dirancang secara khusus untuk mengatasi permasalahan gradien yang menghilang (*vanishing gradient problem*) pada pemodelan data sekuens dengan ketergantungan jangka panjang.

Mekanisme pengendalian aliran informasi pada LSTM dilakukan melalui struktur sel memori unik yang terdiri dari tiga gerbang utama: gerbang masukan (*input*), gerbang pelupakan (*forget*), dan gerbang keluaran (*output*). Mekanisme ini memungkinkan model untuk mengatur aliran informasi dengan menentukan data yang perlu disimpan dalam memori jangka panjang maupun informasi yang harus diabaikan, sehingga LSTM sangat sesuai digunakan untuk menyusun kalimat deskripsi yang memiliki keterkaitan konteks antar kata (Alzubaidi et al., 2021b).



Gambar 2.3 Struktur Sel Long Short-Term Memory (LSTM) (Sumber: GeeksforGeeks, 2025)

Berdasarkan Gambar 2.3, struktur sel Long Short-Term Memory (LSTM) dikendalikan oleh tiga gerbang utama, yaitu forget gate, input gate, dan output gate. Ketiga gerbang tersebut memiliki peran yang berbeda dalam mengatur aliran informasi di dalam sel LSTM.

1. Forget Gate (Gerbang Pelupakan):

Berfungsi untuk menentukan informasi dari status sebelumnya yang perlu dipertahankan atau dihapus dari memori sistem. Keputusan ini dilakukan melalui fungsi aktivasi sigmoid yang menghasilkan nilai dalam rentang antara 0 dan 1.

2. Input Gate (Gerbang Masukan):

Berperan dalam menyeleksi informasi baru yang akan ditambahkan ke dalam memori sel (*cell state*). Gerbang ini memastikan bahwa hanya informasi yang relevan secara kontekstual yang disimpan untuk memperkaya memori jangka panjang.

3. Output Gate (Gerbang Keluaran):

Mengatur informasi akhir yang akan dikeluarkan sebagai status tersembunyi (*hidden state*). Hasil dari gerbang ini akan digunakan pada proses selanjutnya atau sebagai keluaran sementara dari jaringan saraf tersebut.

Secara matematis, mekanisme kerja *Long Short-Term Memory* (LSTM) tidak hanya melibatkan tiga gerbang utama, yaitu input gate, forget gate, dan output gate, tetapi juga mencakup proses pembaruan cell state dan pembentukan hidden state. Tiga gerbang utama berfungsi untuk mengatur informasi yang masuk, dilupakan, dan dikeluarkan, sedangkan cell state dan hidden state merepresentasikan memori serta keluaran yang digunakan pada langkah waktu berikutnya. Oleh karena itu, persamaan LSTM dapat dijelaskan melalui lima komponen utama sebagai berikut:

a. Input Gate

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{(t-1)}) \quad (2)$$

b. Forget Gate

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{(t-1)}) \quad (3)$$

c. Output Gate

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{(t-1)}) \quad (4)$$

d. Cell State

$$c_t = f_t \odot c_{(t-1)} + i_t \odot \tanh(W_{cx}x_t + W_{cm}m_{(t-1)}) \quad (5)$$

e. Hidden State

$$m_t = o_t \odot \tanh(c_t) \quad (6)$$

Keterangan:

i_t : input gate pada waktu ke- t

f_t : forget gate pada waktu ke- t

o_t : output gate pada waktu ke- t

c_t : cell state atau memori sel pada waktu ke- t

c_{t-1} : cell state atau memori sel pada waktu sebelumnya

m_t : hidden state atau keluaran tersembunyi pada waktu ke- t

m_{t-1} : hidden state atau keluaran tersembunyi pada waktu sebelumnya

x_t : masukan pada waktu ke- t

W : bobot model yang dipelajari selama proses pelatihan

σ : fungsi aktivasi sigmoid

$\tanh$: fungsi aktivasi hyperbolic tangent

$\odot$: operasi perkalian elemen atau element-wise multiplication

Melalui mekanisme tersebut, LSTM mampu mempertahankan informasi penting dari urutan kata sebelumnya dan mengabaikan informasi yang tidak relevan. Kemampuan ini menjadikan LSTM sesuai digunakan sebagai *decoder* dalam sistem *image captioning*, karena proses pembangkitan *caption* dilakukan secara sekuensial dari satu kata ke kata berikutnya.

2.2.4 Dataset Flickr8k

Dataset yang digunakan sebagai acuan evaluasi dalam penelitian ini adalah Flickr8k. Berdasarkan survei dataset *image captioning* yang dilakukan oleh Stefanini et al. (2022), Flickr8k direkomendasikan untuk penelitian berskala menengah karena memiliki ukuran yang relatif efisien, yaitu terdiri atas 8.000 citra, namun tetap kaya akan anotasi. Setiap citra pada dataset ini dilengkapi dengan lima

kalimat deskripsi yang berbeda, sehingga total tersedia sekitar 40.000 kalimat referensi.

Penggunaan Flickr8k memungkinkan model untuk mempelajari variasi linguistik dalam deskripsi citra tanpa memerlukan sumber daya komputasi yang besar sebagaimana dibutuhkan oleh dataset berskala lebih besar, seperti MS COCO.

Struktur data pada dataset Flickr8k bertindak sebagai *ground truth* (kebenaran dasar) dalam pelatihan model *Image Captioning*. Kehadiran lima variasi deskripsi tekstual untuk setiap citra berperan sangat penting dalam membantu model mempelajari penggunaan sinonim serta keberagaman susunan kalimat untuk merepresentasikan konteks visual yang sama. Dalam penerapannya, deskripsi teks dari dataset ini harus melalui tahap preprocessing, di mana salah satu langkah krusialnya adalah penambahan token penanda khusus pada awal dan akhir kalimat (seperti token start dan end). Secara teoretis, token-token ini berfungsi sebagai sinyal navigasi bagi model *Long Short-Term Memory* (LSTM) untuk mengetahui secara pasti kapan harus memulai proses memprediksi kata pertama, dan kapan proses pembangkitan kalimat harus dihentikan.

2.2.5 BLEU Score (*Bilingual Evaluation Understudy*)

Evaluasi kinerja model dalam penelitian ini dilakukan menggunakan metrik *BLEU Score*. Berdasarkan survei metrik evaluasi *Natural Language Generation* (NLG) yang dilakukan oleh Sai et al. (2023), *BLEU* masih menjadi salah satu standar evaluasi utama karena memiliki tingkat korelasi yang mampu tinggi dengan penilaian manusia, khususnya dalam mengukur presisi n-gram.

Prinsip kerja *BLEU Score* adalah menghitung tingkat kesesuaian antara kalimat deskripsi yang dihasilkan oleh sistem (*Candidate*) dengan kalimat referensi (*Reference*) berdasarkan kecocokan potongan kata (n-gram). Selain itu, metrik ini juga menerapkan brevity penalty untuk memberikan penalti terhadap kalimat yang dihasilkan apabila panjangnya terlalu pendek dibandingkan dengan kalimat referensi.

Secara matematis, *BLEU Score* dihitung berdasarkan presisi n-gram antara kalimat kandidat yang dihasilkan oleh sistem dan kalimat referensi yang dibuat oleh manusia. Presisi n-gram menunjukkan tingkat kecocokan potongan kata atau frasa

pada kalimat kandidat terhadap kalimat referensi. Rumus presisi n-gram dapat dituliskan sebagai berikut:

$$p_n = \frac{\text{jumlah } n - \text{gram kandidat yang cocok dengan referensi}}{\text{jumlah seluruh } n - \text{gram kandidat}} \quad (7)$$

Keterangan:

p_n : presisi n-gram

n-gram cocok: jumlah n-gram kandidat yang sesuai dengan referensi

seluruh n-gram kandidat: total n-gram pada kalimat hasil prediksi model

Selain menghitung presisi n-gram, *BLEU Score* juga menggunakan *brevity penalty* untuk memberikan penalti terhadap kalimat kandidat yang terlalu pendek dibandingkan dengan kalimat referensi. Rumus *brevity penalty* dapat dituliskan sebagai berikut:

$$BP = 1, \text{ jika } c > r \quad (8)$$

$$BP = e^{1 - \frac{c}{r}}, \text{ jika } c \leq r \quad (9)$$

Adapun rumus akhir BLEU Score adalah sebagai berikut:

$$BLEU = BP \cdot \exp(\sum w_n \log p_n) \quad (10)$$

Keterangan:

p_n : presisi n-gram

Countclip(ngram): jumlah n-gram kandidat yang dibatasi berdasarkan kemunculannya pada referensi

Count(ngram): jumlah n-gram pada kalimat kandidat

c : panjang kalimat kandidat

r : panjang kalimat referensi

BP: brevity penalty

w_n : bobot untuk setiap n-gram

N: jumlah maksimum n-gram yang digunakan dalam evaluasi

Dalam penelitian ini, *BLEU Score* digunakan untuk mengevaluasi kesesuaian deskripsi berbahasa Inggris yang dihasilkan oleh model terhadap deskripsi referensi pada dataset Flickr8k. Evaluasi dilakukan menggunakan BLEU-1 hingga BLEU-4, di mana BLEU-1 mengukur kecocokan tingkat kata, sedangkan BLEU-4 mengukur kecocokan frasa yang lebih panjang.

Skor BLEU direpresentasikan dalam rentang nilai 0 hingga 1 (atau sering dikonversi ke skala 0 hingga 100). Secara teoretis, sangat sulit bagi model machine learning untuk mencapai skor sempurna (1 atau 100) karena tingginya keanekaragaman dan subjektivitas bahasa manusia dalam mendeskripsikan suatu gambar. Oleh karena itu, dalam konteks pemrosesan bahasa alami dan Image Captioning, skor yang diperoleh tidak harus bernilai 1 untuk dianggap berhasil. Secara umum, nilai BLEU di bawah 0,30 menunjukkan kualitas tata bahasa yang masih sangat kaku, sedangkan skor di antara 0,30 hingga 0,50 dianggap dapat dipahami dengan kualitas menengah. Sebuah model komputasi sudah dikategorikan memiliki kinerja yang sangat baik (berkualitas tinggi) apabila mampu meraih skor BLEU-1 pada rentang 0,50 hingga 0,60 (50% - 60%). Pada rentang ini, deskripsi yang dihasilkan oleh sistem sudah sangat memadai, koheren, dan relevan dengan makna aslinya.

Untuk memperjelas mekanisme penilaian tersebut, Tabel 2.1 berikut menyajikan ilustrasi sederhana perhitungan kecocokan n-gram.

Tabel 2.1 Ilustrasi Konsep Pencocokan N-Gram pada *BLEU Score*

Komponen	Kalimat / Hasil Evaluasi
Referensi A (Manusia)	"The cat is on the mat"
Referensi B (Manusia)	"There is a cat on the mat"
Kandidat (Prediksi Sistem)	"The cat is on the mat"
Evaluasi 1-Gram (Kata)	100% Cocok (Semua kata "The", "cat", "is", "on", "the", "mat" muncul di Referensi)

Evaluasi 2-Gram (Frasa)	100% Cocok (Frasa "The cat", "cat is", "is on", "on the", "the mat" muncul berurutan)
-------------------------	---

2.2.6 Beam Search

Beam Search merupakan salah satu strategi decoding yang digunakan dalam proses pembangkitan teks untuk memilih urutan kata terbaik berdasarkan probabilitas gabungan. Berbeda dengan *greedy search* yang hanya memilih satu kata dengan probabilitas tertinggi pada setiap langkah prediksi, *Beam Search* mempertahankan beberapa kandidat urutan kata terbaik secara bersamaan. Dengan cara ini, model memiliki peluang lebih besar untuk menghasilkan *caption* yang lebih stabil dan koheren. Dalam penelitian ini, *Beam Search* digunakan pada tahap inference untuk menentukan *caption* akhir berdasarkan kandidat urutan kata terbaik yang dihasilkan oleh model.

2.2.7 Pengolahan Data Awal (Preprocessing)

Sebelum data digunakan dalam proses pelatihan model, tahap pengolahan awal perlu dilakukan untuk mengubah data mentah menjadi format yang dapat diproses oleh mesin. Menurut Sarker (2021), kualitas data masukan memiliki pengaruh yang signifikan terhadap kinerja dan keberhasilan model *Deep Learning*.

1. Pengolahan Citra (Image Preprocessing)

Model VGG16 yang digunakan sebagai *encoder* memiliki spesifikasi masukan yang bersifat ketat. Oleh karena itu, setiap citra dalam dataset Flickr8k harus melalui beberapa tahapan pengolahan awal sebagai berikut:

a. Resizing:

Mengubah dimensi citra asli yang bervariasi menjadi ukuran tetap 224×224 piksel agar sesuai dengan input layer pada arsitektur VGG16.

b. Normalisasi:

Proses penyesuaian nilai intensitas piksel secara spesifik mengikuti standar pengolahan awal bawaan arsitektur VGG16 (melalui fungsi `preprocess_input`). Normalisasi pada VGG16 tidak

menggunakan penskalaan rentang 0 hingga 1, melainkan menerapkan teknik mean subtraction (pengurangan nilai rata-rata) dan konversi format warna dari RGB ke BGR. Nilai intensitas setiap piksel pada citra akan dikurangi dengan nilai rata-rata piksel yang diperoleh dari dataset ImageNet, yakni 103.939 untuk kanal Biru (Blue), 116.779 untuk kanal Hijau (Green), dan 123.680 untuk kanal Merah (Red). Teknik pemusatan data (zero-centering) ini bertujuan untuk menstabilkan dan mempercepat proses konvergensi gradien selama pelatihan model.

2. Pengolahan *Caption* (Text Preprocessing)

Data deskripsi citra (*caption*) dalam bentuk teks perlu diubah menjadi representasi numerik melalui beberapa tahapan berikut (Stefanini et al., 2022):

a. *Cleaning*:

Menghapus karakter non-alfabet dan tanda baca (seperti koma, titik, atau tanda seru), serta mengonversi seluruh huruf kapital menjadi huruf kecil (*lowercase*). Sebagai contoh, jika kalimat asli (*raw text*) di dalam dataset berbunyi: "*A brown dog, running through the grass!!*", maka setelah proses *cleaning*, kalimat tersebut akan diubah menjadi: "*a brown dog running through the grass*". Langkah ini bertujuan untuk menyeragamkan format teks dan menghilangkan variasi karakter yang tidak esensial (*noise*), sehingga model dapat berfokus pada pengenalan kata-kata utama.

b. Tokenisasi:

Menambahkan token penanda awal *startseq* dan token penanda akhir *endseq* pada setiap kalimat, kemudian memecah kalimat menjadi unit kata (*token*). Melanjutkan contoh pada tahap sebelumnya, kalimat "*a brown dog running through the grass*" akan ditambahkan penanda sekuensial menjadi: "*startseq a brown dog running through the grass endseq*". Selanjutnya, kalimat tersebut dipecah ke dalam bentuk himpunan atau *array* kata individual sebagai berikut: ['startseq', 'a', 'brown', 'dog', 'running', 'through',

'the', 'grass', 'endseq']. Penambahan token khusus ini sangat krusial karena berfungsi sebagai sinyal bagi model LSTM untuk mengetahui secara pasti kapan harus memulai memprediksi kata pertama dan kapan harus menghentikan proses pembangkitan kalimat.

c. *Padding*:

Menyamakan panjang seluruh urutan data (*sequence*) dengan menambahkan nilai nol (0) pada kalimat yang memiliki panjang lebih pendek dari panjang maksimum dalam dataset.

2.2.8 Optimasi Model (*Model Optimization*)

Proses pelatihan model pada penelitian ini bertujuan untuk meminimalkan tingkat kesalahan prediksi yang dihasilkan oleh sistem sehingga mampu menghasilkan deskripsi yang akurat. Untuk mencapai performa optimal, penelitian ini menerapkan konfigurasi optimasi sebagai berikut:

1. Fungsi Kerugian (*Loss Function*)

Penelitian ini menggunakan *Sparse Categorical Crossentropy* sebagai fungsi kerugian utama. Pemilihan fungsi ini didasarkan pada karakteristik proses pembangkitan caption yang memprediksi satu kata berikutnya dari sejumlah kosakata yang tersedia. Setiap kata target direpresentasikan dalam bentuk indeks bilangan bulat, sehingga fungsi *Sparse Categorical Crossentropy* sesuai digunakan untuk menghitung selisih antara prediksi probabilitas model dan label kata yang sebenarnya. Nilai kesalahan tersebut kemudian digunakan untuk memperbarui bobot model selama proses pelatihan.

2. *Algoritma Optimasi (Optimizer)*

Pembaruan bobot pada jaringan saraf dilakukan menggunakan algoritma Adam (*Adaptive Moment Estimation*). Menurut Teng et al. (2022), Adam merupakan optimizer yang sangat efisien dalam menangani dataset berskala besar serta arsitektur berbasis recurrent seperti LSTM. Keunggulan utama Adam terletak pada kemampuannya untuk menyesuaikan tingkat pembelajaran (*learning rate*) secara adaptif pada setiap parameter, sehingga proses konvergensi menuju nilai minimum

fungsi kerugian dapat tercapai lebih cepat dibandingkan algoritma optimasi konvensional lainnya.

2.2.9 Lingkungan Pengembangan (*Development Environment*)

Implementasi sistem dalam penelitian ini dilakukan dengan memanfaatkan ekosistem perangkat lunak dan pustaka pendukung yang berstandar industri guna memastikan modularitas dan stabilitas sistem:

1. Bahasa Pemrograman Python: Digunakan sebagai bahasa utama karena memiliki ekosistem pustaka kecerdasan buatan (*artificial intelligence*) yang sangat luas serta dukungan komunitas pengembang yang aktif.
2. TensorFlow dan Keras: Bertindak sebagai *framework* utama dalam membangun, melatih, dan menguji arsitektur CNN VGG16 dan LSTM. Pustaka ini dipilih karena kemampuannya dalam melakukan abstraksi lapisan jaringan saraf secara modular dan terstruktur.
3. Google Translate API: Dimanfaatkan pada tahap pasca-pemrosesan (*post-processing*) untuk melakukan translasi otomatis terhadap hasil deskripsi citra yang awalnya terbangun dalam Bahasa Inggris ke dalam Bahasa Indonesia. Hal ini dilakukan guna meningkatkan aksesibilitas dan kemudahan pemahaman bagi pengguna lokal.

BAB III

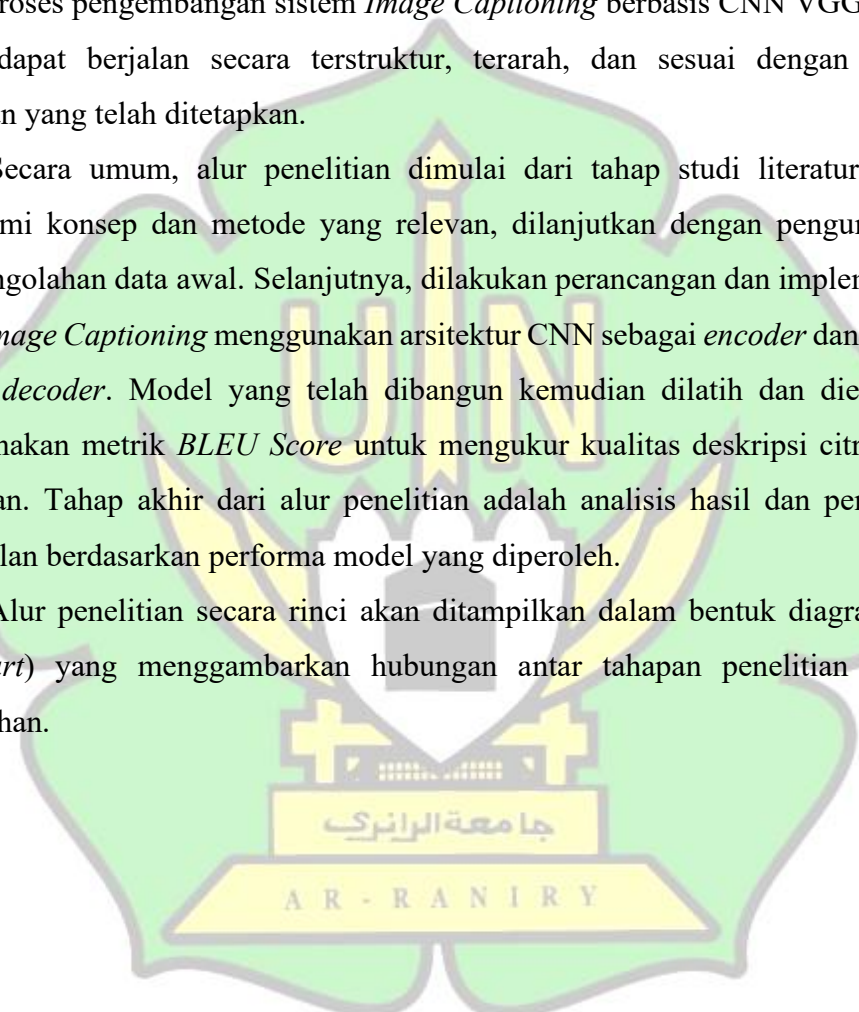
METODE PENELITIAN

3.1 Alur Penelitian

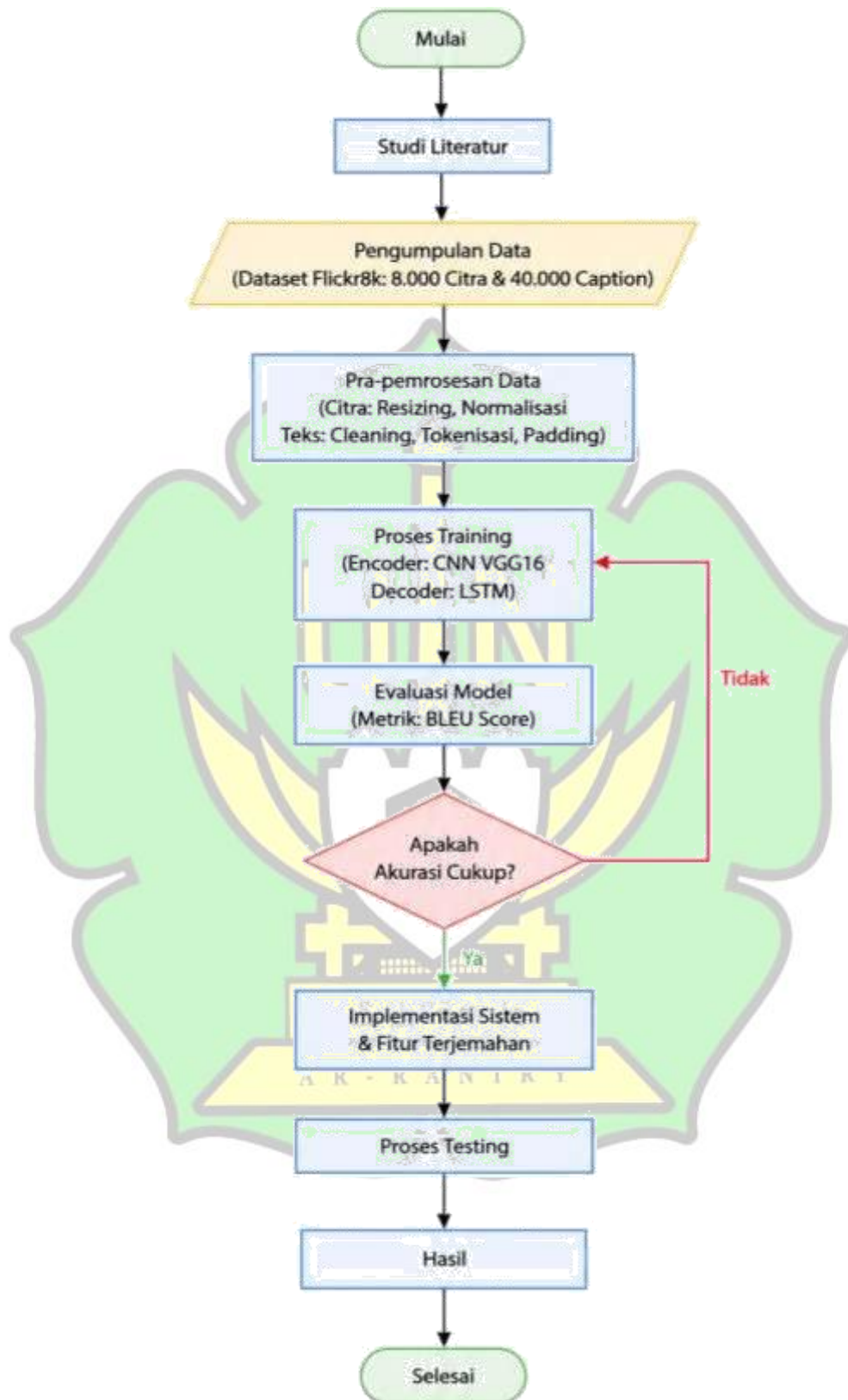
Alur penelitian merupakan tahapan sistematis yang dilakukan penulis untuk menyelesaikan permasalahan penelitian mulai dari tahap perencanaan hingga penarikan kesimpulan. Penyusunan alur penelitian ini bertujuan untuk memastikan bahwa proses pengembangan sistem *Image Captioning* berbasis CNN VGG16 dan LSTM dapat berjalan secara terstruktur, terarah, dan sesuai dengan tujuan penelitian yang telah ditetapkan.

Secara umum, alur penelitian dimulai dari tahap studi literatur untuk memahami konsep dan metode yang relevan, dilanjutkan dengan pengumpulan serta pengolahan data awal. Selanjutnya, dilakukan perancangan dan implementasi model *Image Captioning* menggunakan arsitektur CNN sebagai *encoder* dan LSTM sebagai *decoder*. Model yang telah dibangun kemudian dilatih dan dievaluasi menggunakan metrik *BLEU Score* untuk mengukur kualitas deskripsi citra yang dihasilkan. Tahap akhir dari alur penelitian adalah analisis hasil dan penarikan kesimpulan berdasarkan performa model yang diperoleh.

Alur penelitian secara rinci akan ditampilkan dalam bentuk diagram alir (*flowchart*) yang menggambarkan hubungan antar tahapan penelitian secara keseluruhan.



3.1.1 Diagram Alir Penelitian (*Flowchart*)



Gambar 3.1 Alur Penelitian

Berdasarkan diagram alir penelitian yang telah disajikan, rincian kegiatan pada setiap tahapan penelitian dapat dijelaskan sebagai berikut:

1. Studi Literatur

Tahap ini merupakan langkah awal penelitian, di mana penulis mempelajari konsep teoretis dan penelitian terdahulu yang relevan dengan topik yang diangkat. Fokus kajian meliputi prinsip kerja *Convolutional Neural Network* (CNN) dalam ekstraksi fitur citra, *Long Short-Term Memory* (LSTM) dalam pemrosesan data sekuensial teks, serta metrik evaluasi *BLEU Score* untuk menilai kualitas hasil pembangkitan teks. Referensi diperoleh dari buku, jurnal ilmiah nasional dan internasional, serta artikel prosiding yang diterbitkan dalam kurun waktu lima tahun terakhir.

2. Pengumpulan Dataset

Data yang digunakan dalam penelitian ini adalah dataset publik Flickr8k. Dataset tersebut diperoleh dari *repository Kaggle* dan terdiri dari 8.000 citra digital yang masing-masing dilengkapi dengan anotasi deskripsi. Selanjutnya, dataset dibagi menjadi tiga bagian, yaitu:

- a. *Training Set* (Himpunan Pelatihan): Data utama yang digunakan untuk melatih model dalam mempelajari hubungan antara fitur visual citra dan urutan kata pada *caption*. Pada tahap ini, model menyesuaikan parameter internal seperti bobot dan bias berdasarkan pola yang terdapat pada data latih.
- b. *Validation Set* (Himpunan Validasi): Data yang digunakan untuk memantau performa model selama proses pelatihan. Himpunan validasi berfungsi untuk mengevaluasi kemampuan generalisasi model, membantu pengaturan *hyperparameter*, serta mencegah terjadinya *overfitting*.
- c. *Test Set* (Himpunan Uji): Data yang digunakan untuk mengevaluasi performa akhir model setelah proses pelatihan selesai. Data uji tidak digunakan dalam proses pelatihan maupun validasi, sehingga dapat mencerminkan kemampuan model dalam menghasilkan deskripsi terhadap citra yang belum pernah dilihat sebelumnya.

3. Pengolahan Data Awal (Preprocessing)

Sebelum data digunakan dalam proses pelatihan model, dilakukan tahap pengolahan awal untuk menyesuaikan format data agar kompatibel dengan arsitektur jaringan saraf yang digunakan, yaitu:

- a. Citra: Dilakukan pengubahan ukuran (*resizing*) citra menjadi 224×224 piksel sesuai dengan spesifikasi input model VGG16, serta normalisasi nilai piksel untuk meningkatkan stabilitas proses pelatihan.
- b. Teks: Dilakukan pembersihan tanda baca, pengubahan huruf menjadi huruf kecil (*case folding*), proses tokenisasi untuk mengonversi kata menjadi indeks numerik, serta *padding* untuk menyamakan panjang setiap kalimat deskripsi.

4. Perancangan dan Pelatihan Model

Proses pelatihan model dilakukan menggunakan optimizer Adam dan fungsi kerugian Sparse Categorical Crossentropy. Fungsi kerugian tersebut digunakan karena model melakukan prediksi kata berikutnya berdasarkan indeks kosakata yang telah dibentuk pada tahap tokenisasi.

5. Implementasi Sistem

Setelah model mencapai performa yang optimal, model diimplementasikan ke dalam bentuk prototipe aplikasi sederhana menggunakan bahasa pemrograman Python. Pada tahap ini, ditambahkan fitur penerjemahan otomatis menggunakan *Google Translate API* untuk mengonversi hasil deskripsi dari bahasa Inggris ke bahasa Indonesia agar lebih mudah dipahami oleh pengguna.

6. Pengujian dan Evaluasi

Tahap pengujian dan evaluasi bertujuan untuk mengukur kinerja sistem yang telah dibangun. Pengujian dilakukan melalui dua skenario, yaitu:

- a. Evaluasi Kuantitatif: Menggunakan metrik BLEU Score untuk mengukur tingkat kemiripan antara teks deskripsi yang dihasilkan sistem dengan teks referensi pada data uji.

- b. Evaluasi Kualitatif: Menguji sistem menggunakan citra dunia nyata, seperti gambar jurnalistik atau berita, untuk menilai kemampuan model dalam mengenali objek dan aktivitas utama secara visual dibandingkan dengan deskripsi yang dibuat oleh manusia.

Keputusan “Apakah Akurasi mampu?” pada flowchart digunakan untuk menentukan apakah model telah memenuhi indikator keberhasilan yang ditetapkan. Apabila hasil evaluasi belum memenuhi target, maka proses kembali ke tahap pelatihan untuk dilakukan penyesuaian model, seperti pengaturan parameter, jumlah epoch, atau konfigurasi pelatihan. Namun, apabila hasil evaluasi telah memenuhi target, maka model terbaik digunakan pada tahap implementasi sistem dan fitur penerjemahan. Pada tahap ini, model menghasilkan caption berbahasa Inggris, kemudian caption tersebut diterjemahkan ke dalam bahasa Indonesia agar lebih mudah dipahami oleh pengguna.

7. Penarikan Kesimpulan

Tahap akhir penelitian adalah menganalisis seluruh hasil pengujian dan evaluasi yang telah dilakukan. Berdasarkan analisis tersebut, penulis menarik kesimpulan mengenai kinerja metode CNN VGG16 dan LSTM yang diusulkan serta memberikan saran untuk pengembangan penelitian di masa mendatang.

3.2 Metode Pengumpulan Data

Data yang digunakan dalam penelitian ini merupakan data sekunder yang bersumber dari dataset standar publik Flickr8k, yang umum digunakan dalam penelitian *Image Captioning*.

1. Sumber Data

Dataset Flickr8k diperoleh dari repositori resmi Universitas Illinois serta tersedia secara publik melalui platform Kaggle. Dataset ini berisi kumpulan foto yang diambil dari situs Flickr dan telah dianotasi secara manual oleh manusia.

2. Spesifikasi Data

Karakteristik dataset yang digunakan dalam penelitian ini adalah sebagai berikut:

- a. Jumlah Citra: Dataset terdiri dari 8.000 citra digital yang menampilkan beragam objek, aktivitas manusia, serta konteks visual sehari-hari.
- b. Jumlah Deskripsi: Setiap citra dilengkapi dengan 5 kalimat deskripsi (*caption*) berbahasa Inggris, sehingga total terdapat 40.000 kalimat deskripsi dalam dataset ini.



Gambar 3.2 Sampel Citra dan Anotasi Deskripsi pada Dataset Flickr8k (Sumber: Stefanini et al., 2022)

Gambar 3.2 memperlihatkan secara langsung struktur data aktual pada dataset Flickr8k yang digunakan sebagai ground truth. Setiap satu citra dilengkapi dengan lima variasi deskripsi tekstual yang disusun oleh manusia, yang sangat berguna bagi model untuk mempelajari penggunaan sinonim dan variasi struktur kalimat.

3. Pembagian Data (Data Splitting)

Sesuai dengan standar eksperimen pada dataset Flickr8k, data dibagi menjadi tiga himpunan sebagai berikut:

- a. Data Latih (*Training Set*): Sebanyak 6.000 citra (80%) digunakan untuk melatih bobot model agar mampu mempelajari hubungan antara fitur visual dan deskripsi tekstual.
- b. Data Validasi (*Development/Validation Set*): Sebanyak 1.000 citra (10%) digunakan untuk memantau kinerja model selama proses pelatihan serta mencegah terjadinya *overfitting*.
- c. Data Uji (*Test Set*): Sebanyak 1.000 citra (10%) digunakan khusus untuk pengujian akhir dan tidak pernah diperlihatkan kepada model selama proses pelatihan.

3.3 Tahapan Pengolahan Data Awal (Preprocessing)

Sebelum digunakan dalam proses pelatihan model CNN VGG16 dan LSTM, data citra dan data teks terlebih dahulu melalui tahap pengolahan data awal atau preprocessing. Tahap ini dilakukan karena data mentah tidak dapat langsung diproses oleh model jaringan saraf. Citra memiliki ukuran dan nilai piksel yang bervariasi, sedangkan *caption* masih berbentuk kalimat bahasa alami. Oleh karena itu, data perlu diseragamkan dan diubah ke dalam bentuk numerik agar sesuai dengan kebutuhan arsitektur model. Dalam penelitian ini, preprocessing dibagi menjadi dua bagian utama, yaitu pengolahan citra dan pengolahan *caption*.

3.3.1 Pengolahan Citra (Image Preprocessing)

Pengolahan citra dilakukan untuk menyesuaikan format gambar agar dapat diterima oleh model CNN VGG16. Dataset Flickr8k terdiri dari citra dengan ukuran yang berbeda-beda, sedangkan VGG16 membutuhkan input citra dengan ukuran tetap. Oleh karena itu, setiap citra harus melalui proses resizing, konversi ke bentuk array, dan normalisasi sebelum fitur visualnya diekstraksi oleh model. Tahapan yang dilakukan meliputi:

1. Pengubahan Ukuran (*Resizing*)

Setiap citra dengan dimensi asli yang bervariasi diubah ukurannya

menjadi 224×224 piksel agar sesuai dengan ukuran input standar pada arsitektur VGG16.

2. Konversi ke Bentuk Array

Citra yang telah diubah ukurannya kemudian dikonversi ke dalam bentuk array numerik tiga dimensi yang merepresentasikan kanal warna Red, Green, dan Blue (RGB) sehingga dapat diproses oleh jaringan saraf.

3. Normalisasi dan *Preprocessing* VGG16

Nilai intensitas piksel citra dinormalisasi menggunakan fungsi `preprocess_input` bawaan VGG16. Proses ini umumnya melibatkan penyesuaian nilai piksel berdasarkan rata-rata kanal RGB dari dataset ImageNet, dengan tujuan untuk mempercepat proses konvergensi dan meningkatkan stabilitas pelatihan model.

3.3.2 Pengolahan *Caption* (Text Preprocessing)

Pengolahan *caption* dilakukan untuk mengubah kalimat deskripsi menjadi bentuk numerik yang dapat diproses oleh model LSTM. Hal ini diperlukan karena LSTM tidak membaca teks dalam bentuk kalimat biasa, melainkan dalam bentuk urutan angka yang merepresentasikan kata. Oleh karena itu, *caption* terlebih dahulu dibersihkan, diberi token awal dan akhir, diubah menjadi indeks numerik melalui tokenisasi, serta diseragamkan panjangnya menggunakan padding. Tahapan pengolahan awal teks yang dilakukan dalam penelitian ini meliputi:

1. Pembersihan Teks (Text Cleaning)

Seluruh kalimat deskripsi diubah ke dalam format huruf kecil (*lowercase*), tanda baca dihapus, serta karakter non-alfabet seperti angka dan simbol dibuang guna mengurangi variasi data yang tidak relevan.

2. Penambahan Token Khusus

Pada setiap kalimat ditambahkan token penanda awal *startseq* dan token penanda akhir *endseq*. Token ini berfungsi sebagai sinyal bagi model LSTM untuk menentukan awal dan akhir proses pembangkitan kalimat.

3. Tokenisasi (*Tokenization*)

Setiap kata unik dalam dataset dipetakan menjadi bilangan bulat (*integer index*) berdasarkan kamus kosakata (*vocabulary*) yang dibangun

dari data latih. Proses ini bertujuan untuk mengubah teks menjadi urutan numerik yang dapat diproses oleh jaringan saraf.

4. *Padding*

Panjang seluruh urutan kata diseragamkan dengan menambahkan nilai nol (*zero padding*) pada kalimat yang lebih pendek hingga mencapai panjang maksimum kalimat dalam dataset. Langkah ini dilakukan agar dimensi input model menjadi konsisten.

3.4 Perancangan Arsitektur Model

Sistem *Image Captioning* yang dikembangkan dalam penelitian ini dirancang menggunakan arsitektur *Encoder-Decoder*, yang mengombinasikan dua komponen utama jaringan saraf tiruan, yaitu *Convolutional Neural Network* (CNN) sebagai *Encoder* dan *Long Short-Term Memory* (LSTM) sebagai *Decoder*.

3.4.1 Encoder (CNN VGG16)

Bagian *Encoder* berfungsi untuk mengekstraksi representasi fitur visual dari citra masukan. Penelitian ini memanfaatkan model VGG16 yang telah dilatih sebelumnya (*pre-trained*) menggunakan dataset *ImageNet*, sehingga model memiliki kemampuan awal dalam mengenali pola visual umum.

Lapisan klasifikasi akhir pada VGG16 tidak digunakan karena penelitian ini tidak berfokus pada klasifikasi citra. Output fitur visual diambil dari lapisan sebelum klasifikasi akhir, sehingga menghasilkan vektor fitur berdimensi 4.096. Vektor fitur tersebut kemudian diproses melalui lapisan Dense untuk menghasilkan representasi visual yang lebih sesuai sebagai masukan bagi *decoder* LSTM. Output yang digunakan berasal dari lapisan terakhir sebelum proses klasifikasi, yang merepresentasikan informasi visual tingkat tinggi dari citra. Selanjutnya, fitur tersebut diproses melalui lapisan Dense untuk menghasilkan vektor fitur berdimensi 256 yang digunakan sebagai representasi visual gambar.

3.4.2 Decoder (LSTM)

Bagian *decoder* dirancang menggunakan arsitektur functional yang bertugas mengintegrasikan fitur visual dari citra dan urutan kata pada *caption* untuk

membangkitkan deskripsi kalimat. Penggunaan LSTM dalam penelitian ini berkaitan dengan karakteristik *caption* yang berbentuk data sekuensial, yaitu data yang memiliki urutan dan saling bergantung antar elemen. Dalam proses pembangkitan *caption*, kata berikutnya tidak diprediksi secara terpisah, tetapi dipengaruhi oleh kata-kata yang telah dihasilkan sebelumnya serta fitur visual yang diekstraksi dari citra. Oleh karena itu, LSTM digunakan sebagai *decoder* karena mampu memproses urutan kata dan mempertahankan informasi konteks melalui mekanisme memori internal. Berdasarkan model yang telah diimplementasikan, *Decoder* memproses dua cabang input utama sebagai berikut:

1. Cabang Fitur Gambar (*Vision Branch*)

Menerima input berupa vektor fitur berukuran 4.096 yang diekstraksi oleh VGG16. Fitur visual yang dihasilkan oleh VGG16 selanjutnya diproses melalui lapisan Dropout dengan nilai 0,4. Nilai tersebut menunjukkan bahwa 40% neuron dinonaktifkan secara acak selama proses pelatihan. Penggunaan Dropout bertujuan untuk mengurangi risiko overfitting, yaitu kondisi ketika model terlalu menghafal pola pada data latih sehingga kurang mampu melakukan generalisasi terhadap citra baru. Setelah itu, fitur diteruskan ke lapisan Dense dengan 256 unit dan fungsi aktivasi ReLU. ReLU digunakan karena memiliki proses komputasi yang sederhana serta membantu model mempelajari representasi fitur secara lebih efektif pada jaringan saraf tiruan.

2. Cabang Urutan Teks (*Text Branch*)

Cabang urutan teks menerima input berupa urutan kata dengan panjang maksimum yang diperoleh dari hasil tahap pengolahan awal teks. Input tersebut diproses melalui lapisan Embedding untuk mengubah indeks numerik kata menjadi vektor padat berdimensi 256. Selanjutnya, hasil embedding diproses melalui lapisan Dropout dengan nilai 0,4 dan lapisan LSTM dengan 256 unit memori. Lapisan LSTM ini bertugas mempelajari ketergantungan jangka panjang antar kata di dalam kalimat, sehingga model dapat memahami hubungan konteks pada urutan *caption* yang dibangkitkan.

Kedua fitur dari cabang yang berbeda tersebut kemudian disatukan menggunakan lapisan Add. Karena kedua cabang telah memiliki dimensi

yang sama, yaitu 256, sistem dapat menggabungkan konteks visual dari citra dan konteks bahasa dari urutan kata secara langsung.

Hasil penggabungan kemudian diproses oleh lapisan Dense tambahan dengan 256 unit sebelum masuk ke lapisan keluaran atau output layer. Lapisan keluaran menggunakan fungsi aktivasi *Softmax* untuk memprediksi probabilitas kemunculan kata berikutnya berdasarkan jumlah kosakata yang terbentuk pada tahap tokenisasi.

Pada tahap pembangkitan deskripsi citra, model tidak hanya memilih satu kata dengan probabilitas tertinggi pada setiap langkah prediksi.

Sistem menggunakan metode *Beam Search* untuk mempertahankan beberapa kandidat urutan kata terbaik secara bersamaan. Dalam penelitian ini, *Beam Search* digunakan dengan parameter *beam width* sebesar 5, sehingga model dapat mengevaluasi beberapa kemungkinan susunan kata sebelum menentukan *caption* akhir. Penggunaan metode ini bertujuan untuk menghasilkan deskripsi citra yang lebih stabil, koheren secara struktur kalimat, dan relevan dengan informasi visual yang diekstraksi oleh CNN.

3.4.3 Fitur Penerjemahan (Translation)

Karena model *Image Captioning* dilatih menggunakan dataset berbahasa Inggris, sistem dilengkapi dengan modul tambahan untuk menerjemahkan hasil deskripsi ke dalam bahasa Indonesia. Teks keluaran model dalam bahasa Inggris dikirimkan ke Google Translate API pada tahap *post-processing*. Selanjutnya, hasil terjemahan ditampilkan sebagai deskripsi akhir pada antarmuka aplikasi agar dapat dipahami oleh pengguna lokal.

3.5 Skenario Pengujian

Untuk mengevaluasi kinerja sistem *Image Captioning* yang dikembangkan, penelitian ini menerapkan dua skenario pengujian, yaitu pengujian kuantitatif dan pengujian kualitatif. Pendekatan ini digunakan agar evaluasi tidak hanya berfokus pada nilai numerik semata, tetapi juga pada kualitas makna deskripsi yang dihasilkan.

Pengujian kuantitatif dilakukan dengan mengukur tingkat kesesuaian antara deskripsi yang dihasilkan oleh sistem dan deskripsi referensi menggunakan metrik evaluasi statistik. Sementara itu, pengujian kualitatif dilakukan melalui analisis visual dan semantik terhadap hasil deskripsi pada citra dunia nyata untuk menilai ketepatan pengenalan objek dan aktivitas secara deskriptif.

Pendekatan ganda ini diharapkan mampu memberikan gambaran yang lebih komprehensif mengenai performa sistem, baik dari sisi akurasi berbasis data uji standar maupun dari sisi keandalan sistem dalam kondisi penggunaan nyata.

Pada tahap pengujian, *caption* yang dihasilkan model diperoleh melalui proses decoding menggunakan metode *Beam Search* sebagaimana telah dijelaskan pada bagian perancangan *decoder*. Strategi ini digunakan untuk memilih susunan kata terbaik dari beberapa kandidat prediksi, sehingga hasil deskripsi yang dievaluasi merupakan *caption* akhir yang telah melalui proses pemilihan urutan kata secara lebih optimal.

Selain *beam width*, sistem juga menggunakan parameter *beam alpha* sebesar 0,7 untuk membantu mengatur panjang *caption* yang dihasilkan, sehingga kalimat tidak terlalu pendek maupun terlalu panjang.

3.5.1 Pengujian Kuantitatif (Quantitative Testing)

Pengujian kuantitatif dilakukan untuk mengevaluasi tingkat akurasi *sistem Image Captioning* secara objektif berdasarkan data uji standar. Evaluasi ini bertujuan untuk mengetahui sejauh mana model mampu menghasilkan deskripsi yang sesuai dengan anotasi referensi yang telah tersedia.

1. Data Uji

Pengujian menggunakan sebanyak 1.000 citra yang berasal dari test set dataset Flickr8k. Seluruh citra pada himpunan data ini tidak pernah digunakan selama proses pelatihan maupun validasi, sehingga hasil pengujian mencerminkan kemampuan generalisasi model terhadap data yang belum pernah dilihat sebelumnya.

2. Metrik Evaluasi

Kinerja model diukur menggunakan metrik *BLEU Score (Bilingual Evaluation Understudy)*. Perhitungan BLEU dilakukan dengan

membandingkan kalimat deskripsi yang dihasilkan oleh sistem (*candidate sentence*) dengan lima kalimat referensi yang dibuat oleh manusia (*reference sentences*) untuk setiap citra.

3. Indikator Keberhasilan

Nilai BLEU berada pada rentang 0 hingga 1 (atau dapat direpresentasikan dalam skala 0–100). Semakin tinggi nilai BLEU yang diperoleh, semakin besar tingkat kesesuaian antara deskripsi yang dihasilkan sistem dengan deskripsi referensi manusia. Dalam penelitian ini, evaluasi dilakukan menggunakan BLEU-1 hingga BLEU-4, di mana BLEU-1 mengukur akurasi tingkat kata atau unigram, sedangkan BLEU-4 menilai kesesuaian frasa atau n-gram yang lebih kompleks. Model dinyatakan berhasil apabila mampu meraih skor BLEU-1 pada rentang 0,50 hingga 0,60, yang merupakan standar performa arsitektur CNN VGG16 dan LSTM pada dataset Flickr8k.

Penting untuk dicatat bahwa pengukuran *BLEU Score* dilakukan secara eksklusif pada deskripsi Bahasa Inggris (output asli model) terhadap deskripsi referensi Flickr8k. Fitur penerjemahan ke Bahasa Indonesia merupakan fitur tambahan untuk antarmuka pengguna akhir dan tidak disertakan dalam perhitungan akurasi model menggunakan *BLEU Score*.

3.5.2 Pengujian Kualitatif (Qualitative Testing)

Pengujian kualitatif pada penelitian ini dilaksanakan untuk mengevaluasi kinerja dan keandalan sistem saat dihadapkan pada citra dunia nyata, yakni citra yang secara fundamental tidak terdistribusi di dalam himpunan data latih maupun data uji Flickr8k. Berbeda dengan pengujian kuantitatif yang berfokus pada perhitungan metrik matematis, tahapan ini menitikberatkan pada analisis komparatif-deskriptif secara komprehensif terhadap representasi kalimat yang dibangkitkan oleh model. Pendekatan evaluasi kualitatif ini diklasifikasikan ke dalam tiga aspek tinjauan utama berdasarkan standar evaluasi *Machine Learning*, yaitu:

1. Akurasi Pengenalan Objek Visual (*Object Recognition*)

Tahapan ini menganalisis sejauh mana arsitektur *Encoder* (CNN VGG16) mampu mengekstrak fitur spasial dan mengidentifikasi entitas atau objek utama yang terdapat di dalam citra masukan secara presisi. Kesalahan komputasi pada aspek ini umumnya diklasifikasikan sebagai *Object Hallucination*, yakni sebuah anomali di mana model menghasilkan deskripsi teks yang memuat entitas fiktif atau objek yang secara visual faktual tidak hadir di dalam gambar.

2. Ketepatan Identifikasi Atribut dan Aktivitas (*Attribute and Action Recognition*)

Evaluasi pada tahap ini berfokus pada kemampuan sistem dalam mengenali representasi karakteristik pendukung dari objek yang terdeteksi. Hal ini mencakup identifikasi warna, kondisi lingkungan fisik, interaksi spasial antar entitas, hingga aktivitas dinamis yang sedang berlangsung. Kegagalan model dalam menafsirkan postur, gerakan, atau relasi objek secara tepat akan didefinisikan sebagai Action Error atau *Attribute Error*, yang umumnya dipicu oleh tingkat kerumitan latar belakang citra (*cluttered background*) maupun oklusi objek.

3. Kelancaran Tata Bahasa dan Keutuhan Semantik (*Linguistic Fluency and Semantic Gap*)

Aspek pengujian ketiga bertujuan untuk meninjau struktur sintaksis dari kalimat berbahasa Inggris yang dihasilkan oleh arsitektur *Decoder* (LSTM). Analisis linguistik dilakukan untuk memastikan apakah probabilitas urutan kata yang dirangkai membentuk kaidah kalimat yang logis memiliki struktur subjek, predikat, dan objek yang koheren atau sekadar kumpulan leksikon acak tanpa makna. Ketidakmampuan model dalam merangkai makna yang utuh akibat batasan algoritma dalam memahami konteks visual yang kompleks akan dikaji sebagai fenomena Semantic Gap.

Hasil observasi dari ketiga parameter pengujian kualitatif tersebut selanjutnya akan ditabulasikan dan disandingkan secara manual dengan kondisi visual citra aktual. Temuan-temuan yang dihasilkan, baik berupa tingkat keberhasilan maupun limitasi model (seperti bias prediksi akibat variasi distribusi

data latih), akan dibahas secara mendalam untuk menjustifikasi efektivitas implementasi kombinasi arsitektur *Encoder-Decoder* yang diusulkan dalam penelitian ini.

3.6 Lingkungan Pengembangan

Penelitian dan pengembangan sistem ini dilakukan menggunakan spesifikasi perangkat keras dan perangkat lunak sebagai berikut:

3.6.1 Perangkat Keras (*Hardware*)

Spesifikasi komputer yang digunakan untuk proses pelatihan model (*training*) dan pembuatan aplikasi adalah:

Tabel 3.1 Perangkat keras

Komponen	Spesifikasi Minimum (Untuk Menjalankan Program)	Spesifikasi yang Digunakan (Komputer Pengembangan)
Prosesor (CPU)	Intel Core i3 / AMD Ryzen 3	Laptop ASUS VivoBook X421EQ dengan prosesor Intel® Core™ i5-1135G7 (11th Generation) @ 2.40 GHz
RAM	4 GB	8 GB DDR4 3200 MHz
Penyimpanan (Storage)	Minimal 5 GB ruang kosong (<i>Free Space</i>) pada HDD / SSD	512 GB SSD

3.6.2 Perangkat Lunak (*Software*)

Perangkat lunak dan pustaka pendukung yang digunakan meliputi:

1. Sistem Operasi: Windows 10 / 11 (64-bit).
2. Bahasa Pemrograman: Python versi 3.8 atau lebih baru.

3. Framework *Deep Learning*: *TensorFlow* 2.x dan Keras.
4. Library Pendukung:
 - a. *NumPy & Pandas* (Manipulasi data array).
 - b. *Matplotlib* (Visualisasi grafik akurasi).
 - c. *OpenCV / Pillow* (Pengolahan citra).
 - d. *NLTK* (Pengolahan teks).
5. Tools: Jupyter Notebook / Google Colab (untuk eksperimen model) dan Visual Studio Code (untuk pembuatan kode aplikasi).
6. API Eksternal: *Google Translate API* (untuk fitur penerjemahan bahasa Indonesia).

3.7 Waktu dan Tempat Penelitian

3.7.1 Tempat Penelitian

Penelitian ini merupakan penelitian berbasis pengembangan perangkat lunak dan eksperimen komputasi, sehingga pelaksanaannya tidak sepenuhnya bergantung pada satu lokasi fisik tertentu. Namun demikian, kegiatan penelitian tetap dilakukan pada lingkungan yang mendukung proses pengembangan dan pengujian sistem.

Adapun tempat pelaksanaan penelitian ini adalah sebagai berikut:

1. Laboratorium Komputasi Program Studi Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Islam Negeri (UIN) Ar-Raniry Banda Aceh, yang digunakan sebagai lokasi utama untuk kegiatan pengembangan sistem, pelatihan model (model *training*), serta konsultasi akademik dengan dosen pembimbing.
2. Fasilitas Komputasi Mandiri, yang dimanfaatkan sebagai lingkungan kerja tambahan untuk kegiatan pendukung seperti penulisan kode program (coding), eksperimen lanjutan, analisis hasil pengujian, serta penyusunan laporan penelitian secara fleksibel.

Pemilihan lokasi tersebut dinilai memadai karena menyediakan fasilitas komputasi dan lingkungan kerja yang mendukung pelaksanaan penelitian *Image Captioning* berbasis CNN VGG16 dan LSTM secara optimal.

3.7.2 Waktu Penelitian

Penelitian ini telah dilaksanakan secara bertahap dalam kurun waktu kurang lebih 6 (enam) bulan, terhitung sejak tahap pengajuan judul hingga penyusunan laporan akhir skripsi. Rangkaian kegiatan penelitian diawali dengan studi literatur dan pengumpulan dataset Flickr8k, yang kemudian dilanjutkan dengan tahap pra-pemrosesan data. Setelah data siap, fokus penelitian beralih pada perancangan arsitektur dan pelatihan model CNN VGG16 dan LSTM. Tahap akhir dari waktu penelitian ini dialokasikan untuk implementasi antarmuka sistem, pengujian evaluasi model menggunakan metrik *BLEU Score*, serta analisis hasil dan penyusunan laporan akhir.



BAB IV

HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil penelitian yang diperoleh dari proses penerapan model *Convolutional Neural Network* dan *Long Short-Term Memory* pada sistem deskripsi citra otomatis. Hasil penelitian dijelaskan mulai dari keluaran tahap persiapan data, implementasi arsitektur model, proses pelatihan, hingga pengujian model dalam menghasilkan deskripsi citra. Pembahasan pada bab ini tidak hanya menampilkan hasil secara numerik, tetapi juga menguraikan makna dari hasil yang diperoleh berdasarkan kemampuan model dalam mengenali objek visual dan membentuk kalimat deskripsi.

Penelitian ini menggunakan dataset Flickr8k sebagai sumber data utama. Citra pada dataset diproses menggunakan model VGG16 untuk memperoleh fitur visual, sedangkan data teks berupa *caption* digunakan sebagai masukan bagi model LSTM dalam mempelajari pola urutan kata. Setelah proses pelatihan selesai, model diuji menggunakan evaluasi kuantitatif melalui *BLEU Score* dan evaluasi kualitatif melalui pengamatan terhadap kesesuaian *caption* yang dihasilkan dengan isi citra.

4.1 Gambaran Umum Hasil Penelitian

Penelitian ini menghasilkan sebuah model *image captioning* yang dirancang untuk menerima citra sebagai masukan dan menghasilkan deskripsi teks secara otomatis. Secara umum, sistem bekerja melalui beberapa tahapan utama, yaitu pemrosesan citra, ekstraksi fitur visual, pemrosesan teks *caption*, pelatihan model CNN VGG16 dan LSTM, pembangkitan *caption*, serta penerjemahan hasil deskripsi ke dalam bahasa Indonesia. Tahapan tersebut saling berkaitan karena kualitas *caption* yang dihasilkan sangat dipengaruhi oleh kualitas fitur citra dan pola bahasa yang dipelajari model selama proses pelatihan.

Model yang dikembangkan menggunakan VGG16 sebagai bagian *encoder* untuk mengambil representasi visual dari citra. Fitur visual tersebut kemudian dikombinasikan dengan representasi teks pada bagian *decoder* berbasis LSTM. Melalui mekanisme tersebut, model berusaha mempelajari hubungan antara objek atau aktivitas dalam gambar dengan susunan kata yang sesuai. Dengan demikian,

keluaran utama dari penelitian ini adalah *caption* berbahasa Inggris yang menggambarkan isi citra, kemudian dilengkapi dengan hasil terjemahan agar lebih mudah dipahami oleh pengguna berbahasa Indonesia.

Hasil penelitian pada bab ini disajikan secara bertahap agar proses pengembangan model dapat dipahami dengan jelas. Pembahasan dimulai dari hasil persiapan dan pengolahan dataset, dilanjutkan dengan hasil implementasi arsitektur model, hasil pelatihan, hasil pengujian menggunakan BLEU Score, serta analisis kualitatif terhadap beberapa contoh citra uji. Urutan pembahasan tersebut digunakan untuk menunjukkan hubungan antara tahapan pengolahan data, proses pembelajaran model, dan kualitas *caption* yang dihasilkan.

4.2 Hasil Pengolahan Dataset

Dataset yang digunakan dalam penelitian ini adalah Flickr8k yang terdiri dari 8.000 citra dan 40.000 *caption* berbahasa Inggris. Setiap citra memiliki lima *caption* referensi yang menggambarkan isi gambar dari sudut pandang berbeda. Dataset ini digunakan sebagai data utama untuk melatih, memvalidasi, dan menguji model *image captioning* berbasis CNN VGG16 dan LSTM.

Pada tahap pengolahan dataset, data dibagi menjadi tiga bagian, yaitu data latih, data validasi, dan data uji. Data latih digunakan untuk melatih model dalam mempelajari hubungan antara fitur visual citra dan urutan kata pada *caption*. Data validasi digunakan untuk memantau performa model selama proses pelatihan, sedangkan data uji digunakan untuk mengevaluasi kemampuan akhir model terhadap citra yang belum pernah digunakan pada proses pelatihan. Ringkasan pembagian dataset ditampilkan pada Tabel 4.1.

Tabel 4.1 Pembagian Dataset Flickr8k

Jenis Data	Jumlah Citra	Jumlah <i>Caption</i>
Data Latih	6.000 Citra	30.000 <i>Caption</i>
Data Validasi	1.000 Citra	5.000 <i>Caption</i>
Data Uji	1.000 Citra	5.000 <i>Caption</i>

Berdasarkan Tabel 4.1, jumlah *caption* pada setiap bagian dataset diperoleh dari jumlah citra dikalikan lima, karena setiap citra memiliki lima *caption* referensi. Data latih memiliki jumlah terbesar karena digunakan sebagai data utama dalam proses pembelajaran model, sedangkan data validasi dan data uji digunakan untuk memantau serta mengevaluasi performa model pada data yang berbeda.

Selain pembagian dataset, pemeriksaan ukuran awal citra juga dilakukan sebelum citra dimasukkan ke dalam model. Berdasarkan hasil pemeriksaan menggunakan program *Python*, ukuran awal citra pada dataset Flickr8k tidak seragam. Dari 8.000 citra yang diperiksa, terdapat 596 variasi ukuran citra. Lebar citra berada pada rentang 164 hingga 500 piksel, sedangkan tinggi citra berada pada rentang 127 hingga 500 piksel. Beberapa ukuran awal citra yang paling banyak muncul ditampilkan pada Tabel 4.2.

Tabel 4.2 Hasil Pemeriksaan Ukuran Awal Citra

No	Ukuran Awal Citra	Jumlah Citra
1	500 × 333 piksel	1.499 citra
2	500 × 375 piksel	1.285 citra
3	333 × 500 piksel	648 citra
4	375 × 500 piksel	422 citra
5	500 × 334 piksel	342 citra
6	500 × 332 piksel	243 citra
7	334 × 500 piksel	169 citra
8	500 × 335 piksel	121 citra
9	500 × 400 piksel	115 citra
10	332 × 500 piksel	110 citra

Berdasarkan Tabel 4.2, citra pada dataset Flickr8k memiliki ukuran awal yang bervariasi. Oleh karena itu, seluruh citra disesuaikan menjadi ukuran 224 × 224 piksel agar sesuai dengan kebutuhan input model VGG16. Setelah ukuran citra diseragamkan, citra dikonversi ke dalam bentuk array numerik dan diproses menggunakan fungsi *preprocessing* VGG16. Hasil dari tahap ini adalah fitur visual dalam bentuk vektor numerik yang digunakan sebagai masukan pada model LSTM.

Selain pengolahan citra, *caption* juga diolah agar dapat digunakan oleh model. Pengolahan *caption* dilakukan dengan mengubah seluruh huruf menjadi huruf kecil, menghapus tanda baca, menambahkan token startseq dan endseq, serta mengubah kata menjadi indeks numerik menggunakan *tokenizer*. Setelah itu, urutan

kata disamakan panjangnya menggunakan *padding* agar dapat diproses oleh model LSTM.

4.3 Hasil Pelatihan Model Berdasarkan Epoch

Setelah tahap pengolahan dataset selesai dilakukan, proses selanjutnya adalah pelatihan model CNN VGG16 dan LSTM. Model dilatih untuk mempelajari hubungan antara fitur visual citra yang diekstraksi menggunakan VGG16 dan urutan kata pada *caption*. Pada tahap ini, model menerima dua jenis masukan, yaitu fitur citra dan urutan kata, kemudian memprediksi kata berikutnya hingga membentuk *caption* secara utuh.

Pelatihan model dilakukan dengan konfigurasi sebagaimana ditampilkan pada Tabel 4.3.

Tabel 4.3 Hasil Konfigurasi Pelatihan Model CNN VGG16 dan LSTM

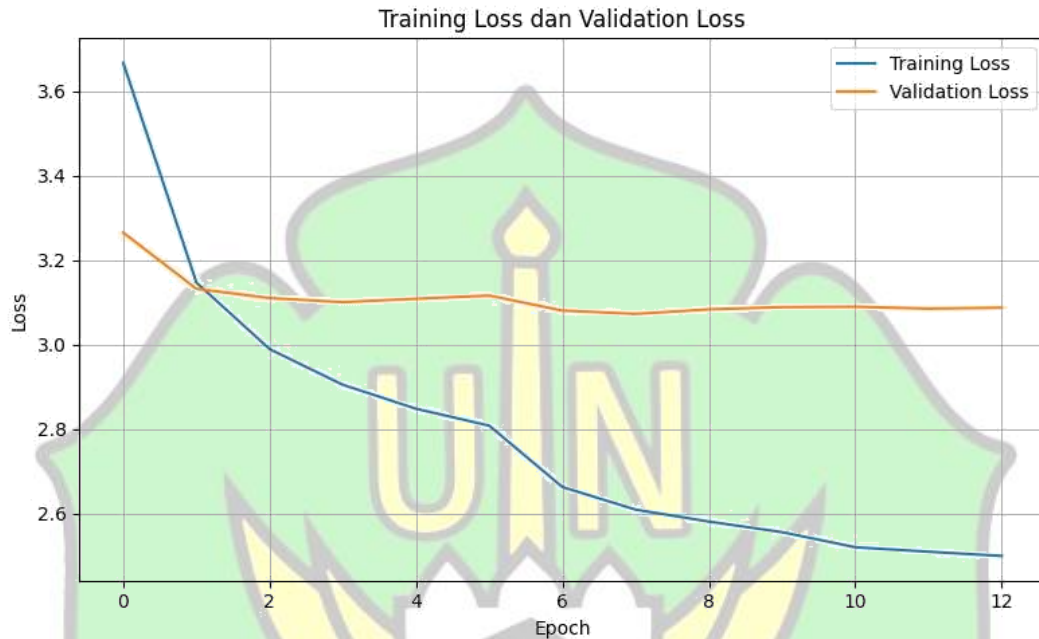
Komponen	Hasil
Arsitektur model	CNN VGG16-LSTM
Optimizer	Adam
Loss function	<i>Sparse Categorical Crossentropy</i>
Epoch maksimum	30 epoch
<i>Learning rate</i> awal	0,001

Berdasarkan Tabel 4.3, model CNN VGG16 DAN LSTM dilatih menggunakan *optimizer Adam* dengan *learning rate* awal sebesar 0,001. Fungsi loss yang digunakan adalah *Sparse Categorical Crossentropy* karena proses pembangkitan *caption* dilakukan dengan memprediksi kata berikutnya berdasarkan kosakata yang tersedia. Jumlah epoch maksimum ditetapkan sebanyak 30 epoch.

Selama proses pelatihan, model tidak berjalan hingga epoch maksimum karena menggunakan mekanisme *early stopping*. Proses pelatihan berhenti pada epoch ke-13 karena nilai *validation loss* tidak menunjukkan perbaikan yang lebih baik setelah beberapa epoch. Model terbaik diperoleh pada epoch ke-8 dengan nilai *validation loss* sebesar 3,0729 dan nilai *validation accuracy* sebesar 0,4002. Setelah pelatihan dihentikan, bobot model dikembalikan ke kondisi terbaik pada epoch ke-

8. Model pada epoch tersebut kemudian digunakan sebagai model akhir untuk evaluasi *BLEU Score* dan pengujian citra luar dataset.

Grafik nilai loss dan *validation loss* selama proses pelatihan ditampilkan di bawah ini.



Berdasarkan Grafik Loss dan Validation Loss Model CNN VGG16 dan LSTM, garis berwarna biru menunjukkan nilai *training loss*, sedangkan garis berwarna orange menunjukkan nilai *validation loss*. Nilai *training loss* mengalami penurunan secara bertahap selama proses pelatihan. Pada epoch awal, nilai *training loss* berada pada kisaran 3,66, kemudian terus menurun hingga mencapai sekitar 2,50 pada epoch ke-13. Penurunan ini menunjukkan bahwa model mampu mempelajari pola hubungan antara fitur visual citra dan urutan kata pada *caption*.

Sementara itu, nilai *validation loss* mengalami penurunan pada beberapa epoch awal, tetapi setelah mencapai nilai terbaik, nilainya cenderung stabil dan tidak mengalami penurunan yang signifikan. Berdasarkan hasil pelatihan, nilai *validation loss* terbaik diperoleh pada epoch ke-8, yaitu sebesar 3,0729. Setelah epoch tersebut, *validation loss* tidak menunjukkan perbaikan yang lebih baik, sehingga proses pelatihan dihentikan pada epoch ke-13 melalui mekanisme *early stopping*.

Dengan demikian, model yang digunakan untuk tahap evaluasi bukan model pada epoch ke-13, melainkan model terbaik pada epoch ke-8. Pemilihan model berdasarkan nilai *validation loss* terendah dilakukan agar model yang digunakan memiliki kemampuan generalisasi yang lebih baik terhadap data validasi.

4.4 Hasil Evaluasi Model Menggunakan *BLEU Score*

Setelah proses pelatihan selesai dilakukan, model terbaik pada epoch ke-8 digunakan untuk tahap evaluasi menggunakan 1.000 citra data uji Flickr8k. Pada tahap ini, model menghasilkan *caption* berbahasa Inggris untuk setiap citra uji, kemudian hasil tersebut dibandingkan dengan *caption* referensi yang tersedia pada dataset.

Caption dihasilkan menggunakan metode *Beam Search*. *Beam Search* adalah metode pencarian urutan kata yang mempertahankan beberapa kandidat *caption* terbaik pada setiap tahap prediksi, bukan hanya memilih satu kata dengan probabilitas tertinggi. Pada penelitian ini, *Beam Search* digunakan agar *caption* yang dihasilkan lebih stabil dan memiliki peluang lebih besar membentuk kalimat yang relevan dengan citra.

Evaluasi dilakukan menggunakan BLEU-1 sampai BLEU-4 karena setiap metrik menilai tingkat kecocokan yang berbeda. BLEU-1 menilai kecocokan kata tunggal, BLEU-2 menilai dua kata berurutan, BLEU-3 menilai tiga kata berurutan, dan BLEU-4 menilai empat kata berurutan. Dengan demikian, evaluasi tidak hanya melihat kesesuaian kata utama, tetapi juga kemampuan model dalam membentuk frasa yang lebih panjang. Hasil evaluasi *BLEU Score* ditampilkan pada Tabel 4.4.

Tabel 4.4 Hasil Evaluasi BLEU Score Model Akhir

Metrik Evaluasi	Nilai
BLEU-1	0,5627
BLEU-2	0,3611
BLEU-3	0,2313
BLEU-4	0,1556

Berdasarkan Tabel 4.4, nilai BLEU-1 memperoleh hasil tertinggi, yaitu sebesar 0,5627. Nilai ini menunjukkan bahwa model mampu menghasilkan kata-kata utama yang sesuai dengan *caption* referensi, seperti objek, aktivitas, atau atribut visual sederhana pada citra.

Nilai BLEU-2, BLEU-3, dan BLEU-4 mengalami penurunan secara bertahap karena semakin tinggi tingkat n-gram yang digunakan, semakin kompleks susunan kata yang harus sesuai dengan *caption* referensi. Oleh karena itu, nilai BLEU-3 dan BLEU-4 yang lebih rendah menunjukkan bahwa model masih memiliki keterbatasan dalam menghasilkan frasa panjang yang sama dengan deskripsi manusia.

Secara umum, hasil BLEU Score menunjukkan bahwa model CNN VGG16 dan LSTM dengan VGG16 dan *Beam Search* mampu menghasilkan *caption* yang sesuai pada tingkat kata utama, tetapi masih terbatas dalam membentuk struktur frasa yang lebih panjang dan kompleks.

4.5 Hasil Pengujian pada Citra Luar Dataset

Selain pengujian menggunakan data uji dari dataset Flickr8k, model juga diuji menggunakan citra luar dataset. Pengujian ini dilakukan untuk melihat kemampuan model dalam menghasilkan *caption* pada citra yang tidak termasuk dalam data latih, data validasi, maupun data uji. Citra luar dataset digunakan sebagai bentuk pengujian kualitatif untuk menilai kesesuaian *caption* yang dihasilkan dengan objek, aktivitas, dan konteks visual pada gambar.

Pada tahap ini, citra masukan diproses terlebih dahulu agar sesuai dengan kebutuhan model VGG16, yaitu diubah ke ukuran 224×224 piksel. Setelah itu, fitur visual citra diekstraksi menggunakan VGG16 dan digunakan oleh model LSTM untuk menghasilkan *caption* berbahasa Inggris. *Caption* yang dihasilkan kemudian diterjemahkan ke dalam bahasa Indonesia agar lebih mudah dipahami oleh pengguna.

Penjelasan lebih rinci terhadap hasil pengujian pada masing-masing citra luar dataset disajikan setelah Tabel 4.5. Hal ini dilakukan agar analisis tidak hanya

terbatas pada ringkasan di dalam tabel, tetapi juga mencakup kesesuaian objek, aktivitas, atribut visual, dan latar pada setiap citra yang diuji.

Hasil pengujian model pada citra luar dataset ditampilkan pada Tabel 4.5.

Tabel 4.5 Hasil Pengujian pada Citra Luar Dataset

No	Citra Uji	Sumber	Caption Bahasa Inggris	Terjemahan Bahasa Indonesia	Analisis Hasil
1		(PNGTree , 2022)	a man in a red shirt be stand in front of a building	seorang pria berbaju merah berdiri di depan sebuah gedung	Sebagian sesuai, tetapi salah mengenali subjek dan latar.
2		(Clear Conscience Pet, 2024)	a black and white dog be run through the grass	seekor anjing hitam putih berlari melewati rerumputan	Objek anjing dikenali, tetapi aktivitas salah.
3		(SWA, 2023)	a man be ride a bike on a track	seorang pria mengendarai sepeda di lintasan	Caption sesuai dengan objek dan aktivitas pada citra.
4		(Blue River OK, 2026)	a man be stand on a snowy mountain	seorang pria berdiri di atas gunung bersalju	Objek dan konteks gunung cukup sesuai, tetapi atribut salju kurang tepat.

5		(Kumparan Bisnis, 2023)	a group of person stand in front of a building	sekelompok orang berdiri di depan sebuah gedung	Objek manusia dikenali, tetapi konteks tempat salah.
---	---	-------------------------	--	---	--

Berdasarkan hasil pengujian pada Tabel 4.5, model menunjukkan kemampuan yang berbeda pada setiap citra luar dataset. Pada citra pertama, model mampu mengenali atribut warna merah dan posisi objek yang sedang berdiri. Namun, model masih salah mengenali subjek sebagai pria, sedangkan citra menampilkan anak kecil. Selain itu, model juga keliru membaca latar sebagai bangunan, padahal citra berada pada area pepohonan.

Pada citra kedua, model berhasil mengenali objek utama berupa anjing dan keberadaan area rumput. Akan tetapi, model belum mampu mengenali aktivitas secara tepat karena *caption* yang dihasilkan menyebutkan anjing sedang berlari, sedangkan pada citra sebenarnya anjing terlihat sedang minum dari mangkuk. Hal ini menunjukkan bahwa model mampu baik dalam mengenali objek utama, tetapi masih terbatas dalam membedakan aktivitas yang lebih spesifik.

Pada citra ketiga, *caption* yang dihasilkan sudah sesuai dengan isi citra. Model berhasil mengenali objek manusia, sepeda, serta aktivitas bersepeda. Hasil ini menunjukkan bahwa model dapat menghasilkan *caption* yang relevan ketika citra memiliki objek utama yang jelas dan aktivitas yang umum ditemukan pada dataset.

Pada citra keempat, model mampu mengenali keberadaan manusia dan konteks pegunungan. Namun, model menambahkan atribut “snowy” atau bersalju, padahal citra tidak memperlihatkan salju secara jelas. Kesalahan ini menunjukkan bahwa model masih dapat menghasilkan atribut visual yang kurang tepat meskipun makna umum *caption* masih mendekati isi citra.

Pada citra kelima, model berhasil mengenali adanya sekelompok orang. Namun, model keliru dalam mengenali konteks tempat karena menghasilkan

deskripsi “di depan sebuah gedung”, sedangkan citra menunjukkan suasana di dalam transportasi umum. Secara umum, hasil pengujian pada citra luar dataset menunjukkan bahwa model mampu mengenali objek utama, tetapi masih memiliki keterbatasan dalam memahami aktivitas, atribut visual, dan latar yang lebih kompleks.

Dari aspek kelancaran tata bahasa, *caption* yang dihasilkan model masih dapat dipahami, meskipun sebagian kalimat memiliki struktur sederhana dan belum sepenuhnya natural seperti deskripsi manusia. Kesalahan utama model lebih banyak terjadi pada pemahaman objek, aktivitas, atribut visual, dan konteks latar, bukan pada keterbacaan kalimat secara umum.

Secara keseluruhan, hasil pengujian pada citra luar dataset menunjukkan bahwa model mampu menghasilkan *caption* yang baik pada citra dengan objek umum dan aktivitas yang jelas, seperti citra pesepeda. Namun, pada beberapa citra lain model masih mengalami kesalahan dalam mengenali subjek, aktivitas, atribut visual, dan konteks latar. Kesalahan tersebut terlihat pada citra anak kecil yang dikenali sebagai pria, anjing yang sedang minum dikenali sedang berlari, serta suasana dalam transportasi umum dikenali sebagai depan gedung. Hal ini menunjukkan bahwa model mampu mengenali objek utama, tetapi masih memiliki keterbatasan dalam memahami konteks visual yang lebih kompleks.

4.6 Tampilan Sistem dan Hasil Implementasi

Setelah model selesai dilatih dan diuji, tahap selanjutnya adalah mengimplementasikan model ke dalam bentuk prototipe sistem sederhana. Sistem ini dibuat agar pengguna dapat memasukkan citra dan memperoleh hasil deskripsi secara otomatis. Pada tahap implementasi, model terbaik yang diperoleh dari proses pelatihan digunakan untuk menghasilkan *caption* berdasarkan citra masukan.

Prototipe sistem dijalankan melalui Google Colab dengan bantuan library Gradio sebagai antarmuka pengguna. Gradio digunakan untuk menyediakan fitur unggah citra, menampilkan hasil *caption* bahasa Inggris, serta menampilkan hasil terjemahan bahasa Indonesia. Tautan publik yang dihasilkan oleh Gradio bersifat

sementara selama sesi Colab berjalan, sehingga tidak dijadikan sebagai komponen utama dalam evaluasi sistem.

Sistem menerima citra dari pengguna, kemudian citra tersebut diproses agar sesuai dengan kebutuhan input model VGG16. Setelah itu, fitur visual citra diekstraksi menggunakan VGG16 dan digunakan oleh model LSTM untuk menghasilkan *caption* berbahasa Inggris. *Caption* yang dihasilkan kemudian diterjemahkan ke dalam bahasa Indonesia sehingga lebih mudah dipahami oleh pengguna.

Tampilan sistem hasil implementasi ditampilkan pada Gambar 4.2.



Gambar 4.2 Tampilan Sistem *Image Captioning* CNN VGG16 dan LSTM dengan Beam Search.

Berdasarkan Gambar 4.2, sistem menampilkan antarmuka sederhana untuk proses deskripsi citra otomatis. Pengguna dapat mengunggah citra pada bagian kiri, kemudian sistem memproses citra tersebut menggunakan model CNN VGG16 dan LSTM yang telah dilatih. Hasil proses ditampilkan pada bagian kanan dalam bentuk *caption* bahasa Inggris dan terjemahan bahasa Indonesia.

Pada contoh pengujian, sistem menghasilkan *caption* “a man be ride a bike on a track” dan terjemahan “seorang pria mengendarai sepeda di lintasan”. *Caption* tersebut sesuai dengan citra masukan yang menampilkan seseorang sedang mengendarai sepeda. Sistem juga menampilkan parameter *Beam Width* 5 dan *Beam Alpha* 0,7. *Beam Width* 5 berarti model mempertahankan lima kandidat *caption* terbaik pada setiap tahap prediksi, sedangkan *Beam Alpha* 0,7 digunakan sebagai pengatur panjang *caption* agar hasil kalimat tidak terlalu pendek atau terlalu panjang. Kedua parameter tersebut digunakan untuk membantu proses *Beam Search* menghasilkan *caption* yang lebih stabil dan relevan terhadap citra.

4.7 Pembahasan

Berdasarkan hasil penelitian yang telah dilakukan, model CNN VGG16 dan LSTM yang dibangun mampu menghasilkan deskripsi citra secara otomatis. Model VGG16 berperan sebagai *encoder* untuk mengekstraksi fitur visual dari citra, sedangkan LSTM berperan sebagai *decoder* untuk membentuk urutan kata menjadi *caption*. Kombinasi kedua arsitektur tersebut memungkinkan sistem untuk menghubungkan informasi visual pada gambar dengan bentuk deskripsi teks.

Pada tahap pengolahan dataset, citra pada dataset Flickr8k memiliki ukuran awal yang bervariasi. Oleh karena itu, seluruh citra disesuaikan menjadi ukuran 224×224 piksel agar dapat diproses oleh VGG16. *Caption* juga diolah dengan mengubah teks menjadi format yang dapat dipahami oleh model, seperti penambahan token startseq dan endseq, tokenisasi, serta *padding*. Tahapan ini penting karena model CNN VGG16 dan LSTM membutuhkan data citra dan data teks dalam bentuk numerik agar dapat diproses selama pelatihan.

Hasil pelatihan menunjukkan bahwa model mengalami proses pembelajaran yang mampu baik. Hal ini terlihat dari nilai *training loss* yang terus menurun selama proses pelatihan. Model terbaik diperoleh pada epoch ke-8 dengan nilai *validation loss* terendah sebesar 3,0729. Setelah epoch tersebut, nilai *validation loss* tidak menunjukkan perbaikan yang lebih baik, sehingga proses pelatihan dihentikan pada epoch ke-13 menggunakan mekanisme *early stopping*. Kondisi ini menunjukkan bahwa model terbaik tidak selalu berada pada epoch terakhir, melainkan pada epoch dengan nilai *validation loss* paling rendah.

Hasil evaluasi menggunakan *BLEU Score* menunjukkan bahwa model memperoleh nilai BLEU-1 yang lebih tinggi dibandingkan BLEU-2, BLEU-3, dan BLEU-4. Nilai BLEU-1 sebesar 0,5627 menunjukkan bahwa model mampu menghasilkan kata-kata utama yang sesuai dengan *caption* referensi. Namun, nilai BLEU-3 dan BLEU-4 yang lebih rendah menunjukkan bahwa model masih memiliki keterbatasan dalam menghasilkan susunan frasa panjang yang sama dengan *caption* manusia. Hal ini wajar terjadi karena proses *image captioning* tidak

hanya membutuhkan kemampuan mengenali objek, tetapi juga kemampuan menyusun hubungan antar objek dan aktivitas dalam struktur kalimat yang tepat.

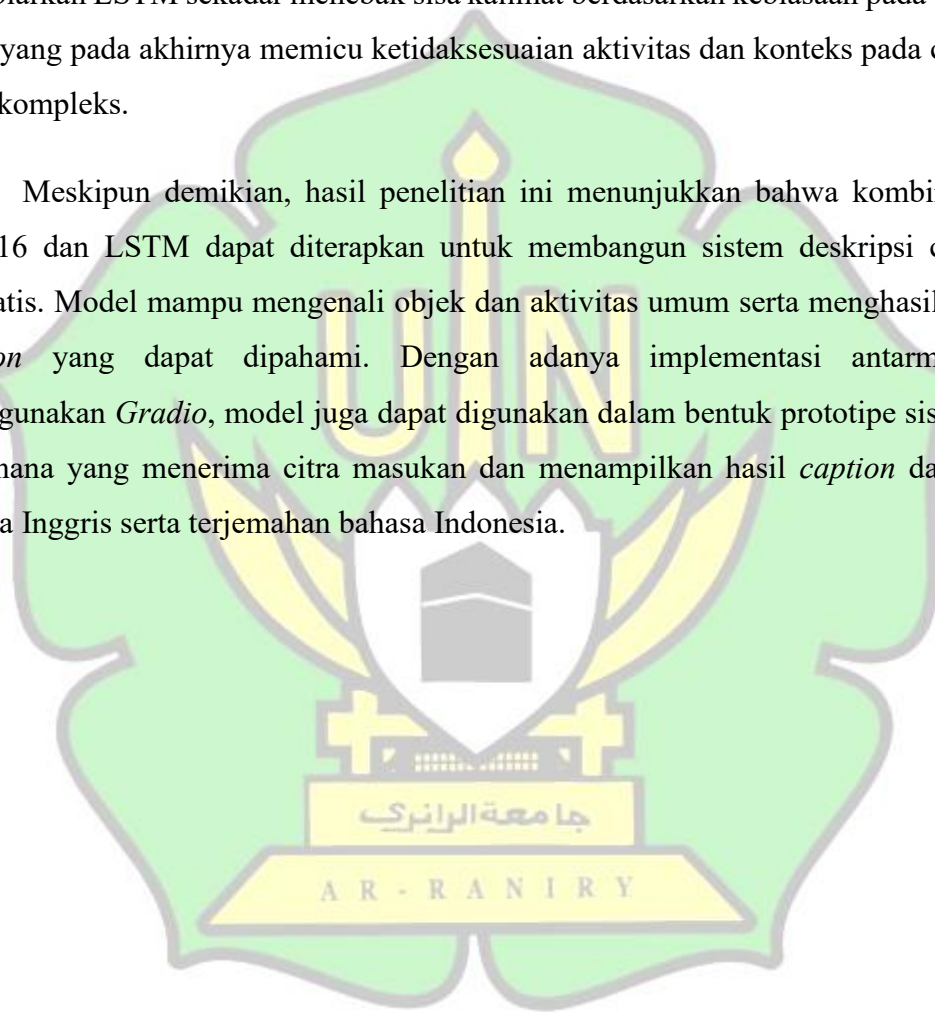
Pada pengujian citra luar dataset, model mampu menghasilkan *caption* yang sesuai pada gambar dengan objek umum dan aktivitas yang jelas, seperti citra pesepeda. Namun, model masih mengalami beberapa kesalahan pada citra yang memiliki konteks lebih kompleks. Kesalahan tersebut meliputi kesalahan pengenalan subjek, kesalahan aktivitas, kesalahan atribut visual, dan kesalahan latar. Contohnya, model mengenali anak kecil sebagai pria, anjing yang sedang minum sebagai anjing yang sedang berlari, serta suasana di dalam transportasi umum sebagai di depan gedung. Dari aspek kelancaran tata bahasa, *caption* yang dihasilkan model masih dapat dipahami, meskipun sebagian kalimat memiliki struktur sederhana dan belum sepenuhnya natural seperti deskripsi manusia. Dengan demikian, kesalahan utama model lebih banyak terjadi pada pemahaman objek, aktivitas, atribut visual, dan konteks latar, bukan pada keterbacaan kalimat secara umum.

Ketidakakuratan model dalam mendeskripsikan citra uji di luar dataset (seperti kesalahan pengenalan subjek anak kecil menjadi pria, atau aktivitas anjing minum menjadi anjing berlari) dapat dijelaskan melalui dua batasan fundamental dari arsitektur Encoder-Decoder dasar.

Pertama, terjadi fenomena bias probabilitas bahasa (dataset bias). Pada tahap *decoding*, model LSTM cenderung bertumpu pada probabilitas statistik kemunculan pasangan kata (*co-occurrence*) yang paling sering ia pelajari di dataset Flickr8k. Sebagai contoh, karena gambar "anjing" di area "rumput" pada data latih mayoritas dideskripsikan sedang "berlari", language model akan memprioritaskan prediksi kata tersebut secara membabi buta, meskipun informasi visual citra aslinya menunjukkan anjing yang sedang minum. Model rentan mengalami "halusinasi teks" (menambahkan atribut salju atau gedung) karena kata-kata tersebut sering muncul berbarengan dengan kata gunung atau pria pada data latih.

Kedua, ketiadaan Mekanisme Atensi (*Attention Mechanism*) memperlebar kesenjangan semantik (*semantic gap*). Arsitektur VGG16 pada sistem ini memampatkan seluruh informasi visual gambar menjadi satu vektor fitur global yang statis. Akibatnya, model kehilangan informasi spasial mengenai relasi antar objek. Tanpa *Attention*, model tidak mampu memusatkan pandangan pada detail lokal yang krusial (seperti interaksi anjing dengan mangkuk air). Hal ini menyebabkan model hanya mengenali objek utama secara garis besar, lalu membiarkan LSTM sekadar menebak sisa kalimat berdasarkan kebiasaan pada data latih, yang pada akhirnya memicu ketidaksesuaian aktivitas dan konteks pada citra yang kompleks.

Meskipun demikian, hasil penelitian ini menunjukkan bahwa kombinasi VGG16 dan LSTM dapat diterapkan untuk membangun sistem deskripsi citra otomatis. Model mampu mengenali objek dan aktivitas umum serta menghasilkan *caption* yang dapat dipahami. Dengan adanya implementasi antarmuka menggunakan *Gradio*, model juga dapat digunakan dalam bentuk prototipe sistem sederhana yang menerima citra masukan dan menampilkan hasil *caption* dalam bahasa Inggris serta terjemahan bahasa Indonesia.



BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, maka kesimpulan yang diperoleh adalah sebagai berikut:

1. Model *image captioning* berbasis CNN VGG16 dan LSTM berhasil dirancang dan diimplementasikan untuk menghasilkan deskripsi citra secara otomatis. VGG16 digunakan sebagai *encoder* untuk mengekstraksi fitur visual dari citra, sedangkan LSTM digunakan sebagai *decoder* untuk membentuk urutan kata menjadi *caption*. Sistem juga dilengkapi dengan fitur terjemahan ke bahasa Indonesia agar hasil deskripsi lebih mudah dipahami oleh pengguna.
2. Hasil evaluasi model pada dataset Flickr8k menggunakan *BLEU Score* menunjukkan bahwa model memperoleh nilai BLEU-1 sebesar 0,5627, BLEU-2 sebesar 0,3611, BLEU-3 sebesar 0,2313, dan BLEU-4 sebesar 0,1556. Nilai BLEU-1 yang lebih tinggi menunjukkan bahwa model mampu mengenali kata-kata utama pada *caption*, sedangkan nilai BLEU-3 dan BLEU-4 yang lebih rendah menunjukkan bahwa model masih memiliki keterbatasan dalam membentuk frasa panjang dan struktur kalimat yang kompleks.
3. Hasil pengujian pada citra dunia nyata di luar dataset menunjukkan bahwa model CNN VGG16 dan LSTM mampu menghasilkan *caption* yang sesuai pada citra dengan objek umum dan aktivitas yang jelas. Dari aspek pengenalan objek visual, model mampu mengenali objek utama seperti manusia, anjing, sepeda, dan konteks umum seperti rumput atau pegunungan. Dari aspek kelancaran tata bahasa, *caption* yang dihasilkan masih dapat dipahami, meskipun struktur kalimatnya sederhana dan belum sepenuhnya natural seperti deskripsi manusia. Sementara itu, dari aspek relevansi aktivitas, model masih memiliki keterbatasan karena beberapa aktivitas tidak dikenali secara tepat, seperti anjing yang sedang minum dikenali sebagai anjing yang sedang berlari, serta konteks di dalam

transportasi umum dikenali sebagai di depan gedung. Dengan demikian, model mampu andal untuk menghasilkan deskripsi umum pada citra luar dataset, tetapi belum sepenuhnya akurat dalam memahami aktivitas spesifik, atribut visual, dan konteks latar yang lebih kompleks.

5.2 Saran

Berdasarkan hasil penelitian dan keterbatasan yang ditemukan, terdapat beberapa saran untuk pengembangan penelitian selanjutnya, yaitu:

1. Penelitian selanjutnya dapat menggunakan dataset yang lebih besar dan beragam, seperti Flickr30k atau MS COCO, agar model dapat mempelajari lebih banyak variasi objek, aktivitas, atribut, dan konteks latar.
2. Pengembangan model dapat dilakukan dengan menggunakan arsitektur *encoder* yang lebih modern, seperti *ResNet*, *EfficientNet*, atau *Vision Transformer*, sehingga fitur visual yang dihasilkan lebih kuat dibandingkan VGG16.
3. Penelitian selanjutnya dapat menambahkan mekanisme *attention* agar model lebih fokus pada bagian penting dari citra saat menghasilkan *caption*, sehingga kesalahan dalam mengenali objek, aktivitas, dan atribut visual dapat dikurangi.
4. Sistem dapat dikembangkan agar mampu menghasilkan *caption* bahasa Indonesia secara langsung tanpa bergantung pada proses penerjemahan otomatis, sehingga hasil deskripsi lebih sesuai dengan struktur bahasa Indonesia.
5. Prototipe sistem dapat dikembangkan menjadi aplikasi web atau mobile yang lebih stabil dan interaktif, serta dapat ditambahkan fitur seperti penyimpanan riwayat hasil *caption* dan *text-to-speech* untuk mendukung kebutuhan aksesibilitas.

DAFTAR PUSTAKA

- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaria, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021a). Review of *deep learning*: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1). <https://doi.org/10.1186/s40537-021-00444-8>
- Jean, J., Sengka, V., Lantang, O. A., Dwisnanto Putro, M., Sam, U., Manado, R., & Kampus Bahu, J. (2025). *Aplikasi Image Captioning Menggunakan Convolutional Neural Network dan Long Short Term Memory* (Vol. 10, Issue 02).
- Kaur, M., & Kaur, H. (2025). An Efficient CNN VGG16 DAN LSTM Based Framework for Improved *Image Captioning*. *Procedia Computer Science*, 258, 3601–3607. <https://doi.org/10.1016/j.procs.2025.04.615>
- Sai, A. B., Mohankumar, A. K., & Khapra, M. M. (2023). A Survey of Evaluation Metrics Used for NLG Systems. In *ACM Computing Surveys* (Vol. 55, Issue 2). Association for Computing Machinery. <https://doi.org/10.1145/3485766>
- Sarker, I. H. (2021a). *Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions*. In *SN Computer Science* (Vol. 2, Issue 6). Springer. <https://doi.org/10.1007/s42979-021-00815-1>
- Sarker, I. H. (2021b). *Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions*. In *SN Computer Science* (Vol. 2, Issue 6). Springer. <https://doi.org/10.1007/s42979-021-00815-1>
- Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2021a). *From Show to Tell: A Survey on Deep Learning-based Image Captioning*. <http://arxiv.org/abs/2107.06912>
- Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2021b). *From Show to Tell: A Survey on Deep Learning-based Image Captioning*. <http://arxiv.org/abs/2107.06912>
- Zakir Hossain, M. D., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of *deep learning* for *image captioning*. In *ACM Computing Surveys* (Vol. 51, Issue 6). Association for Computing Machinery. <https://doi.org/10.1145/3295748>

Teng, M., Zhang, T., & Luo, J. (2022). Fusion of multi-scale local cues and hierarchical attention-based LSTM for stock price trend prediction. *Neurocomputing*, 479, 155–167. <https://doi.org/10.1016/j.neucom.2022.01.077>

PNGTree. (2022). Girl in Red Fashion Girl Cold Photo. Diakses pada Juni 2026, dari https://id.pngtree.com/freebackground/girl-in-red-fashion-girl-cold-photo_7210358.html

Clear Conscience Pet. (2024). Is Your Dog a Picky Drinker? Diakses pada Juni 2026, dari <https://clearconsciencepet.com/blogs/news/is-your-dog-a-picky-drinker>

SWA. (2023). Program Move for Coffee Cause ID Jangkau ke Penjuru Nusantara. Diakses pada Juni 2026, dari <https://swa.co.id>

Blue River OK. (n.d.). Pendakian Gunung Populer untuk Petualang Pemula. Diakses pada Juni 2026, dari <https://blueriverok.com/pendakian-gunung-populer-untuk-petualang-pemula/>

Kumparan Bisnis. (2023). Jumlah Penumpang KRL Diprediksi Membeludak, Tembus 410 Juta Orang hingga 2027. Diakses pada Juni 2026, dari <https://m.kumparan.com/kumparanbisnis/jumlah-penumpang-krl-diprediksi-membeludak-tembus-410-juta-orang-hingga-2027-20mE7Vb3sGy>

