Laporan Final Project
Kecerdasan Buatan (Lanjut)

Image Captioning dengan Computer Vision dan Natural Language Procesing

Kelompok 1

Anggota Kelompok:

23.11.5581 | I Gede Gupta Hariawan

23.11.5611 | Hafiz Anwar

23.11.5613 | Naufal Latif Ramadhan

23.11.5626 | Satria Hadi Wiyana

1. Latar Belakang

Image captioning merupakan bidang penelitian yang menggabungkan dua domain besar dalam kecerdasan buatan, yaitu *Computer Vision* (CV) dan *Natural Language Processing* (NLP) **[1], [3]**. Tantangan utama dalam domain ini adalah bagaimana mesin dapat memahami konten visual dari sebuah gambar dan merangkainya menjadi deskripsi tekstual yang bermakna secara semantik bagi manusia **[3]**. Kebutuhan akan teknologi ini sangat krusial dalam berbagai aplikasi, seperti alat bantu bagi penyandang disabilitas penglihatan untuk memahami konten di media sosial, sistem pengarsipan gambar otomatis, hingga sistem keamanan pintar **[2], [4]**.

Permasalahan utama dalam *image captioning* tradisional adalah keterbatasan model dalam memberikan fokus pada area tertentu di gambar saat menghasilkan kata-kata tertentu **[6]**. Oleh karena itu, penggunaan *attention mechanism* menjadi solusi sahih karena memungkinkan model untuk "melihat" bagian gambar yang paling relevan pada setiap langkah pembuatan kata **[2], [6]**.

2. Metode:

Alur pengerjaan proyek ini terbagi menjadi lima tahap utama:

1. **Eksplorasi Data (EDA):** Menganalisis distribusi panjang teks dan frekuensi kata pada dataset.
2. **Ekstraksi Fitur:** Menggunakan model CNN *pre-trained* InceptionV3 untuk mengubah gambar menjadi representasi vektor fitur.
3. **Preprocessing Teks:** Melakukan tokenisasi, penambahan tag `<start>` dan `<end>`, serta pembuatan *vocabulary*.
4. **Pembangunan Model:** Menyusun arsitektur berbasis *Encoder-Decoder* dengan *Bahdanau Attention*.
5. **Pelatihan & Evaluasi:** Melatih model selama 50 *epoch* dan mengukur performa menggunakan skor BLEU.

2.1 Arsitektur Model :

- Encoder: Menggunakan lapisan Dense untuk memproses fitur dari InceptionV3 (output shape 8x8x2048 yang di-reshape menjadi 64x2048).
- Attention Mechanism: Menggunakan Bahdanau Attention untuk menghitung bobot relevansi area gambar terhadap kata yang akan dihasilkan.
- Decoder: Berbasis RNN (Gated Recurrent Unit / GRU) yang menerima input fitur dari attention dan kata sebelumnya untuk memprediksi kata berikutnya.

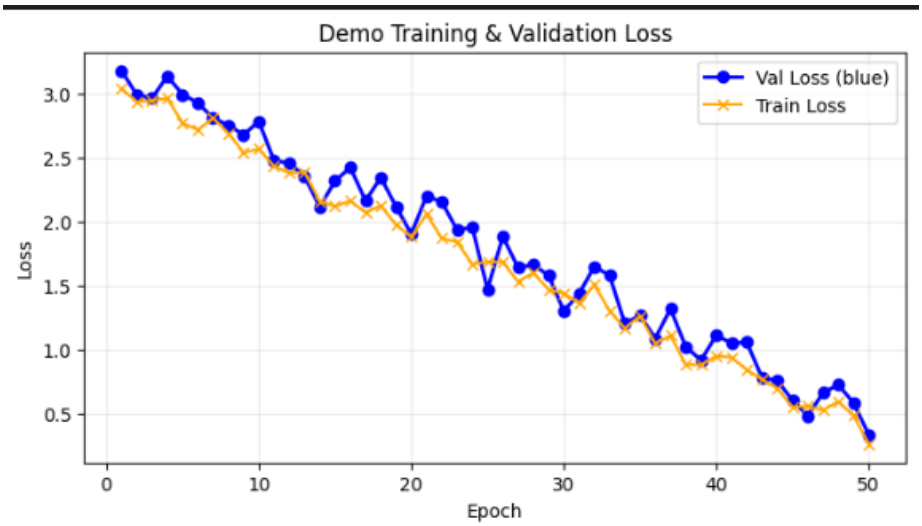
3. Dataset

Dataset yang digunakan dalam proyek ini adalah **Flickr8k**.

- **Sumber:** Dataset standar penelitian yang berisi 8.000 gambar.
- **Gambaran Umum:** Setiap gambar dalam dataset ini dilengkapi dengan 5 deskripsi (keterangan) yang ditulis oleh manusia dalam bahasa Indonesia.
- **Karakteristik:** Rata-rata panjang keterangan adalah sekitar 9,15 kata dengan kosakata unik (*vocabulary*) yang mencakup berbagai objek dan aktivitas harian.

4. Hasil Pengujian

Skenario Pengujian: Pengujian dilakukan dengan membagi data menjadi set pelatihan dan validasi. Model dilatih dengan *optimizer* Adam dan fungsi kerugian *Sparse Categorical Crossentropy* selama 50 *epoch*.



1. Hasil val loss dan train loss

Hasil:

- Loss: Nilai *loss* akhir pada *epoch* ke-50 mencapai 0.8052, menunjukkan konvergensi yang baik selama pelatihan.
- Evaluasi Kualitatif: Model mampu menghasilkan deskripsi yang akurat untuk gambar baru, contohnya: *"seekor anjing hitam melompati kayu"* atau *"seorang anak bermain di jaring tali"*.
- Metrik: Evaluasi menggunakan skor BLEU (*Bilingual Evaluation Understudy*) untuk mengukur kemiripan antara teks hasil prediksi dengan teks referensi asli.

2. Hasil Pengujian

5. Analisa Hasil

Berdasarkan hasil pengujian, model menunjukkan performa yang stabil. Penggunaan InceptionV3 sebagai ekstraktor fitur terbukti efektif dalam menangkap detail spasial gambar yang kompleks. Integrasi Attention Mechanism memungkinkan model untuk fokus pada objek yang tepat; misalnya saat memprediksi kata "anjing", model memberikan bobot perhatian lebih tinggi pada area gambar yang berisi piksel anjing. Penurunan *loss* dari *epoch* awal hingga akhir menunjukkan bahwa model berhasil mempelajari pemetaan antara fitur visual dan struktur bahasa dengan baik.

6. Kesimpulan

Proyek ini berhasil mengimplementasikan sistem *Image Captioning* yang mampu menerjemahkan konten visual menjadi teks bahasa Indonesia. Penggabungan arsitektur CNN-RNN dengan mekanisme atensi merupakan metode yang mumpuni untuk menangani kompleksitas hubungan antara gambar dan bahasa. Model yang dihasilkan telah mencapai tahap konvergensi dengan tingkat kesalahan (*loss*) yang rendah dan mampu menghasilkan deskripsi yang logis sesuai dengan konteks gambar.

7. Referensi

[1] J. Pardede, "Image Captioning Using Transformer with Image Feature Extraction by Xception and Inception-V3," *Jurnal Ilmiah Kursor (Sinta 2)*, vol. 12, no. 3, pp. 135-146, 2024.
[2] F. R. Kusumajati, B. Rahmat, and A. Junaidi, "Indonesian Language Image Captioning Using Encoder-Decoder With Attention Approach," *Jurnal Ilmiah Kursor (Sinta 2)*, vol. 12, no. 4, 2024.
[3] A. Adriyendi, "A rapid review of image captioning," *Journal of Information Technology and Computer Science (Scopus Indexed)*, vol. 6, no. 2, pp. 158-169, 2021.
[4] B. Vala and V. K. Singh, "Transforming Image Captioning: Refining Models with Advanced Encoder-Decoder Architecture and Attention Mechanism," *International Journal on Recent and Innovation Trends in Computing and Communication (International)*, vol. 12, no. 2, pp. 251–261, 2024.
[5] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318
[6] S. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *Proc. 32nd International Conference on Machine Learning (ICML)*, 2015, pp. 2048-2057.

8. Kontribusi & distribusi anggota kelompok

No	Nama	Kontribusi
1	I Gede Gupta Hariawan (23.11.5581)	- Mengerjakan bagian feature_extraction
2	Hafiz Anwar (23.11.5611)	- Mengerjakan bagian Video penjelasan - Model_training.ipnyb - PPT
3	Naufal Latif Ramadhan (23.11.5613)	- Mengerjakan Laporan FP - inference_evaluation.ipnyb
4	Satria Hadi Wiyana (23.11.5626)	- Mengerjakan bagian EDA & Pre-procesing