CNN dan Transformer pada *Image Captioning*

MOHAMMAD FAISHAL DZAKY[1*], JASMAN PARDEDE[1]

[1]Program Studi Informatika, Institut Teknologi Nasional Bandung
Email: mfdzaky7@gmail.com

Received 28 01 2023 | Revised 04 02 2023 | Accepted 04 02 2023

ABSTRAK

Teknologi kecerdasan buatan kini berkembang pesat dan dimanfaatkan secara luas untuk mendukung aktivitas manusia. Penggunaan teknologi kecerdasan buatan untuk pembuatan teks dari suatu gambar adalah salah satunya. Image captioning adalah proses otomatis untuk membuat keterangan gambar. Caption menampilkan teks bahasa alamiah berdasarkan gambar. Image captioning didefinisikan sebagai proses menghasilkan deskripsi tekstual untuk sebuah gambar. Dimulai dengan computer vision yang digunakan untuk mengidentifikasi objek, atribut, dan hubungannya, kemudian Natural Language Processing digunakan untuk memonitor sintaks dan semantik, dan terkahir machine learning digunakan untuk menghasilkan teks. Penelitian dilakukan untuk membangun model yang dapat melakukan image captioning dengan pengekstrak fitur CNN dan membangkitkan kalimatnya menerapkan arstektur Transformer. Dataset yang digunakan adalah Flickr8k yang memiliki total dataset 8000 citra beserta masing-masing captionnya. Penelitian menguji kemampuan model yang dihasilkan dari proses pelatihan menggunakan kedua metode dalam membangkitkan kalimat. Kemampuan model diukur menggunakan skor BLEU. Pengujian menghasilkan nilai BLEU-1 0.718427, BLEU-2 0.608269, BLEU-3 0.563714, dan BLEU-4 0.472956.

***Kata kunci**: Image Captioning, CNN, Transformer*

ABSTRACT

Artificial intelligence technology is now developing rapidly and is widely used to support human activities. The use of artificial intelligence technology to create text from an image is one of them. Captioning is an automated process for creating captions. Captions display natural language captions based on images. Image captioning is defined as the process of generating a textual description for an image. Starting with computer vision which is used to identify objects, attributes, and substitutions, then Natural Language Processing is used to monitor syntax and semantics, and finally machine learning is used to generate text. The research was conducted to build a model that can perform image captioning with the CNN feature extractor and generate the sentence using the Transformer architecture. The dataset used is Flickr8k which has a total dataset of 8000 images and their respective captions. The research tested the ability of the models resulting from the training process to use both methods in generating sentences. The ability of the model is measured using the BLEU score. The test yielded a value of BLEU-1 0.718427, BLEU-2 0.608269, BLEU-3 0.563714, and BLEU-4 0.472956.

***Keywords**: Image Captioning, CNN, Transformer*

1. PENDAHULUAN

Teknologi kecerdasan buatan kini berkembang pesat dan dimanfaatkan secara luas untuk mendukung aktivitas manusia. Penggunaan teknologi kecerdasan buatan untuk pembuatan teks dari suatu gambar adalah salah satunya.

Image Captioning merupakan salah satu topik penelitian yang paling diminati dalam kecerdasan buatan yang berkaitan dengan pemahaman citra dan mendeskripsikan citra tersebut. Kalimat hasil *Image captioning* seharusnya terbentuk dengan baik (Zakir Hossain et al., 2019).

Image captioning adalah salah satu masalah *computer vision* yang telah terlihat memiliki banyak kemajuan (Vinyals Google et al., 2015). Pada kehidupan sehari-hari, mendeskripsikan gambar secara otomatis bisa sangat berguna, contohnya dalam membantu individu penyandang disabilitas untuk memahami isi dari suatu gambar. Namun demikian, menggambarkan suatu gambar adalah pekerjaan yang sangat sulit. Model untuk *image captioning* harus sepenuhnya memahami gambar yang diberikan untuk menangkap interaksi kompleks antara *item* dan mengidentifikasi apa yang menonjol. Selain itu, mereka harus mampu menginterpretasikan reperesentasi citra ke dalam bahasa dan memiliki hubungan antara bahasa dan citra. Terarkhir, frase yang dibuat untuk menyampaikan informasi yang ditangkap harus sealami mungkin (Jiang et al., 2018).

Ali Ashraf Mohamed melakukan penelitian terhadap image captioning berjudul "Image Caption using CNN & LSTM" dimana penelitian menggunakan metode CNN dengan model Xception sebagai Feature Extractor dan LSTM sebagai sequence processor-nya, lalu membandingkannya dengan model CNN VGG16. Penelitian ini menghasilkan BLEU score sebesar 55.01% menggunakan metode CNN dengan model Xception. Di penelitian ini penulis berharap penelitian selanjutnya memanfaatkan arsitektur model lainnya untuk meningkatkan hasil BLEU score nya (Ashraf Mohamed, 2020).

Pemrosesan bahasa alami adalah serangkaian teknik komputasi yang dimotivasi secara teoritis untuk mengevaluasi dan merepresentasikan teks yang terjadi secara alami pada satu atau lebih tingkat analisis linguistik (Joseph et al., 2016). LSTM dan GRU adalah model berbasis RNN yang paling umum digunakan dalam pemodelan bahasa (Dai et al., 2019). Namun, keduanya memiliki masalah dengan *long term dependability.* Untuk mengatasi kelemahan dari pendekatan berbasis *RNN,* model *Transformer* dikembangkan. Teknik *self attention* yang digunakan oleh model *Transformer,* yang tidak lagi menghitung hasil secara berurutan, memungkinkan pengurangan waktu pelatihan (Dowdell & Zhang, 2020).

Dalam hal *long term dependencies*, pemodelan bahasa berbasis *RNN* memiliki keterbatasan yang menyulitkan untuk pemberian keterangan gambar. Berdasarkan uraian tersebut, penelitian akan dilakukan dengan tujuan menggunakan pemodelan bahasa *Transformer* untuk *image captioning*, yang diharapkan dapat mengatasi masalah dengan model arsitektur *RNN*.

Dataset *Flickr8k* yang tersedia di Kaggle.com digunakan pada penelitian ini. Dataset yang digunakan terdiri dari 8000 citra dengan 5 *caption* yang masing-masing memberikan deskripsi tentang citra tersebut. Dataset dibagi menjadi tiga yaitu 6000 dataset latih, 1000 dataset validasi, dan 1000 dataset uji.

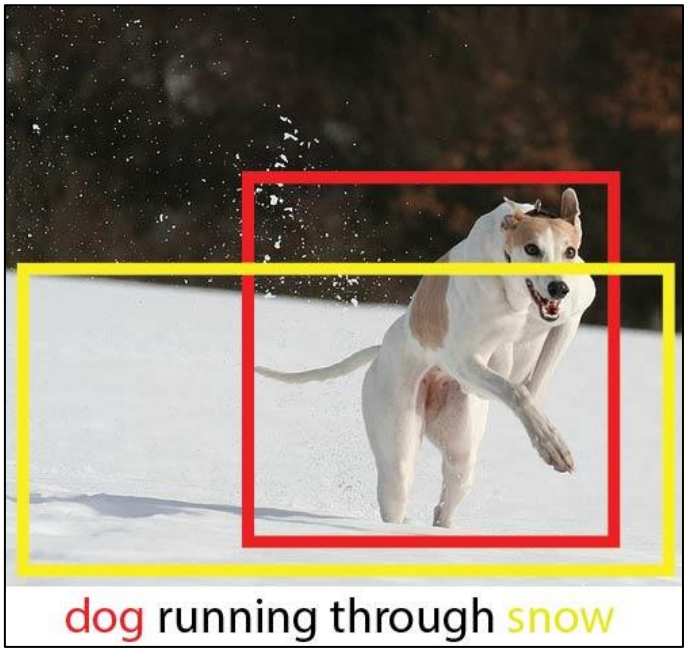
Penelitian ini menerapkan ekstraksi fitur dari EfficientNetB5 dan NLP dari Transformer untuk membangkitkan kalimat dari suatu citra.

2. METODE PENELITIAN

2.1 *Image Captioning*

Image captioning adalah proyek penelitian baru yang populer dalam analisis gambar dan analisis teks. Image captioning adalah proses otomatis untuk membuat keterangan gambar. Caption menampilkan teks bahasa alamiah berdasarkan gambar. Image captioning didefinisikan sebagai proses menghasilkan deskripsi tekstual untuk sebuah gambar. Dimulai dengan computer vision yang digunakan untuk mengidentifikasi objek, atribut, dan hubungannya, kemudian Natural Language Processing digunakan untuk memonitor sintaks dan semantik, dan terkahir machine learning digunakan untuk menghasilkan teks (Adriyendi, 2021).

Ada tiga jenis *image captioning* yang berbeda yaitu teknik berbasis *template*, yang pertama-tama mengidentifikasi ciri-ciri adegan, item, dan *template* kalimat sebelum mengisinya; metode berbasis pemulihan, yang mencari gambar yang mirip satu sama lain; dan yang ketiga adalah metode yang didasarkan pada jaringan syaraf dan mempekerjakan *CNN* untuk mengkasifikasi gambar sebagai *encoder* sebelum menggunakan pemodelan bahasa sebagai *decoder* untuk mentransfer infomasi gambar ke gambar target (Jiang et al., 2018). Gambar 1 merupakan contoh dari citra yang digunakan sebagai citra uji beserta caption yang dihasilkan setelah dilakukan proses image captioning.



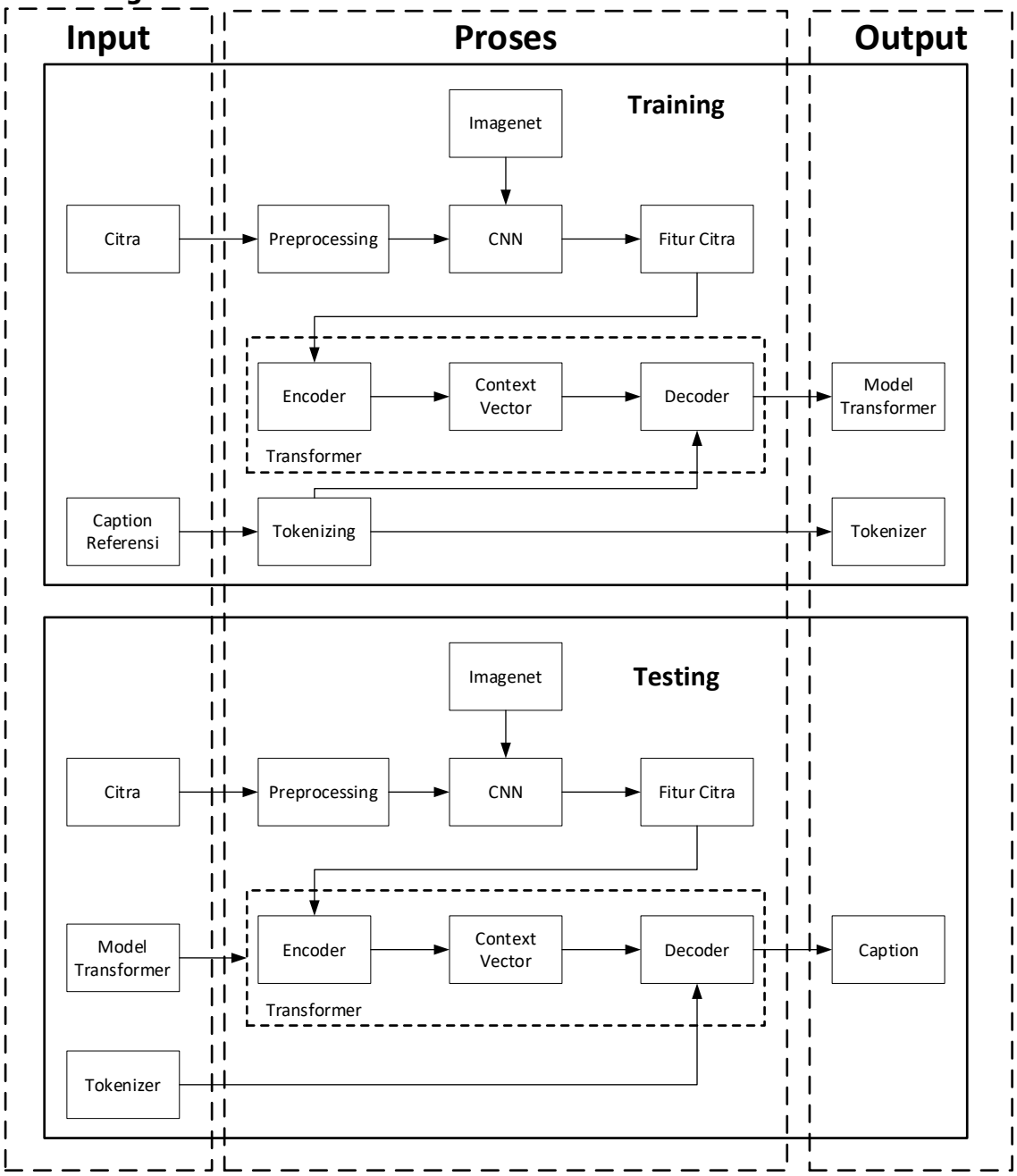
Gambar 1. Image Captioning

2.2 *Transformer*

Transformer adalah konsep arsitektur yang sepenuhnya bergantung pada proses *attention* untuk menarik ketergantungan global antara input dan output. *Transformer* memungkinkan paralelisasi yang lebih besar lagi. *Transformer* pertama kali diperkenalkan oleh Vaswani dan kawan-kawan pada tahun 2017. *Transformer* dapat mencapai tingkat kualitas terjemahan yang baik hanya dalam dua belas jam pada delapan *GPU P100* (Vaswani et al., 2017).

Pada model Transformer, encoder memetakan urutan input representasi simbol (x1, …, xn) ke urutan representasi kontinu z = (z1, …, zn). Kemudian decoder menghasilkan urutan output (y1, …, ym) dari simbol satu elemen pada satu waktu (Vaswani et al., 2017).

2.4 Blok Diagram



Gambar 2. Blok Diagram

Pada Gambar 2 diilustrasikan blok diagram mengenai proses kerja sistem dengan dua tahapan, yaitu training dan testing. Pada tahap training, berikut adalah tahapan-tahapan yang dilakukan:

1. Menyiapkan dataset yang akan digunakan yang terdiri dari dataset citra dan dataset deskripsi yang saling terhubung dengan citranya masing-masing berdasarkan nama citra.

2. Dataset citra di pra proses dengan dua tahapan yaitu 1) mengubah ukuran citra menjadi 299px x 299px dan 3 *channel,* 2) menormalisasi citra sehingga citra memiliki pixel dengan jarak antara -1 hingga 1.

3. Dataset citra yang telah disiapkan kemudian diinputkan ke *CNN* dengan arsitektur *EfficientNetB5* yang menggunakan bobot imagenet untuk melakukan proses ekstraksi fitur yang kemudian menghasilkan fitur citra yang merepresentasikan *object of interest* dari masing-masing citra, kemudian fitur citra di *encode* menggunakan *encoder Transformer* dan menghasilkan *context vector*.

4. Dataset deskripsi yang berupa teks juga dilewatkan melalui proses preprocessing yaitu 1) mengubah semua kalimat menjadi kalimat berhuruf kecil supaya memudahkan perhitungan jumlah kata unik dalam dataset, 2) menambahkan token "<start>" pada awal kalimat dan token "<end>" pada akhir kalimat supaya sistem dapat memulai menghasilkan kalimat saat diberikan token "<start>" dan berhenti menghasilkan kalimat saat diberikan token "<end>", 3) memisahkan kalimat menjadi kata yang individu untuk memberikan kosa kata dari seluruh kata unik dalam dataset, 4) mengubah kata ke dalam indeks kata berurutan untuk menghasilkan vektor yang menunjukkan urutan kata dalam bentuk kalimat, 5) memberikan angka 0 apabila panjang dari vektor urutan kata kutang dari yang ditentukan dan menghapus akhir vektor jika panjang melebihi yang ditentukan, 6) menggeser vektor urutan kata sebanyak satu langkah sehingga model mampu belajar menghasilkan kata berikutnya.

5. Selanjutnya, fitur map dari citra yang sudah di encode dan tokenizer dari hasil preprocessing kalimat deskripsi di decode menggunakan decoder dari Transformer dan disimpan untuk digunakan pada proses testing.

6. Hasil dari decode akan menghasilkan sebuah model.

Pada tahap testing, berikut adalah tahapan-tahapan yang dilakukan:

1. Pengguna menginputkan sebuah citra uji yang diambil dari dataset uji.

2. Citra uji dilewatkan melalui proses preprocessing yaitu 1) mengubah ukuran citra menjadi 299px x 299px, dan 2) menormalisasi citra sehingga citra memiliki pixel dengan jarak antara -1 hingga 1

3. Dataset citra yang telah disiapkan kemudian dilewatkan ke CNN dengan arsitektur dari EfficientNetB5 yang menggunakan bobot dari Imagenet untuk melakukan proses feature extraction yang menghasilkan feature mapping (vektor) yang merepresentasikan object of interest dari masing-masing citra, kemudian vektor tersebut akan di encode menggunakan encoder dari Transformer dan menghasilkan context vector

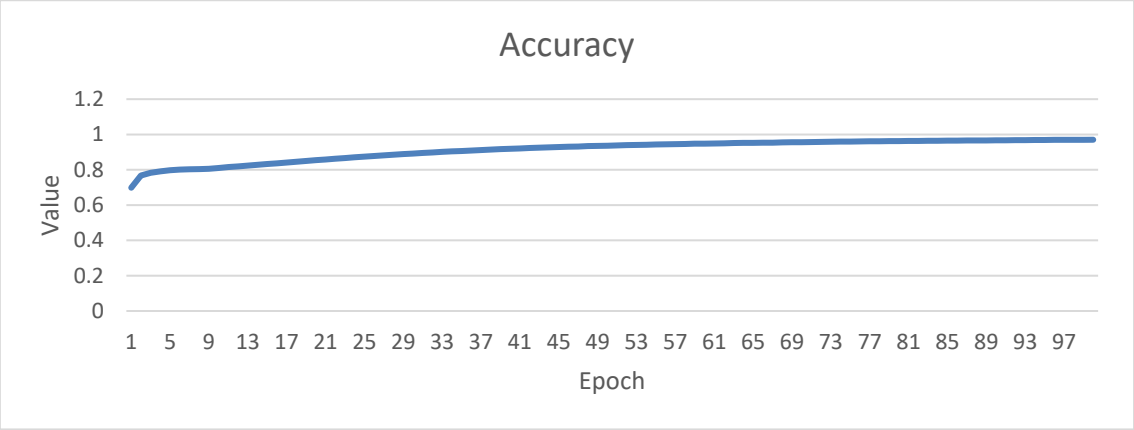
4. Model dan tokenizer yang telah dihasilkan dari tahap training dimuat ke sistem.

Context vector dan tokenizer didecode menggunakan decoder dari Transformer untuk membangun caption baru berdasarkan citra inputan dan prediksi dari model dan tokenizer.

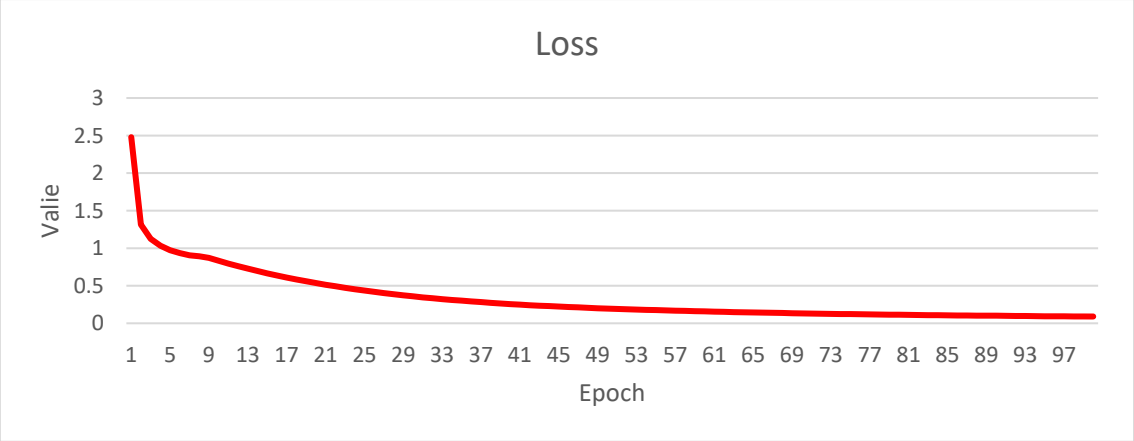
3. HASIL DAN PEMBAHASAN

3.1 *Training* Sistem

Training dilakukan dengan menggunakan 6000 citra latih beserta masing-masing caption nya dari dataset Flickr8k yang diambil dari Kaggle.com. Training dilakukan selama 100 epoch, model yang diambil adalah model yang terbaik berdasarkan validasi menggunakan 1000 citra validasi beserta masing-masing caption nya, model yang terbaik adalah model dengan capaian learning rate yang paling baik dalam satu epoch. Grafik dari learning rate ditampilkan pada Gambar 3 dan Gambar 4.



Gambar 3. Learning Rate Accuracy



Gambar 4. Learning Rate Loss

Berdasarkan learning rate yang ditunjukkan pada Gambar 3 dan Gambar 4, didapatkan model dengan accuracy tertinggi (0.9702) dan loss (0.0902) terendah pada epoch ke-100.

3.2 Skema Pengujian Sistem

Keefektifan kinerja model *image captioning* dinilai menggunakan metrik penilaian yang menghitung jumlah istilah yang muncul di teks yang dihasilkan oleh *image captioning* dengan referensi (Cui et al., 2018). Pada penelitian ini akan dilakukan pengukuran kinerja model menggunakan *Billingual Evaluatin Understudy*.

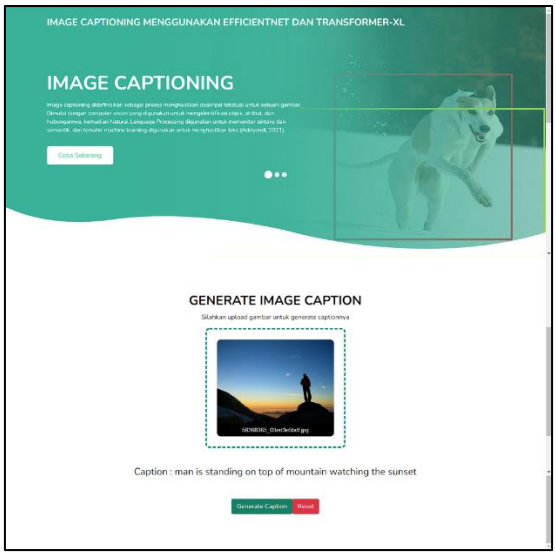
Billingual Evaluation Understudy (BLEU) merupakan nilai untuk membandingkan sebuah kalimat dengan satu atau lebih kalimat referensi. Metrik *BLEU* memiliki rentang skor 0 hingga 1, dengan 0 mewakili skor terendah dan 1 terbesar. Skor meningkat dengan jumlah kutipan terjemahan yang digunakan dalam kalimat. Dalam kehidupan nyata, skor sempurna tidak memungkinkan karena terjemahannya harus sama persis dengan referensi. Bahkan dengan bantuan penerjemah manusia, ini tidak mungkin. Metode ini bekerja dengan menghitung jumlah *n-gram* dalam teks referensi, di mana setiap token sesuai dengan 1 gram atau *unigram*, dan setiap pasangan kata sesuai dengan perbandingan *bigram* (Papineni et al., 2002). Berikut ini merupakan persamaan untuk menghitung nilai BLEU, ditunjukkan pada persamaan (1).

$$P_n = \frac{\sum_{C \in \{Candidates\}} \sum_{ngram \in C} Count_{clip}(ngram)}{\sum_{C' \in \{Candidates\}} \sum_{ngram' \in C'} Count_{clip}(ngram')} \quad (1)$$

Pengujian kinerja sistem dilakukan menggunakan 1000 citra uji. Data uji dimuat satu per satu ke program inferensi untuk dibangkitkan kalimat caption nya, kemudian hasil pembangkitan kalimat disimpan dalam sebuah array dan dilakukan scoring menggunakan BLEU dengan cara membandingkan kalimat hasil pembangkitan dan kalimat referensi yang tersedia pada dataset. Pengujian menghasilkan skor *BLEU-1* 0.718427, *BLEU-2* 0.608269, *BLEU-3* 0.563714, dan *BLEU-4* 0.472956.

3.3 Sistem Website

Pembuatan website dilakukan menggunakan *framework bootstrap, framework flask*¸ dan editor *pycharm* untuk membangun *front-end* website. Pada halaman utama website, pada bagian *header* menampilkan *slider* yang berisi penjelasan *image captioning*, *efficientnet,* dan *transformer*. Jika pengguna menekan tombol "coba sekarang" pada bagian *header* maka pengguna akan diarahkan ke bagian bawah dari website yaitu bagian *generate image caption* dimana pengguna diminta untuk *drag* sebuah gambar ke kotak yang telah disediakan ataupun klik kotak tersebut untuk memilih sebuah gambar dari direktori dan menekan tombol *generate caption* untuk melanjutkan proses.



Gambar 5. Implemetasi Website

Gambar 5 menunjukkan bahwa sistem website berhasil melakukan *image captioning* berdasarkan citra yang diinputkan oleh pengguna.

4. KESIMPULAN

Pada penelitian ini telah diimplementasikan model *pre-trained EfficientNetB5* sebagai *feature extractor* dan *Transformer* sebagai *encoder-decoder*. Metode-metode yang digunakan mampu melakukan pembangkitan kalimat dari suatu citra. Setelah model dilakukan pengujian menggunakan skor BLEU, model mampu mencapai skor tertinggi dengan skor *BLEU*-1, *BLEU*-2, *BLEU*-3, dan *BLEU*-4 (0.718427, 0.608269, 0.563714, 0.472956).

Dengan skor BLEU yang berhasil dicapai, maka dapat disimpulkan bahwa model yang dihasilkan memiliki kelayakan untuk melakukan image captioning.

DAFTAR PUSTAKA

Adriyendi. (2021). A Rapid Review of Image Captioning. In *Journal of Information Technology and Computer Science* (Vol. 6, Issue 2). www.jitecs.ub.ac.id

Ashraf Mohamed, A. (2020). *Image Caption using CNN & LSTM*. https://www.researchgate.net/publication/342407897

Cui, Y., Yang, G., Veit, A., Huang, X., & Belongie, S. (2018). *Learning to Evaluate Image Captioning*.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. v., & Salakhutdinov, R. (2019). *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*. http://arxiv.org/abs/1901.02860

Dowdell, T., & Zhang, H. (2020). *Language Modelling for Source Code with Transformer-XL*. http://arxiv.org/abs/2007.15813

Jiang, W., Ma, L., Chen, X., Zhang, H., & Liu, W. (2018). *Learning to Guide Decoding for Image Captioning*. www.aaai.org

Joseph, S. R., Hlomani, H., Letsholo, K., Kaniwa, F., & Sedimo, K. (2016). Natural Language Processing: A Review. In *International Journal of Research in Engineering and Applied Sciences* (Vol. 6, Issue 3). http://www.euroasiapub.org

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*.

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*.

Vinyals Google, O., Toshev Google, A., Bengio Google, S., & Erhan Google, D. (2015). *Show and Tell: A Neural Image Caption Generator*.

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