Implementasi Arsitektur *Transformer* pada *Image Captioning* dengan Bahasa Indonesia

Umar Abdul Aziz Al-Faruq
Program Studi Informatika Program Sarjana
Universitas Islam Indonesia
Sleman, Indonesia
umar.faruq@student.uii.ac.id

Dhomas Hatta Fudholi
Jurusan Informatika
Universitas Islam Indonesia
Sleman, Indonesia
hatta.fudholi@uii.ac.id

***Abstract*—Penelitian *image captioning* untuk menghasilkan deskripsi yang baik pada gambar dalam Bahasa Inggris banyak dilakukan. Sedikit penelitian yang ditemukan mengenai *image captioning* untuk menghasilkan deskripsi gambar dalam Bahasa Indonesia. *Image Captioning* bermanfaat untuk membantu manusia, memahami konten visual, seperti memberikan deskripsi suatu gambar yang kemudian menggunakan teknologi *text-to-speech* untuk mengubah hasil deskripsi menjadi suara sehingga bisa membantu tunanetra. Penelitian ini akan berfokus dalam mengembangkan model generative yang menggabungkan *natural language preprocessing* dan *computer vision* untuk menghasilkan deskripsi gambar dalam Bahasa Indonesia. Model yang digunakan dalam penelitian ini adalah *attention mechanism* dengan arsitektur *transformer*. Model penelitian ini menggunakan dataset MS COCO captions 2014 yang sudah diterjemahkan dan memperoleh skor rata-rata BLEU-1, BLEU-2, BLEU-3, BLEU-4 masing-masing adalah 31.12, 32.31, 42.39, 46.16.**

Keywords—Indonesia image captioning, attention mechanism, transformer

I. PENDAHULUAN

Image captioning adalah kemampuan mendeskripsikan isi sebuah gambar dalam bentuk kalimat [1]. *Image captioning* memerlukan kemampuan dari dua bidang *artificial intelligence*, yaitu *computer vision* untuk memahami isi gambar yang diberikan dan *natural language processing* untuk mengubah isi pada gambar menjadi bentuk kalimat yang natural [14].

Dalam satu gambar, biasanya memiliki beberapa objek yang setiap objeknya memiliki atribut, posisi, dan bagaimana objek teresbut terhubung dengan yang lain. Kalimat yang dihasilkan sendiri harus menjelaskan semua hal tersebut. Lebih lanjut, kalimat yang dihasilkan harus sealami bahasa manusia. Contoh gambar dan kalimatnya dapat dilihat pada Gambar. 1.

Image captioning sudah memiliki dampak positif dalam banyak bidang, contohnya analisis gambar (contoh: pencarian gambar) dan membantu orang yang memiliki masalah penglihatan atau tuna netra untuk dapat berinteraksi dengan konten visual pada situs media social [11]. *Image captioning* juga memiliki potensi untuk memberikan perubahan positif dalam hal lain misalnya interaksi manusia dengan computer dan keamanan.

Dengan perkembangan *machine translation*, munculah arsitektur baru, yaitu *transformer* [9]. Arsitektur ini menggunakan *self-attention mechanism* dan telah menjadi *state-of-the-art* pada bidang NLP (*Natural Language Processing*). *Transformer* mengakselerasi proses pelatihan dan menggunakan *self-attention* untuk menarik dependensi global antara masukkan yang berbeda. Perkembangan ini membuat model *image captioning* mulai mengadopsi arsitektur *transformer*, seperti *Meshed-Memory Transformer for Image Captioning* [2], *Boosted Transformer for Image Captioning* [8], dan *Image Captioning through Image Transformer* [1].

a man on a skateboard rides on the inside wall of an empty swimming pool
A man riding a skate board in an empty swimming pool.
A man on a skateboard doing tricks in a pool
A skate board enthusiast enjoying riding inside an old swimming pool
a person riding a skate board in an empty pool

Gambar 1. Contoh gambar dan deskripsinya

Penelitian ini akan menggunakan arsitektur *transformer* sebagai basis model *image captioning* untuk menghasilkan deskripsi gambar dalam Bahasa Indonesia. Hasil deskripsi diharapkan memiliki kalimat sealami mungkin seperti bahasa manusia. Evaluasi metrik yang akan digunakan untuk penelitian ini adalah BLEU, untuk menilai seberapa bagus deskripsi yang dihasilkan.

II. TINJAUAN PUSTAKA

A. Image Captioning

Pendekatan *image captioning* baru-baru ini menggunakan framework *deep encoder decoder*, yang terinspirasi dari pengembangan *neural machine translation*. Contohnya framework *end-to-end* digunakan dengan CNN mengkodekan gambar ke fitur vektor dan LSTM mendekodekannya menjadi kalimat. Dalam [8], *Hierarchical Attention Network* diperkenalkan yang membuat *attention* dapat dihitung dalam hirarki piramida dari fitur secara sinkronus. Dalam [10], *adaptive attention mechanism* diperkenalkan untuk memutuskan kapan mengaktifkan *visual attention*.

Didapatkan dua penelitian *image captioning* dengan Bahasa Indonesia, yaitu [3, 17]. Pendekatan *image captioning* yang dilakukan pada penelitian [3] menggunakan metode *CNN+GRU*. Pada penelitian [17] metode yang digunakan adalah *CNN+LSTM*.

B. *Attention Mechanism*

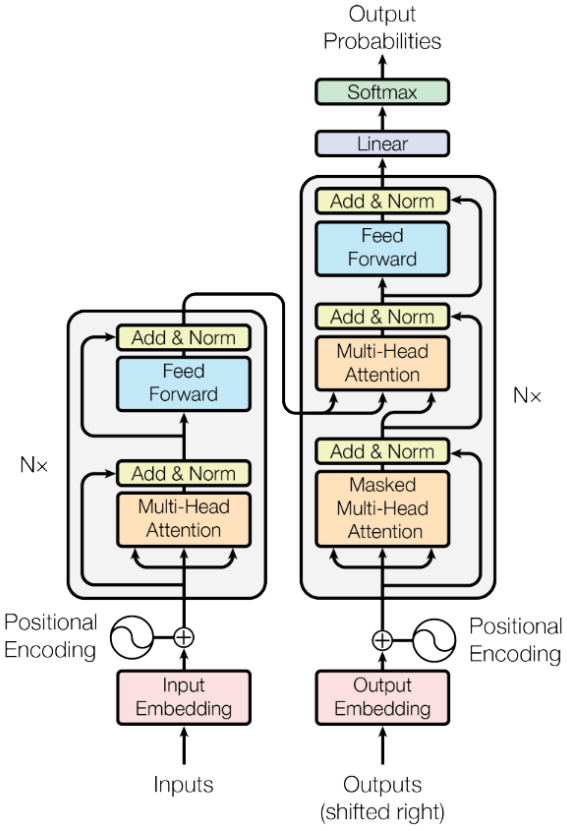
Mekanisme *attention*, yang merupakan turunan dari intuisi manusia, telah diterapkan secara luas dan menghasilkan peningkatan yang signifikan untuk masalah-masalah yang berhubungan dengan berbagai pembelajaran yang berurutan.

Attention dapat digambarkan sebagai pemetaan kueri dan sekumpulan pasangan *key-value* ke sebuah output, di mana kueri, *keys*, *values*, dan output semuanya adalah vektor. Output dihitung sebagai jumlah bobot dari *values*, di mana bobot yang ditetapkan untuk setiap *value* dihitung oleh fungsi kompatibilitas kueri dengan *key* yang sesuai [9]. Contoh penelitian yang menerapkan *attention* seperti adaptive attention, stacked attention, multi-level attention, multi-head attention and self-attention.

Penelitian oleh Vaswani et at. [9] menunjukan bahwa *self-attention* dapat mencapai hasil terbaik untuk *machine translation*. Beberapa penelitian menggunakan gagasan tersebut untuk diterapkan ke dalam *computer vision*, yang mana menginspirasi penulis untuk menerapkan *self-attention* ke *image captioning*.

C. *Transformer*

Penelitian "Attention is All You Need" [9] memperkenalkan arsitektur baru yang dinamai *transformer*. Sesuai judul makalah tersebut, *transformer* menggunakan mekanisme *attention* seperti yang dijelaskan sebelumnya. *Transformer* adalah arsitektur yang untuk mengubah satu urutan ke urutan lain dengan bantuan dua bagian, yaitu *encoder* dan *decoder*.



Gambar 2. Model arsitektur transformer [9]

Transformer adalah model transduksi pertama yang sepenuhnya mengandalkan *self-attention* untuk menghitung representasi input dan outputnya tanpa menggunakan RNN atau konvolusi. *Transformer* menggunakan lapisan *self-attention* dan *point-wise* yang ditumpuk, *fully connected* untuk *encoder* dan *decoder* yang masing-masing ditunjukan pada bagian kiri dan kanan Gambar. 2.

Transformer berbasis model *image captioning* menggunakan *dot-product attention mechanism* untuk menghubungkan daerah informatif secara implisit. Beberapa contoh penerapan *transformer* dalam *image captioning*, yaitu AoANet [4] menggunakan arsitektur lapisan *transformer* internal asli, dengan tambahan *gated linear layer* di atas *multi-head attention*. Model *entangled transformer* memiliki fitur *transformer* paralel ganda untuk mengkodekan dan menyempurnakan informasi visual dan semantic dalam gambar, yang digabungkan melalui *gated bilateral controller*.

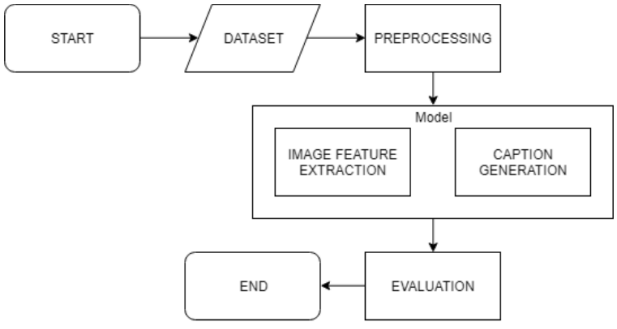
D. Penelitian Terdahulu

Kami melakukan literatur *review* pada 10 makalah penelitian *image captioning* dengan Bahasa Inggris yang menggunakan model *transformer* dan *attention mechanism* yang nantinya hasil review tersebut akan digunakan sebagai referensi untuk penelitian ini. Semua makalah yang diperoleh menggunakan dataset yang sama yaitu, *MS COCO caption 2014* dengan pembagian dataset mengikuti Karpathy [16], dengan 113,287 gambar pada *traning set*, 5,000 gambar pada *validation set*, dan 5,000 gambar pada *test set.*

Penelitian [1, 2, 5, 6] menggunakan arsitektur *transformer* sebagai model dasarnya, meski sama tiap penelitian menggunakan metodenya masing-masing. Metode tersebut antara lain [1] menggunakan *image transformer*, [2] menggunakan *meshed-memory transformer*, [5] menggunakan *boosted transformer*, dan [6] menggunakan *multimodal transformer with multi-view visual representation*. Dari penelitian tersebut metode penelitian [6] memperoleh skor BLEU paling tinggi dengan skor 81.7, 66.8, 52.4, 40.4.

Penelitian [4, 7, 8, 10, 13, 14] menggunakan *attention mechanism* sebagai model dasarnya. Metode yang digunakan pada penelitian tersebut, yaitu [4] menggunakan *attention on attention*, [7] menggunakan *hierarchy parsing*, [8] menggunakan *hierarchical attention network*, [10] menggunakan *visual sentinel*, [13] menggunakan *entangled attention*, [14] menggunakan *bottom-up and top-down attention*. Dari penelitian tersebut metode [7] mendapatkan skor BLEU paling tinggi dengan skor 81.6, 66.2, 51.5, 39.3.

III. METODOLOGI



Gambar 3. Desain sistem image captioning Bahasa Indonesia [3]

Ada empat proses utama dalam metodologi penelitian ini, yaitu *preprocessing* (gambar dan kalimat), *image feature extraction*, *caption generation*, dan *evaluation*. Proses dari *image captioning* dalam Bahasa Indonesia menggunakan *transformer* diilutrasikan pada Gambar 3.

A. Dataset

Dataset menggunakan MS COCO 2014 yang diambil dari situs cocodataset.org, yang kemudian diterjemahkan ke dalam Bahasa Indonesia menggunakan Google Translate untuk digunakan dalam penelitian ini. Dataset ini memilki 82.783 gambar dengan 413.915 kalimat dalam *training*, 40.504 gambar dengan 202.520 kalimat dalam *validation*, dan 40.775 gambar dengan 379.249 kalimat dalam *testing*. Setiap gambar dideskripsikan dengan minimal 5 kalimat. Pada penelitian ini jumlah dataset yang digunakan hanya 10.000 gambar dengan 50.000 kalimat referensi.

B. Praproses Gambar

Praproses dilakukan dengan 2 tahap, yaitu

- Mengubah ukuran gambar menjadi 299px x 299px
- Praproses gambar menggunakan metode *preprocess_input* untuk menormalisasikan gambar sehingga gambar memiliki pixel dengan jarak antara -1 hingga 1

Nilai RBG adalah diantara [0, 255]. Ini tidak ideal untuk *neural network*, umumnya nilai input perlu diperkecil. Praproses tersebut akan menormalisasikan nilai RGB ke [-1, 1].

Dua proses tersebut dilakukan untuk menyesuaikan format gambar yang digunakan pada *pretrained model InceptionV3* untuk proses ekstraksi gambar.

C. Praproses Kalimat

Praproses kalimat yang dilakukan ada 6 tahapan, yaitu

1) *Lower Case*

Mengubah semua kalimat menjadi kalimat berhuruf kecil. Proses ini bertujuan untuk memudahkan perhitungan jumlah kata unik dalam dataset.

2) *Mark Caption*

Menambahkan token "<start>" pada awal kalimat dan token "<end>" pada akhir kalimat. Ini dilakukan supaya sistem dapat mulai menghasilkan kalimat saat diberikan token "<start>" dan berhenti menghasilkan kalimat saat menemui token "<end>".

3) *Tokenizing*

Memisahkan kalimat menjadi kata yang individu. Ini dilakukan untuk memberikan kosa kata dari seluruh kata unik dalam *dataset*.

4) *Text to Sequence*

Mengubah kata ke dalam indeks kata berurutan. Proses ini akan menghasilkan vektor yang menunjukan urutan kata dalam bentuk kalimat.

5) *Pad Sequence*

Memberikan angka 0 apabila panjang dari vektor urutan kata kurang dari yang ditentukan dan menghapus akhir vektor jika panjang melebihi dari yang ditentukan.

6) *Shift Sequence*

Menggeser vektor urutan kata sebanyak satu langkah sehingga model mampu belajar menghasilkan kata berikutnya.

D. Model

InceptionV3, adalah arsitektur CNN dari keluarga *inception* yang membuat beberapa peningkatan termasuk menggunakan *Label Smoothing*, konvolusi *Factorized 7x7*, dan penggunaan pengklasifikasi tambahan untuk menyebarkan informasik label ke bawah jaringan [18].

Model I*nceptionV3* digunakan untuk melakukan ektraksi fitur pada gambar. Karena hanya perlu mengekstrak vektor gambar, lapisan *softmax* dari model ini perlu dihapus. Sebelum diberikan kepada model, semua gambar harus melakukan praproses dengan menyamakan ukuran gambar 299X299. Bentuk output dari model ini adalah 8x8x2048.

Positional Encoding, *positional encoding* menggunakan fungsi *sin* dan *cos* dari frekuensi yang berbeda. Fungsi *cos* dibuat jika posisi indeks ganjil pada vektor input, sedangkan fungsi *sin* dibuat jika posisi indeks genap pada vektor input. Menambahkan vektor-vektor tersebut ke *embedding input* yang sesuai yang berhasil memberikan informasi jaringan pada posisi setiap vektor.

$$PE_{(pos,2i)} = sin\left(pos/10000^{2i/d_model}\right) \quad (1)$$

$$PE_{(pos,2i+1)} = cos\left(pos/10000_{2i/d_model}\right) \quad (2)$$

pos adalah posisi dan *i* adalah dimensi. Setiap dimensi dari *positional encoding* sesuai dengan *sinusoid*. Panjang gelombang membentuk deret geometri dari 2π ke $10.000 \cdot 2\pi$. Fungsi ini dipilih karena hipotesisnya memungkinkan model dapat dengan mudah belajar posisi relarif, karena untuk setiap *offset* tetap *k*, PE_{pos+k} dapat direpresentasikan sebagai fungsi linier dari PE_{pos} [9].

Multi-Head Attention, menghitung bobot *attention*. Q (query), K (key), V (value) harus memiliki dimensi utama yang cocok. K dan V harus memiliki dimensi kedua dari belakang yang cocok. Penutup memiliki bentuk yang berbeda tergantung pada jenisnya (*padding* atau *look-ahead*). Terdapat 4 bagian dalam *multi-head attention* ini, yaitu

- Lapisan linier
- *Scalied dot-product attention*
- *Concatenation*
- Lapisan linier akhir

E. Loss Function

Fungsi sparse cross entropy digunakan sebagai *loss function* dalam penelitian ini.

$$L(I,S) = -\sum_{t=1}^{N} log\ p_t(S_t) \quad (3)$$

F. Skor BLEU

Skor BLEU (*Bilingual Evaluation Understudy*) [12] adalah metrik yang sering digunakan untuk mengevaluasi hasil dari model *machine translation* dan juga model *image captioning* dan memberikan hasil yang baik. Metrik ini mengevaluasi kalimat yang dihasilkan ke kalimat target.

Kalimat yang dihasilkan mirip dengan kalimat target akan diberikan skor 1.0 dan apabila kalimatnya tidak mirip akan diberikan skor 0.0.

BLEU menghitung ketepatan menggunakan metrik *n-grams*. Panjang maksimal *n-grams* adalah 4 karena memiliki korelasi paling tinggi dengan penilaian manusia. Skor BLEU diadopsi dari [BLEU] dengan formula:

$$BLEU = min(1, \frac{output - length}{reference - length}) \left(\prod_{i=1}^{4} precision_i \right)^{\frac{1}{4}} \quad (4)$$

IV. HASIL DAN PEMBAHASAN

Hasil dievaluasi dengan perhitungan kuantitatif dan kualitatif. Hasil kuantitatif dihitung menggunakan skor BLEU. Setiap kalimat yang dihasilkan oleh model pada gambar yang diberikan akan dibandingkan dengan 5 kalimat target yang digunakan sebagai referensi kalimat untuk menghitung skor BLEU. Ada 2000 gambar untuk menguji model. Gambar tersebut belum pernah dipelajari oleh model sebelumnya. Sehingga dapat mencari tau apakah model dapat belajar dengan baik atau tidak.

Tabel 1 menggambarkan performa model yang digunakan menggunakan skor BLEU. 3 skenario eksperimen dilakukan untuk menguji lapisan *Multi-Head Attention* berdasarkan jumlah lapisannya. Kami menggunakan 4, 5, dan 6 lapisan. Jumlah lapisan tersebut dipilih karena mengikuti jumlah lapisan yang digunakan dalam penelitian [9] dan sedikit pengurangan dilakukan.

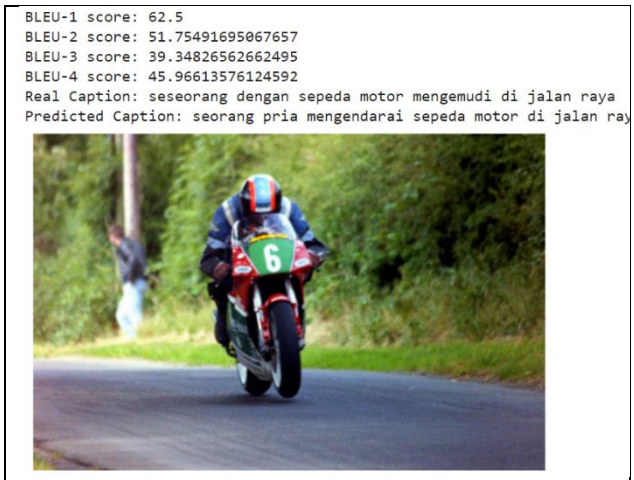
TABEL 1. SKOR RATA-RATA BLEU UNTUK IMAGE CAPTIONING DALAM BAHASA INDONESIA

# lapisan	BLEU-1	BLEU-2	BLEU-3	BLEU-4
4	31.12	32.31	42.39	46.16
5	31.49	32.07	41.38	44.89
6	31.24	32.18	41.87	45.54

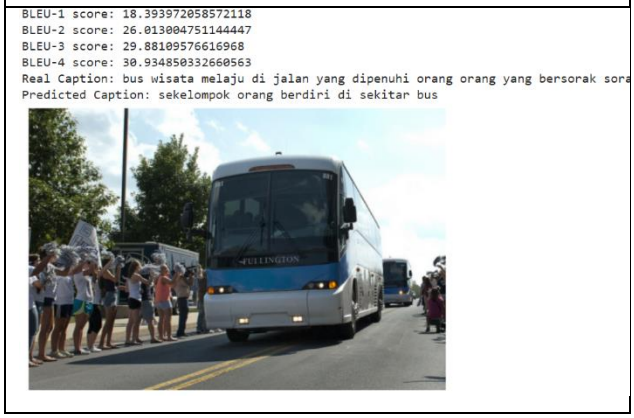
Hasil kualitatif, beberapa gambar dengan kalimat yang dihasilkan oleh model ditunjukan dan kualitasnya juga dijelaskan dalam Tabel 2. Sebagian besar kalimat yang dihasilkan memiliki tata bahasa yang benar.

TABEL 2. HASIL PREDIKSI KALIMAT DAN SKOR BLEU

Keterangan tanpa *error*



Keterangan dengan sedikit *error*



V. KESIMPULAN

Sebagai ringkasan, beberapa hal yang dilaporkan dalam makalah ini adalah 1) meneliti model *image captioning* dengan Bahasa Indonesia, 2) menemui masalah dataset *image captioning* Bahasa Indonesia, 3) menggunakan *encoder-decoder* sebagai model *image captioning* dan pembuatan kalimat dalam Bahasa Indonesia, 4) menggunakan arsitektur *transformer*.

Model yang digunakan dalam penelitian ini terdiri dari *InceptionV3 deep convolutional neural network* yang merupakan *pre-trained* menggunakan *dataset ImageNet*, *pre-trained* model *word embedding* dan 4 lapisan *Multi-Head Attention.* Model ini memperoleh rata-rata skor BLEU-1, BLEU-2, BLEU-3, dan BLEU-4 dengan skor masing-masing adalah 31.12, 32.31, 42.39, 46.16.

Untuk penelitian dan pengembangan kedepan, disarankan menggunakan penelitian terbaru dalam *image captioning* yang umumnya menghasilkan kalimat dalam Bahasa Inggris dan mengimplementaikannya untuk Bahasa Indonesia.

DAFTAR PUSTAKA

[1] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, “Image Captioning through Image Transformer,” 2020, [Online]. Available: http://arxiv.org/abs/2004.14231.

[2] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, “Meshed-Memory Transformer for Image Captioning,” pp. 10575–10584, 2020, doi: 10.1109/cvpr42600.2020.01059.

[3] A. A. Nugraha, A. Arifianto, and Suyanto, “Generating image description on Indonesian language using convolutional neural

network and gated recurrent unit," *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, pp. 98–103, 2019, doi: 10.1109/ICoICT.2019.8835370.

[4] L. Huang, W. Wang, J. Chen, and X. Y. Wei, "Attention on attention for image captioning," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, pp. 4633–4642, 2019, doi: 10.1109/ICCV.2019.00473.

[5] J. Li, P. Yao, L. Guo, and W. Zhang, "Boosted Transformer for Image Captioning," *Appl. Sci.*, vol. 9, p. 3260, 2019, doi: 10.3390/app9163260.

[6] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *arXiv*, vol. 14, no. 8, pp. 1–12, 2019, doi: 10.1109/tcsvt.2019.2947482.

[7] T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy parsing for image captioning," *Proc. IEEE Int. Conf. Comput. Vis.*, vol.

[8] W. Wang, Z. Chen, and H. Hu, "Hierarchical attention network for image captioning," 33rd AAAI Conf. Artif. Intell. AAAI 2019, 31st Innov. Appl. Artif. Intell. Conf. IAAI 2019 9th AAAI Symp. Educ. Adv. Artif. Intell. EAAI 2019, pp. 8957–8964, 2019, doi: 10.1609/aaai.v33i01.33018957.

[9] A. Vaswani et al., "Attention is all you need," Adv. Neural Inf. Process. Syst., vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.

[10] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 3242–3250, 2017, doi: 10.1109/CVPR.2017.345.

[11] V. Voykinska, S. Azenkot, S. Wu, and G. Leshed, "How blind people interact with visual content on social networking services," Proc. ACM Conf. Comput. Support. Coop. Work. CSCW, vol. 27, pp. 1584–1595, 2016, doi: 10.1145/2818048.2820013.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, no. October 2002, pp. 311–318, 2001, doi: 10.3115/1073083.1073135.

[13] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2019-Octob, no. c, pp. 8927–8936, 2019, doi: 10.1109/ICCV.2019.00902.

[14] P. Anderson *et al.*, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 6077–6086, 2018, doi: 10.1109/CVPR.2018.00636.

[15] X. Chen *et al.*, "Microsoft COCO Captions: Data Collection and Evaluation Server," *arXiv*, pp. 1–7, 2015, [Online]. Available: http://arxiv.org/abs/1504.00325.

[16] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 664–676, 2017, doi: 10.1109/TPAMI.2016.2598339.

[17] A. M. Nugroho and A. F. Hidayatullah, "Keterangan Gambar Otomatis Berbahasa Indonesia dengan CNN dan LSTM," *Automata*, vol. 2, no. 1, pp. 0–3, 2021.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.