Pemahaman Visual di Dalam Ruangan dengan *Image Captioning* berbasis Transformer

Royan Abida Nur Nayoan
20917031

Tesis diajukan sebagai syarat untuk meraih gelar Magister Komputer
Konsentrasi Sains Data
Program Studi Informatika Program Magister
Fakultas Teknologi Industri
Universitas Islam Indonesia
2022

Lembar Pengesahan Pembimbing

Pemahaman Visual di Dalam Ruangan dengan *Image Captioning* berbasis Transformer

Royan Abida Nur Nayoan
20917031

Yogyakarta, Juli 2022

Pembimbing

Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D.

Lembar Pengesahan Penguji

Pemahaman Visual di Dalam Ruangan dengan *Image Captioning* berbasis Transformer

Royan Abida Nur Nayoan

20917031

Yogyakarta, Juli 2022

Tim Penguji,

Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D.
Ketua

Chandra Kusuma Dewa, S.Kom., M.Cs., Ph.D
Anggota I

Irving Vitra Paputungan, S.T., M.Sc., Ph.D.
Anggota II

Mengetahui,

Ketua Program Studi Informatika Program Magister

Universitas Islam Indonesia

Izzati Muhimmah, S.T., M.Sc., Ph.D.

Abstrak

Pemahaman Visual di Dalam Ruangan dengan *Image Captioning* berbasis Transformer

Model enkoder-dekoder telah menjadi model standar untuk digunakan sebagai framework untuk menyelesaikan masalah *image captioning* dengan CNN sebagai enkoder dan RNN sebagai dekoder. Namun RNN memiliki kekurangan dalam dependensi jangka panjang dalam jaringannya dan menyebabkan RNN kesulitan dalam mengingat urutan panjang yang kemudian diperbaiki dengan munculnya Transformer dengan mekanisme *attention*. Transformer telah banyak digunakan dalam tugas *image captioning* pada dataset berbahasa Inggris seperti MSCOCO dan Flickr. Namun begitu, penelitian terkait *image captioning* dengan Bahasa Indonesia masih sedikit dan menggunakan penerjemah untuk mendapatkan dataset berbahasa Indonesia. Pada penelitian ini, digunakan model Transformer untuk memprediksi deskripsi gambar pada dataset modifikasi MSCOCO dan Flickr berbahasa Indonesia untuk mendapatkan pemahaman visual di dalam ruangan. Dataset yang digunakan merupakan dataset yang telah dimodifikasi dengan membuat *captions* menjadi *captions* baru Berbahasa Indonesia dengan menuliskan deskripsi yang mengandung nama objek, warna, posisi/lokasi (sudut pandang pengguna), karakteristik, dan objek sekitarnya. Dilakukan eksperimen dengan menggunakan varian model pre-trained CNN untuk mendapatkan fitur gambar sebelum dilanjutkan pada model Transformer. Kemudian dilakukan pengaturan *hyperparameter* pada model dengan mengubah ukuran batch, dropout, dan attention heads untuk mendapatkan model terbaik. Matriks evaluasi yang digunakan yakni BLEU-n, METEOR, CIDEr, dan ROUGE-L untuk mengevaluasi model. Dari penelitian ini, didapatkan model dengan memanfaatkan fitur ekstraktor IncepResNetV2 yang memiliki ukuran batch dengan nilai 128, dropout dengan nilai 0.1, dan attention heads dengan nilai 4 mampu mendapatkan skor terbaik di semua matriks evaluasi. Model IncepResNetV2 mendapatkan skor tertinggi pada BLEU-1 dengan skor 0.6971, BLEU-2 dengan skor 0.5246, BLEU-3 dengan skor 0.3921, BLEU-4 dengan skor 0.2831, METEOR dengan skor 0.2468, CIDEr dengan skor 0.4801, dan ROUGE-L dengan skor 0.5114.

Kata kunci

bahasa Indonesia, berbasis cnn, ekstraksi fitur, image captioning, pemahaman visual, ruang dalam, transformer

Abstract

Transformer-based Image Captioning for Indoor Environment Visual Understanding

The encoder-decoder model has become a standard model to be used as a framework to solve image captioning problems. This model usually uses CNN as the encoder and RNN as the decoder. However, RNN has a drawback in long-term dependencies in their network that make them difficult to remember longer words. Transformer is developed to overcome RNN's drawback usin an attention mechanism. Transformer has been widely used in image captioning tasks to generate image descriptions using English datasets such as MSCOCO and Flickr. However, there are only a few image captioning research using Indonesian. These previous researches use datasets that are already available such as MSCOCO and Flickr and use English-Indonesia professional translators or Google translator to obtain the datasets. In this study, we use the Transformer model to predict image's caption using our modified MSCOCO and Flickr datasets using Indonesian to gain indoor visual understanding. Here, we make our own modified dataset using MSCOCO and Flickr pictures and remove their original captions. We then added our own captions that contains object names, colors, positions/locations (user point of view), characteristics, and its surrounding objects. We also conducted experiments using several pre-trained CNN variants to obtain image features before feeding them to the Transformer model. The models we use are InceptionV3, ResNet50, Xception, DenseNet201, and IncepResNetV2. We also adjust the hyperparameters on the model by changing the batch size, dropout, and attention heads to get the best model. We use the evaluation matrices of BLEU-n, METEOR, CIDEr, and ROUGE-L to get the best model. From our experiment, we found that the model by utilizing IncepResNetV2 as feature extractor and by setting the hyperparameter of batch to the size of 128, dropout with a value of 0.1, and attention heads with a value of 4, we are able to get the best score in all evaluation matrices. The IncepResNetV2 model got the highest score on BLEU-1 with a score of 0.6971, BLEU-2 with a score of 0.5246, BLEU-3 with a score of 0.3921, BLEU-4 with a score of 0.2831, METEOR with a score of 0.2468, CIDEr with a score of 0.4801, and ROUGE-L with a score of 0.5114.

Keywords

bahasa indonesia, cnn-based, feature extraction, image captioning, indoor environment, visual understanding, transformer

Pernyataan Keaslian Tulisan

Dengan ini saya menyatakan bahwa tesis ini merupakan tulisan asli dari penulis, dan tidak berisi material yang telah diterbitkan sebelumnya atau tulisan dari penulis lain terkecuali referensi atas material tersebut telah disebutkan dalam tesis. Apabila ada kontribusi dari penulis lain dalam tesis ini, maka penulis lain tersebut secara eksplisit telah disebutkan dalam tesis ini.

Dengan ini saya juga menyatakan bahwa segala kontribusi dari pihak lain terhadap tesis ini, termasuk bantuan analisis statistik, desain survei, analisis data, prosedur teknis yang bersifat signifikan, dan segala bentuk aktivitas penelitian yang dipergunakan atau dilaporkan dalam tesis ini telah secara eksplisit disebutkan dalam tesis ini.

Segala bentuk hak cipta yang terdapat dalam material dokumen tesis ini berada dalam kepemilikan pemilik hak cipta masing-masing. Apabila dibutuhkan, penulis juga telah mendapatkan izin dari pemilik hak cipta untuk menggunakan ulang materialnya dalam tesis ini.

Yogyakarta, Juli 2022

Royan Abida Nur Nayoan, S. Kom

Daftar Publikasi

Fudholi, D. H., & Nayoan, R. A. N. (2022). The Role of Transformer-based Image Captioning for Indoor Environment Visual Understanding. International Journal of Computing and Digital Systems.

Kontributor	Jenis Kontribusi
Dhomas Hatta Fudholi	Memberikan ide (70%) Mendesain eksperimen (30%) Menulis dan memperbaiki paper (30%)
Royan Abida N. Nayoan	Memberikan ide (30%) Mendesain eksperimen (70%) Menulis paper (30%)

Fudholi, D. H., Zahra, A., & Nayoan, R. A. N. (2022). A Study on Visual Understanding Image Captioning using Different Word Embeddings and CNN-Based Feature Extractions. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 91-98.

Kontributor	Jenis Kontribusi
Dhomas Hatta Fudholi	Memberikan ide (70%) Mendesain eksperimen (30%) Mereview paper (20%)
Annisa Zahra	Memberikan ide (30%) Mendesain eksperimen (70%) Menulis paper (50%)
Royan Abida N. Nayoan	Menganalisis dan menulis paper (30%)

Halaman Kontribusi

Terima kasih penulis sampaikan kepada Bapak Dhomas Hatta Fudholi selaku pembimbing Tesis yang telah membimbing dan memberikan saran kepada penulis dalam melaksanakan penelitian ini.

Halaman Persembahan

Segala puji syukur bagi Allah SWT yang atas izinNya, Penulis mampu menyelesaikan Pendidikan Magister di Universitas Islam Indonesia. Dengan rasa syukur yang mendalam, Penulis mempersembahkannya kepada:

1. Keluarga yang senantiasa memberi dukungan dan do'a.
2. Bapak Dhomas Hatta Fudholi selaku pembimbing yang telah memberikan saran, semangat, serta mendorong penulis untuk berkembang.

Kata Pengantar

Dengan menyebut nama Allah SWT yang Maha Pengasih lagi Maha Penyayang. Segala puji bagi Allah, Tuhan semesta alam. Dengan rahmat dan karunia-Nya Penulis dapat menyelesaikan Tugas Akhir dengan judul "*Pemahaman Visual di Dalam Ruangan dengan Image Captioning berbasis Transformer*" yang disusun sebagai syarat akhir yang harus ditempuh untuk menyelesaikan pendidikan pada jenjang Strata Dua (S2) Jurusan Informatika Universitas Islam Indonesia.

Ucapan terimakasih tak lupa Penulis haturkan kepada segenap pihak baik yang secara langsung maupun tidak langsung turut serta dalam membantu menyelesaikan Tugas Akhir ini. Secara khusus Penulis menyampaikan terima kasih yang sebesar-besarnya kepada Bapak Dhomas Hatta Fudholi selaku pembimbing yang senantiasa memberikan bimbingan, masukan, dan diskusi yang intensif dalam penulisan dan penggunaan model yang digunakan dalam tugas *image captioning*.

Besar harapannya Tugas Akhir ini dapat bermanfaat untuk menambah wawasan serta ilmu pengetahuan terutama dalam bidang *image captioning*, baik bagi Penulis maupun para pembaca pada umumnya. Penulis menyadari bahwa tulisan ini jauh dari kata sempurna, oleh karena itu, diharapkan adanya kritik, saran, dan usulan untuk menjadi lebih baik ke depannya.

Yogyakarta, Juli 2022

Royan Abida N. Nayoan

Daftar Isi

Daftar Tabel

Daftar Gambar

BAB 1
Pendahuluan

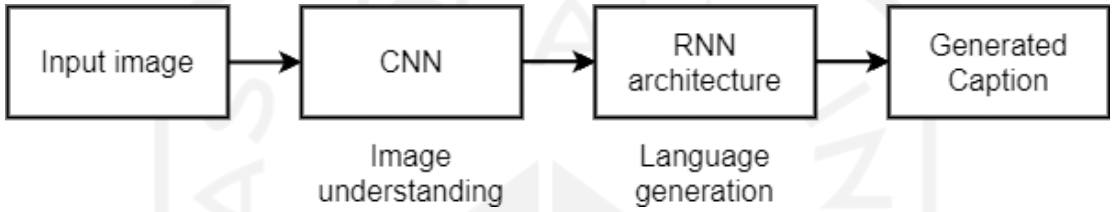
1.1 Latar Belakang

Image captioning sudah sangat popular di bidang kecerdasan buatan yang membantu dalam mengenerasi deskripsi gambar. Untuk mendeskripsikan gambar, *image captioning* menggabungkan visi computer, Natural Language Processing (NLP), dan pembelajaran mesin. *Image captioning* memiliki beberapa peran penting untuk berbagai macam tujuan serta dapat diaplikasikan dalam berbagai skenario seperti menambahkan subtitle pada video, video tanya jawab, pencarian gambar (Wang et al., 2021), dan aplikasi bantuan untuk tunanetra. Bagi tunanetra dan orang yang kesulitan dalam melihat, *image captioning* mampu menyumbangkan manfaat yang besar dalam membantu memberikan gambaran atas apa yang sedang terjadi dan apa yang ada di sekitar mereka. Berangkat dari ide tersebut, pada penelitian ini dibangun model Transformer untuk membantu mendeskripsikan lingkungan di dalam ruangan menggunakan Bahasa Indonesia untuk mencapatkan pemahaman visual.

Seiring dengan berkembangnya pembelajaran mesin, semakin banyak pula penelitian yang dilakukan pada tugas *image captioning* untuk meningkatkan performa dalam mendapatkan deskripsi yang akurat dari gambar dengan memanfaatkan arsitektur-arsitektur baru. Namun kebanyakan metode *image captioning* menggunakan *framework* enkoder-dekoder yang terdiri dari dua tahap sederhana (F. Chen et al., 2021; H. Sharma et al., 2020). Sesuai namanya, tahap pertama yakni enkoder. Pada tahap ini, CNN biasa digunakan sebagai enkoder untuk meng-enkode gambar dan mengubahnya menjadi *vector embedding*. Tahap kedua yakni dekoder. Dekoder berguna untuk mengenerasi teks per kata. Model *recurrent* biasanya sering digunakan sebagai dekoder. Model enkoder-dekoder ini telah digunakan pada studi sebelumnya (Jia et al., 2015; Vinyals et al., 2014) dengan menggunakan LSTM untuk menghasilkan teks yang berkualitas dan CNN sebagai enkoder untuk memetakan fitur gambar menjadi representasi vektor gambar. Gambar 1.1 menampilkan ilustrasi dari arsitektur enkoder-dekoder yang biasa digunakan pada tugas *image captioning*.

RNN sudah banyak digunakan sebagai bagian decoder dalam tugas *image captioning.* Namun begitu RNN masih memiliki kesulitan dalam hal dependensi jangka panjang dan juga lama Ketika melakukan training. Pada tahun 2017, Vaswani et al. (Vaswani et al., 2017) mengenalkan Transformer yang menwarkan solusi dan menyelesaikan masalah RNN. Sejak itu, beberapa model terobosan baru diciptakan berdasarkan dari model

Transformer seperti BERT (Devlin et al., 2019). Hal ini menunjukkan bahwa Transformer dengan memanfaatkan *self-attention* mampu memberikan hasil yang mumpuni dibandingkan model-model RNN. Hal ini juga membantu Transformer dalam mendapatkan popularitas dan digunakan sebagai arsitektur standar dalam berbagai tugas pemahaman Bahasa, termasuk *image captioning* yang juga merupakan masalah *sequence.* Beberapa penelitian *image captioning* yang memanfaatkan Transformer juga menghasilkan skor dan teks generasi yang menjanjikan (G. Li et al., 2019; J. Li et al., 2019).



Gambar 1.1 Arsitektur enkoder-dekoder pada *image captioning* (Al-Malla et al., 2022).

Tidak banyak data gambar dengan keterangan teks berbahasa Indonesia yang dapat digunakan untuk mendukung *image captioning* menggunakan bahasa Indonesia (Mahadi et al., 2020). Untuk mendapatkan model yang mampu menghasilkan teks yang baik dan natural, maka diperlukan dataset yang sudah ditranslasikan dengan baik pula. Penelitian sebelumnya menggunakan dataset berbahasa Indonesia dengan memanfaatkan mesin Google translate atau translator professional Inggris-Indonesia untuk mentranslasi dataset berbahasa Inggris dari MSCOCO atau Flickr (Mahadi et al., 2020; Mulyanto et al., 2019). Berbeda dari penelitian-penelitian *image captioning* yang telah disebutkan, dataset yang digunakan pada penelitian ini mengambil dari data yang disediakan oleh MSCOCO lalu menghapus keterangan teks aslinya. Kemudian dibuat keterangan gambar baru dengan menggunakan bahasa Indonesia dengan mencantumkan nama objek, posisi/lokasi (sesuai posisi pengguna), warna, karakteristik, dan objek terdekatnya.

Pada penelitian ini, dibangun model untuk mengenerasi teks dari gambar yang diambil di dalam ruangan. Model yang digunakan yakni model Transformer yang kemudian diubah dan disesuaikan modelnya dengan melakukan *hyper-parameter tuning* untuk mendapatkan model terbaik dalam mengenerasi deskripsi gambar serta melakukan komparasi dari beberapa penggunaan model ekstraksi fitur gambar. Dilakukan pula beberapa eksperimen dalam menggunakan beberapa varian CNN untuk mendapatkan fitur gambar sebelum dilanjutkan ke dalam model Transformer. Model ini berkontribusi untuk membantu

mengidentifikasi objek indoor untuk mendapatkan pemahaman visual di dalam ruangan indoor.

1.2 Rumusan Masalah

Dalam pembuatan *image captioning* untuk mendapatkan pemahaman visual di dalam ruangan, penulis merumuskan beberapa rumusan masalah di antaranya:

a. Bagaimana mendapatkan metode terbaik dalam melakukan *image captioning* menggunakan data MSCOCO dan Flickr dengan keterangan gambar berbahasa Indonesia?
b. Bagaimana hasil implementasi *image captioning* menggunakan arsitektur terbaik untuk mendapatkan pemahaman visual di dalam ruangan?

1.3 Tujuan Penelitian

Dari rumusan masalah yang telah dirumuskan untuk melakukan *image captioning*, maka dapat dituliskan tujuan penelitian sebagai berikut:

a. Mengetahui hasil terbaik dari penggunaan metode yang digunakan untuk melakukan *image captioning* dengan menggunakan data MSCOCO dan Flickr dengan keterangan gambar berbahasa Indonesia
b. Mengetahui hasil implementasi *image captioning* dari penggunaan arsitektur terbaik untuk mendapatkan pemahaman visual di dalam ruangan

1.4 Batasan Penelitian

Dalam melakukan penelitian *image captioning*, digunakan beberapa batasan penelitian, yaitu:

a. Dataset yang digunakan hanya objek di dalam ruangan, yakni tempat tidur, wastafel, kursi, sofa, meja, televisi, kulkas, tanaman rumah, oven, tangga, lampu gantung, rak gantung, pintu dan jendela, mesin cuci, dan telepon genggam yang didapatkan dari MSCOCO dan melalui *scraping* situs Flickr
b. Digunakan *captions* baru dengan bahasa Indonesia untuk setiap gambar dan menghapus *captions* asli MSCOCO
c. Model yang digunakan untuk melakukan ekstraksi fitur dalam *image captioning* adalah Densenet201, Resnet50, InceptionResnetv2, InceptionV3, dan Xception

d. Matriks evaluasi yang digunakan adalah BLEU-n, METEOR, CIDEr, dan ROUGE-L saja.

1.5 Sistematika Penulisan

Dalam menuliskan laporan penelitian, penulis menggunakan sistematika penulisan sebagai berikut.

BAB 1 PENDAHULUAN

Bagian Pendahuluan memuat materi terkait latar belakang penelitian, rumusan masalah, tujuan penelitian, dan batasan penelitian.

BAB 2 TINJAUAN PUSTAKA

Bagian Tinjauan Pustaka menyuguhkan landasan teori yang berupa teori pendukung penelitian *image captioning* dan berisi kajian pustaka terhadap penelitian-penelitian sebelumnya yang berkaitan dengan *image captioning*.

BAB 3 METODOLOGI PENELITIAN

Bagian Metodologi Penelitian menyajikan metode atau langkah-langkah yang digunakan penulis untuk melakukan penelitian. Pada bagian ini memuat penjelasan singkat mengenai data, pengolahan data, model, dan matriks yang digunakan untuk mengevaluasi model *image captioning*.

BAB 4 HASIL DAN PEMBAHASAN

Bagian Hasil dan Pembahasan menyajikan hasil serta pembahasan dari implementasi *image captioning.* Pada bagian ini penulis menjelaskan implementasi *image captioning* secara rinci dengan menyantumkan *code* yang digunakan untuk mendapatkan data, pengolahan data, model *image captioning* yang digunakan, dan hasil dari matriks evaluasi pada model *image captioning.*

BAB 5 PENUTUP

Bagian Penutup memuat butir-butir kesimpulan penting yang dirangkum penulis dari penelitian yang telah dilakukan serta saran penelitian yang dapat dilakukan ke depannya di bidang *image captioning*.

BAB 2
Tinjauan Pustaka

2.1 Landasan Teori

2.1.1 Image Captioning

Image captioning merupakan sebuah tugas yang cukup popular dalam bidang kecerdasan buatan yang bertujuan untuk menghasilkan deskripsi dari sebuah gambar secara otomatis. Dalam sebuah gambar terdapat beberapa objek didalamnya, di mana setiap objek memiliki atribut, posisi dan hubungan antara objek-objek tersebut (Nugraha et al., 2019). Oleh karena itu, sebuah model *image captioning* diharapkan mampu untuk mendeteksi dan mengenali objek dan atributnya, memahami tipe skenario, lokasi, dan hubungan antara objek yang ada di dalam gambar. Selain itu, *image captioning* juga perlu untuk menghasilkan kalimat deskripsi yang sesuai dengan gambar secara sintetik maupun semantic (Zakir Hossain et al., 2019).

Menurut survei terkait *Image captioning* saat ini, banyak penelitian yang memanfaatkan model berbasis *deep learning* yang dinilai mampu untuk mengatasi kompleksitas tugas ini (Zakir Hossain et al., 2019). *Image captioning* membutuhkan pixel gambar sebagai masukan yang kemudian melalui tahap enkode visual, pixel akan dienkode sebagai vektor fitur. Tahap selanjutnya adalah mengenerasi Bahasa dimana akan dihasilkan urutan kata atau sub kata yang didekode sesuai dengan kosakata yang diberikan.

Penelitian *image captioning* biasa menggunakan framework enkoder-dekoder dengan teknik *deep learning* yang dinilai mampu untuk mempelajari data latih yang besar seperti gambar maupun video. Saat ini, model yang biasa digunakan sebagai enkoder adalah Convolutional Neural Networks (CNN) yang biasa digunakan untuk mengekstraksi fitur dari lapisan konvolusi terakhir kemudian diikuti Recurrent Neural Networks (RNN) untuk mengenerasi deskripsi teks. CNN masih banyak digunakan karena model mampu untuk mendapatkan berbagai aspek dari objek dan hubungannya di dalam gambar sehingga mampu merepresentasikan gambar dalam level tinggi (Katiyar & Borgohain, 2021). Namun begitu, semenjak munculnya Transformer dengan performa mumpuni di bidang pengolahan Bahasa (Vaswani et al., 2017), kini RNN digantikan dengan Transformer sebagai dekoder karena kemampuannya dalam melakukan pelatihan secara parallel dan hasil performa yang

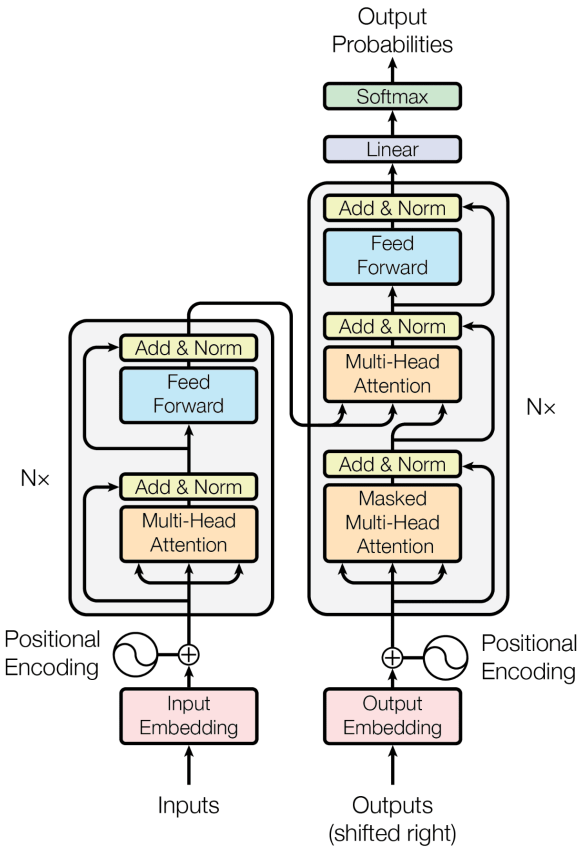
memuaskan. Sedangan penggunaan model varian CNN hingga saat ini masih tetap digunakan sebagai bagian enkoder untuk mendapatkan fitur gambar.

2.1.2 Transformer

Transformer pertama kali diajukan pada paper "Attention is All You Need" oleh Vaswani et al. (Vaswani et al., 2017). Sesuai dengan judulnya, Transformer menggunakan mekanisme *attention* untuk meningkatkan kecepatan. *Attention* dikenal karena kemampuannya untuk mengikuti bagaimana otak bekerja, yang secara selektif memilih informasi penting dan mengabaikan informasi lain. *Attention* bekerja dengan selektif untuk mencari urutan penting di setiap tahapannya di masukan sekuensial. Transformer merupakan arsitektur Seq2Seq dengan dua bantuan bagian, enkoder dan dekoder, namun beda dengan arsitektur Seq2Seq biasanya karena Transformer tidak membutuhkan model *recurrent*. Transformer merupakan model transduksi yang mengandalkan *self-attention* untuk mengkomputasi representasi masukan dan keluaran.

Sebelum Transformer, RNN menjadi model standar yang digunakan sebagai dekoder dalam tugas *image captioning.* Namun RNN memiliki kekurangan dalam dependensi jangka panjang dalam jaringannya yang menyebabkan RNN kesulitan dalam mengingat kata jauh sebelumnya. Transformer hadir untuk mengatasi mengatasi hal ini dengan mekanisme *attention.* Transformer tidak hanya menggunakan satu *Attention* saja, namun model ini menggunakan distribusi banyak *attention* dan banyak keluaran untuk satu masukan. Transformer juga menggunakan lapisan normalisasi dan koneksi residual untuk optimisasi yang lebih cepat dan mudah. Dalam memberikan posisi dalam masukan, Transformer menggunakan enkoding posisi eksplisit.

Enkoder dan dekoder di Transformer terbuat dari tumpukan enkoder dan dekoder sesuai ilustrasi yang ditunjukkan pada Gambar 2.1. Model Transformer tersusun dari 6 enkoder dan 6 dekoder yang tersusun bertumpukan di atasnya. Blok enkoder terdiri dari satu lapisan *multi-head attention* (MHA) dan lapisan *feed forward*. Blok dekoder memiliki lapisan yang sama dengan enkoder, namun dekodernya memiliki satu ekstra MHA yang terletak di antara layernya. MHA memungkinkan model untuk melihat posisi lain di masukan untuk mendapatkan enkoding terbaik untuk katanya.



Gambar 2.1 Ilustrasi Arsitektur Transformer (Vaswani et al., 2017).

Transformer memiliki beberapa komponen penting untuk *image captioning* seperti Multi Head Attention (MHA), *positional-wise feed forward layer* dan *positional encoding.* MHA merupakan komponen inti dari lapisan *self-attention* dan *cross-attention* (Bahdanau et al., 2015) dengan *multi head* dengan bobot yang dipelajari berbeda. *Attention* diterapkan menggunakan *scaled dot-products* untuk menghitung *similarity* (Vaswani et al., 2017) sementara *key*, *query*, dan *value* dihitung melalui transformasi linier. Pada tahap *encoding*, urutan vektor fitur visual digunakan untuk menyimpulkan *query*, *key*, dan *value*, sehingga menciptakan pola *self-attention* di mana hubungan visual berpasangan dimodelkan. Pada tahap *decoding*, urutan kata digunakan untuk menyimpulkan *query*, dan elemen visual digunakan sebagai *key* dan *value*.

Model Transformer memanfaatkan *attention* dengan menggunakan *scaled dot-product* yang efisien tempat dan lebih cepat dari yang lain. Attention ini terdiri dari tiga nilai yakni kunci K, nilai V, dan *query* Q dengan formula (1.1) hingga (1.3):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \qquad (1.1)$$

Dimana Q merupakan matriks hasil dari komputasi *fungsi attention* seluruh *queries.* K dan V merupakan matriks kunci dan nilai yang disatukan menjadi matriks. dk merupakan dimensionalitas *query* dan kunci.

Mekanisme *multi-head self-attention* dalam transformer digunakan untuk mencari nilai relevan berdasarkan kunci dan *query.* Kunci, nilai, dan *query* dapat digunakan sebagai masukan embedding. Digunakan multi-head attention karena single attention tidak cukup dalam mendapatkan aspek-aspek berbeda dari masukan.

$$h_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (1.2)$$

$$H = Concat(h_1, h_2, \dots, h_n)W^O \quad (1.3)$$

Di mana, $W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{model} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{model} \times d_v}$, $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ merupakan masukan multi-head attention. Concat merupakan fungsi concatenation, dan $h_i \in \mathbb{R}^{L \times d_v}$ merupakan keluaran dari fungsi attention scaled-dot product.

Komponen kedua dari *layer* Transformer adalah *point-wise* FC *feed forward* yang diterapkan berdasarkan waktu pada urutan input. Komponen terakhir yakni *positional encoding.* Karena model Transformer tidak memiliki *recurrence* dan tidak ada konvolusi, maka untuk model dapat mengetahui urutan, harus diberikan beberapa informasi tentang posisi relatif atau absolut dari token dalam urutan. Penambahan *positional encoding* diletakkan pada input embedding pada bagian akhir tumpukan *encoder* dan *decoder*.

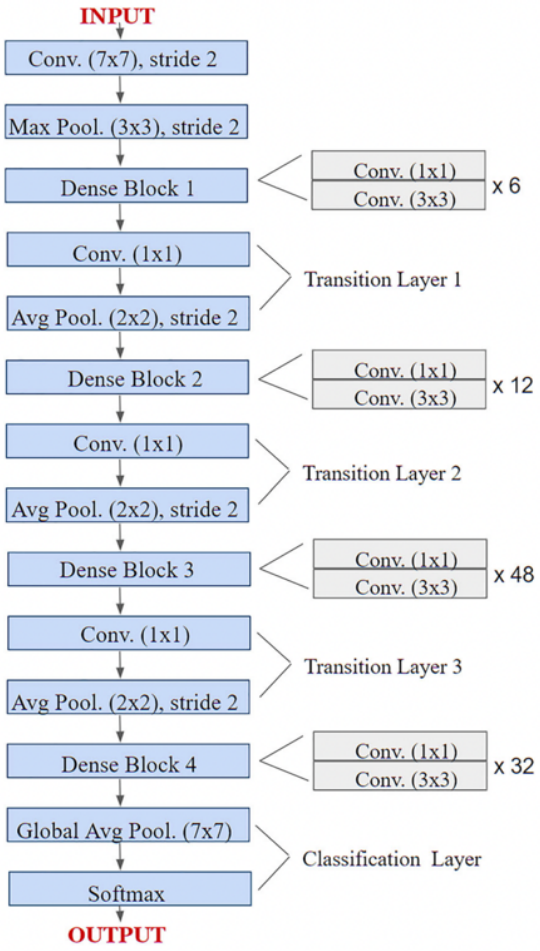
2.1.3 Fitur Ekstraktor

a. DenseNet201

Arsitektur DenseNet (Huang et al., 2016) merupakan salah satu varian Convolutional Neural Networks (CNN) yang menggunakan *dense* di setiap lapisannya, di mana lapisan *dense* merupakan lapisan yang saling terhubung. Setiap lapisan mengambil masukan tambahan dari seluruh lapisan sebelumnya dan kemudian memberikan masukan ke lapisan selanjutnya. Hal ini membuat model mudah dilatih dan sangat efisien secara parametrik karena DenseNet memungkinkan penggunaan kembali fitur dari lapisan yang berbeda sehingga meningkatkan variasi pada masukan lapisan berikutnya dan meningkatkan performa model. Susunan lapisan DenseNet201 dapat dilihat pada Gambar 2.2.

DenseNet memiliki beberapa varian yakni 121, 169, dan 201 dimana angka-angka ini menunjukkan kedalaman dari modelnya. DenseNet201 menunjukkan performa yang mumpuni di penelitian sebelumnya pada dataset ImageNet dan juga CIFAR-100 (Jaiswal et al., 2021). Selain itu, penggunaan DenseNet201 dan Xception sebagai pengekstraksi fitur

gambar juga memberikan hasil akurasi dan f1-score yang menjanjikan pada penelitian *image* captioning penyakit mata sebelumnya (Vellakani & Pushbam, 2020).



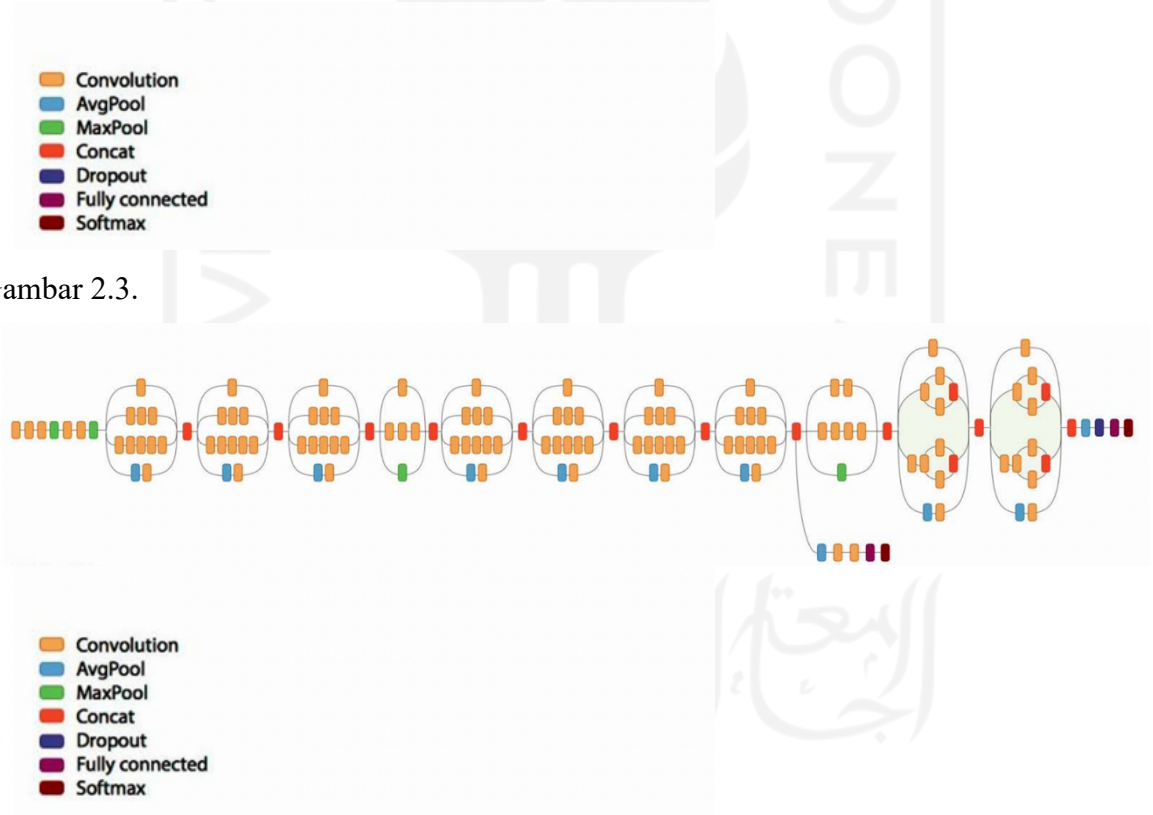
Gambar 2.2 Lapisan Arsitektur DenseNet201 (Jaiswal et al., 2021).

b. ResNet50

ResNet (Residual Network) merupakan model yang menggunakan hubungan residual yang menjumlahkan keluaran dari sebuah blok dari lapisan dengan masukannya dan mengumpankannya sebagai masukan ke lapisan setelahnya (He et al., 2016). ResNet biasa digunakan pada visi komputer untuk menyelesaikan masalah-masalah seperti melakukan klasifikasi gambar, lokalisasi objek, dan mendeteksi objek. ResNet memiliki beberapa varian di antaranya ResNet50, ResNet101, dan ResNet152 di mana setiap angkanya (50, 101, dan 152) menunjukkan jumlah lapisan dari model ResNet. ResNet50 terdiri dari 48 lapisan konvolusi, 1 maxpool dan 1 average pooling. Pada pengujian dengan ImageNet dataset, ResNet50 meraih skor *error* Top-1 yang cukup rendah di antara varian CNN lain dengan nilai 20.47.

c. InceptionResNetV2

InceptionResNet sesuai namanya, merupakan model yang memiliki lapisan CNN yang dalam dengan struktur Inception dengan Residual Network (ResNet) (Szegedy et al., 2016). Namanya sendiri terinspirasi dari film yang disutradarai Christopher Nolan dengan nama dan konsep yang sama "Inception". Arsitektur ini memiliki beberapa unit CNN *Inception* dan diikuti dengan Residual yang berulang-ulang dalam susunannya. Arsitektur ini memiliki lapisan awal yang terdiri dari 3 lapisan konvolusional biasa yang diikuti dengan lapisan maxpooling yang kemudian diikuti kembali dengan 2 lapisan konvolusi dan sebuah lapisan maxpooling. Tahap selanjutnya yakni konvolusi *inception* dimana masukan akan melalui konvolusi secara serentak menggunakan berbagai ukuran filter untuk setiap konvolusinya kemudian melalui *merging* dengan menggabungkan hasilnya bersamaan akan melanjutkannya ke susunan jaringan selanjutnya. Ilustrasi susunan arsitektur model dapat dilihat dari



Gambar 2.3.

Gambar 2.3 Ilustrasi Arsitektur InceptionResNetV2 (Bhatia et al., 2019).

d. InceptionV3

InceptionV3 merupakan varian lain dari Inception yang dikembangkan dan merupakan salah satu model pre-trained *state-of-the-art* (Maeda-Gutiérrez et al., 2020). Model ini tersusun dari 48 lapisan konvolusi yang diciptakan untuk melakukan tugas encoding gambar.

InceptionV3 tersusun dari 11 modul inception yang terdiri dari lapisan konvolusi dan maxpooling. Dalam penggunaannya sebagai ekstraktor fitur gambar, gambar perlu diubah menjadi 299x299 sebelum diberikan kepada model. Karena penggunaan model InceptionV3 diperlukan untuk melakukan ekstraksi fitur, maka *layer* akhir *softmax* untuk mengklasifikasikan gambar dihapus.

e. Xception

Xception (Extreme version of Inception) merupakan varian lain dari Inception yang dikembangkan dan memiliki performa lebih baik dari InceptionV3 (Chollet, 2017). Xception mendapatkan skor yang cukup baik dibandingkan InceptionV3 ketika diujikan pada dataset ImageNet dan jauh melampaui InceptionV3 pada klasifikasi dataset yang memiliki lebih dari 350 juta gambar dan 17 ribu kelas. Model ini dikembangkan dengan parameter yang sama dengan InceptionV3 dengan tujuan efisiensi model parameter. Xception memiliki 71 lapisan dengan metode kedalaman konvolusi yang dapat dimodifikasi. Penelitian sebelumnya mengaplikasikan Xception dan DenseNet201 menunjukkan bahwa dua model yang digunakan mampu memberikan performa terbaik dalam memprediksi teks deskripsi gambar untuk penyakit mata secara benar (Vellakani & Pushbam, 2020).

2.1.4 Maktriks Evaluasi

a. BLEU

BLEU (Papineni et al., 2001) (The Bilingual Evaluation Understudy Score) adalah matriks yang mendefinisikan kemiripan antar teks yang diprediksi dengan referensinya. BLEU mempertimbangkan n-grams (yang biasanya sebanyak 1-4) daripada masing-masing kata, lalu membandingkan okurensi n-gram di deskripsi teks dan pada referensinya. Nilai n-gram yakni 4 karena dalam penemuannya, angka ini memiliki korelasi tertinggi dengan teks yang digenerasi oleh manusia (Nugraha et al., 2019). Dalam mengevaluasi setiap teks, BLEU tidak melihat kebenaran secara sintetik. Jika deskripsi teks yang digenerasi mesin mirip dengan referensi, maka skor yang diberikan adalah 1.0, jika deskripsi teks sama sekali tidak mirip, maka nilai yang diberikan adalah 0.0. Skor BLEU dapat dikalkulasi menggunakan formula (1.4) (Papineni et al., 2001):

$$BLEU = \min\left(1, \frac{output-length}{reference-length}\right)\left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}} \qquad (1.4)$$

b. METEOR

METEOR (Lavie & Agarwal, 2007) (Metric for Evaluation for Translation with Explicit Ordering) merupakan matriks yang berorientasi dengan presisi single dan pemanggilan kata untuk mengatasi kekurangan dari BLEU. Dari perbaikan ini, METEOR menjadi lebih baik dalam mendapatkan korelasi semantic dan semakin relevan dengan penilaian manusia. Matriks METEOR mengkalkulasi akurasi, *recall,* dan f-*score* untuk setiap kata, stem dan sinonim yang sama. Kalkulasi METEOR membutuhkan set jajaran yang sudah didefinisikan, seperti thesaurus WordNet, untuk menilai kata, stem, dan sinonim. Formula yang digunakan oleh METEOR dapat dituliskan pada formula (1.5) hingga (1.9):

$$P = \frac{m}{w_t} \quad (1.5)$$

$$R = \frac{m}{w_r} \quad (1.6)$$

$$F_{mean} = \frac{10PR}{R+9P} \quad (1.7)$$

$$p = 0.5\left(\frac{c}{u_m}\right)^3 \quad (1.8)$$

$$M = F_{mean}(1-p) \quad (1.9)$$

Di mana, P merupakan presisi, R merupakan recall, dan F_{mean} untuk menghitung F score. m merupakan jumlah unigram pada kandidat yang juga ditemukan pada referensi. w_t merupakan jumlah unigram yang ada pada kandidat, w_r merupakan jumlah unigram yang ada pada referensi. p merupakan penalti dengan c sebagai kandidat *chunk* dan u_m sebagai kandidat unigram. Sedangkan M merupakan formula penghitungan METEOR.

c. ROUGE-L

ROUGE (Lin, 2005) (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) merupakan matriks yang membandingkan unit dasar seperti n-gram, urutan kata, dan pasangan kata antara prediksi deskripsi teks dan referensi untuk evaluasi. ROUGE-L merupakan salah satu dari methode untuk evaluasi ROUGE, ROUGE yang lain yakni ROUGE-n (n biasanya bernilai 1,2,3,4, n merepresentasikan jumlah n-gram), ROUGE-W, ROUGE-L, dan skip-bigram (ROUGE-S). Pada penelitian ini, digunakan ROUGE-L yang penilaiannya berdasarkan urutan terpanjang (LCS) pada level kalimat yang tidak membutuhkan kata yang sama secara berurut-urutan. ROUGE-L akan menghitung LCS antara keluaran dan referensi, sehingga dihitung urutan terpanjang yang ada pada keduanya. Formula yang digunakan untuk menghitung ROUGE-L dapat dilihat pada persamaan (1.10) hingga (1.12):

$$recall = \frac{LCS(gram_n)}{count(gram_n)} \quad (1.10)$$

$$precision = \frac{LCS(gram_n)}{count(gram_n)} \quad (1.11)$$

$$F1\ score = 2 \times \frac{precision * recall}{precison + recall} \quad (1.12)$$

d. CIDEr

CIDER (Vedantam et al., 2015) (Consensus-based Image Description Evaluation) mempertimbangkan setiap kalimat terdiri dari n-gram. N-gram ini nantikan akan dienkode dan bobot untuk setiap n-gram dikalkulasi menggunakan *frequency-inverse document frequencies* (TF-IDF) antara deskripsi teks yang diprediksi dengan deskripsi teks referensi untuk mengkalkulasi skor kesamaan *cosine*. CIDEr tidak memperlakukan setiap katanya secara sama seperti BLEU, karena TF dan IDF bekerja untuk saling saling membatasi, sehingga CIDEr hanya akan focus dalam memilih kata penting dan signifikan. Untuk mengevaluasi deskripsi teks yang digenerasi, CIDEr mengubah setiap kata di deskripsi teks prediksi dan kalimat referensi menjadi bentuk akar atau stemnya.

2.2 Kajian Pustaka

Dalam beberapa tahun terakhir, telah banyak *paper image captioning* yang dipublikasikan. Penelitian-penelitian yang diterbitkan kebanyakan menggunakan *deep learning* yang dapat melakukan ekstraksi fitur secara otomatis dari kumpulan data latih. *Deep learning* juga telah banyak diakui kemampuannya dalam menangani data gambar atau video yang besar dan beraneka ragam (Zakir Hossain et al. 2019). Selain itu, deep learning juga bekerja dengan baik dalam mengatasi kompleksitas *image captioning*. Dalam *image captioning*, Convolutional Neural Network (CNN) biasa digunakan sebagai enkoder untuk melakukan ekstraksi fitur, lalu diikuti oleh Recurrent Neural Network (RNN) atau LSTM sebagai decoder untuk mengenerasi keterangan gambar. Hasil deskripsi yang didapatkan dari pendekatan ini lebih ekspresif dengan metode yang dapat diaplikasikan dalam berbagai domain, namun kekurangan dari penggunaan model RNN untuk mengenerasi teks adalah bahwa model tidak memiliki kemampuan untuk menjaga dependensi jangka panjang atara teks yang digenerasi (Nugraha et al., 2019; Stefanini et al., 2021).

Banyak penelitian sebelumnya yang menggunakan enkoder CNN dan dekoder RNN debagai standard untuk melakukan *image captioning.* Penelitian sebelumnya menggunakan dataset berbahasa Cina yang memanfaatkan varian CNN DenseNet berbasis *fuzzy attention* dan LSTM sebagai dekoder yang secara efektif meningkatkan masalah korespondensi antara

fitur gambar dan informasi kontekstual (Lu et al., 2021). Penelitian lain memanfaatkan Efficient Channel Attention (ECA) CNN sebagai enkoder untuk meningkatkan efisiensi penggunaan jaringan CNN dengan memperhatikan *channel* (biasanya memiliki tiga warna RGB; merah, hijau, biru) yang dianggap penting selama konvolusi (Mishra et al., 2022). Penelitian *image captioning* dengan Bahasa Indonesia hingga saat ini masih terbatas juga dengan menggunakan *framework* encoder-decoder CNN (ResNet101, CNN) dan RNN (GRU, LSTM) (Mahadi et al., 2020; Mulyanto et al., 2019; Nugraha et al., 2019). Tabel model yang digunakan pada penelitian-penelitian sebelumnya dapat dilihat pada Tabel 2.1.

Pada tahun 2017, arsitektur baru dengan nama Transformer dikenalkan sebagai salah satu terobosan pada tugas pemahaman bahasa dan secara gampang mendapatkan popularitas karena mampu menyelesaikan masalah model RNN (Vaswani et al., 2017). Transformer merupakan model yang memanfaatkan framework enkoder-dekoder yang menggunakan *attention* (sebuah konsep untuk meningkatkan performa mesin translasi) untuk membantu meningkatkan kecepatan model dalam melatih model. Model ini telah banyak diadopsi oleh banyak peneliti di bidang *image captioning* untuk mendapatkan deskripsi terbaik pada gambar yang diberikan.

Namun begitu, model transformer yang diciptakan bertujuan untuk diterapkan dalam tugas mesin translasi yang menggunakan masukan dan memberikan keluaran teks. Oleh karena itu, dalam tugas *image captioning,* dilakukan pengubahan masukan yang digunakan dengan gambar sebagai *input*. Dari gambar tersebut, akan dilakukan ekstraksi fitur sebagai masukan yang akan digunakan pada model Transformer. Beberapa riset sebelumnya memanfaatkan Transformer yang mampu meningkatkan encoding gambar dan teks yang digenerasi menggunakan Meshed Transformer with Memory (M2M) untuk mendapatkan fitur level rendah dan tinggi yang membantu dalam memprediksi keterangan gambar (Cornia et al., 2020). Penelitian lain dilakukan oleh (J. Li et al., 2019) yang membuat Boosted Transformer dan memanfaatkan Concept-Guided Attention (CGA) untuk membantu memanfaatkan informasi bantuan konsep semantik untuk mendapatkan fitur visual yang kuat serta Visual-Guided Attention (VGA) untuk mendapatkan informasi visual sekuensial sehingga dapat meningkatkan kemampuan model untuk memprediksi deskripsi gambar.

Beberapa penelitian lain mencoba menggabungkan penggunaan Transformer dengan model *pre-trained* CNN. Penelitian sebelumnya pada dataset deskripsi dengan *personality captions* menggunakan beberapa varian ResNet seperti ResNet152 dan ResNeXt-IG-3.5B sebagai fitur ekstraktor sebelum diberikan kepada model Transformer untuk mengenerasi deskripsi gambar (Shuster et al., 2019). Dari hasil penelitian tersebut, didapatkan hasil yang

tinggi pada dataset Flickr 30k dengan menggunakan model ResNeXt-IG-3.5B pada pengujian Karpathy. Penelitian lain juga menggunakan fitur ekstraktor varian ResNet yakni gabungan Inception-ResNetV2 yang memberikan manfaat optimasi melalui koneksi residual dan unit Inception yang efisien secara komputasi (P. Sharma et al., 2018). Model juga membandingkan dekoder dengan LSTM dan Transformer dengan model Transformer yang memiliki performa paling baik.

Tabel 2.1 Perbandingan model yang digunakan dalam penelitian *Image Captioning*.

Judul Penelitian	Penulis	Model
Adaptive Attention Generation for Indonesian Image Captioning	(Mahadi et al., 2020)	ResNet101-LSTM
Generating image description on Indonesian language using convolutional neural network and gated recurrent unit	(Nugraha et al., 2019)	CNN-GRU
Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset	(Mulyanto et al., 2019)	CNN-LSTM
Chinese Image Captioning via Fuzzy Attention-based DenseNet-BiLSTM	(Lu et al., 2021)	Fuzzy Attention-based DenseNet-BiLSTM
Efficient Channel Attention Based Encoder–Decoder Approach for Image Captioning in Hindi	(Mishra et al., 2022)	ECA-NET CNN-GRU
Meshed-memory transformer for image captioning	(Cornia et al., 2020b)	Meshed-memory transformer
Boosted transformer for image captioning	(J. Li et al., 2019)	Boosted transformer
Engaging image captioning via personality	(Shuster et al., 2019)	TransResNet
Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning	(P. Sharma et al., 2018)	InceptionResnetV2 -Transformer

Beberapa penelitian yang telah dilakukan tidak hanya terbatas pada penggunaan dataset MSCOCO(X. Chen et al., 2015) maupun Flickr (Plummer et al., 2015) yang umumnya digunakan sebagai dataset dalam penelitian *image captioning.* Penelitian sebelumnya memanfaatkan *personality-captions* image captioning (Shuster et al., 2019) yang didapatkan dari YFCC100M dataset kemudian digolongkan menjadi 215 sifat (misalkan, manis, khawatir, dramatis, simpatik, arogan, optimis, dlsb) dan menggunakan metode TransResNet-LSTM untuk membedakan sifat deskripsi gambar untuk menghasilkan *captions* yang mendekati kalimat manusia. Pada penelitian (P. Sharma et al., 2018), kombinasi Inception-ResNetv2 untuk mengekstraksi gambar dan model Transformer untuk melakukan generasi Bahasa mampu mendapatkan hasil yang bagus pada dataset dengan *captions* konseptual (dataset yang dikembangkan untuk merepresentasikan variasi yang banyak pada gambar dan gaya keterangan). Penelitian *image captioning* dengan Bahasa Cina memanfaatkan dataset yang disediakan oleh AI Challenger dengan 300,000 dataset gambar dan 1.5 juta deskripsi yang ditulis dalam Bahasa Cina (Lu et al., 2021). Namun begitu, peneltian *image captioning* umumnya masih menggunakan dataset MSCOCO dan Flickr

yang sudah popular digunakan (Zakir Hossain et al., 2019). Beberapa penelitian dengan Bahasa Indonesia juga hingga saat ini masih banyak menggunakan MSCOCO dan Flickr sebagai dataset masukan maupun data uji(Mahadi et al., 2020; Mulyanto et al., 2019; Nugraha et al., 2019).

Tabel 2.2 Bahasa dan dataset yang digunakan dalam penelitian *Image Captioning*.

Judul Penelitian	Penulis	Bahasa	Dataset
Adaptive Attention Generation for Indonesian Image Captioning	(Mahadi et al., 2020)	Indonesia	MSCOCO & Flickr30k
Generating image description on Indonesian language using convolutional neural network and gated recurrent unit	(Nugraha et al., 2019)	Indonesia	Flickr30k
Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset	(Mulyanto et al., 2019)	Indonesia	FEE-ID
Chinese Image Captioning via Fuzzy Attention-based DenseNet-BiLSTM	(Lu et al., 2021)	Cina	*Captions* disediakan AI Challenger (Wu et al., 2017)
Efficient Channel Attention Based Encoder–Decoder Approach for Image Captioning in Hindi	(Mishra et al., 2022)	Hindi	MSCOCO
Meshed-memory transformer for image captioning	(Cornia et al., 2020b)	Inggris	MSCOCO
Boosted transformer for image captioning	(J. Li et al., 2019)	Inggris	MSCOCO
Engaging image captioning via personality	(Shuster et al., 2019)	Inggris	YFCC100M Dataset
Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning	(P. Sharma et al., 2018)	Inggris	Conceptual captions

Penelitian *image captioning* sebelumnya banyak ditemukan dengan menggunakan dataset berbahasa inggris dan hasil yang didapatkan juga sangat baik. Beberapa penelitian *image captioning* juga sudah banyak diaplikasikan ke dalam berbagai Bahasa berbeda yakni Cina dengan memanfaatkan model DenseNet-BiLSTM berbasis *attention* (Lu et al., 2021) dan India dengan memanfaatkan model enkoder-dekoder berbasis *attention* (Mishra et al., 2022). Namun begitu, penelitian *image captioning* dengan Bahasa Indonesia saat ini masih terbatas dikarenakan dataset berbahasa Indonesia yang masih sulit untuk ditemui. Beberapa penelitian yang dilakukan sebelumnya masih menggunakan dataset berbahasa inggris yang kemudian ditranslasikan ke dalam bahasa indonesia dengan memanfaatkan mesin terjemahan google dan mesin terjemahan profesional inggris-indonesia (Mahadi et al., 2020;

Mulyanto et al., 2019; Nugraha et al., 2019). **Error! Reference source not found.** menampilkan ringkasan perbandingan Bahasa dan dataset yang digunakan oleh penelitian-penelitian *image captioning* sebelumnya.

Tabel 2.3 Perbandingan matriks evaluasi yang digunakan dalam penelitian *Image Captioning*.

Judul Penelitian	Penulis	Matriks Evaluasi
Adaptive Attention Generation for Indonesian Image Captioning	(Mahadi et al., 2020)	BLEU-1, BLEU-2, BLEU-3, BLEU-4, dan CIDEr
Generating image description on Indonesian language using convolutional neural network and gated recurrent unit	(Nugraha et al., 2019)	BLEU-1, BLEU-2, BLEU-3, BLEU-4
Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset	(Mulyanto et al., 2019)	BLEU-1, BLEU-2, BLEU-3, BLEU-4
Chinese Image Captioning via Fuzzy Attention-based DenseNet-BiLSTM	(Lu et al., 2021)	BLEU-1, BLEU-4, METEOR, ROUGE-L, dan CIDEr.
Efficient Channel Attention Based Encoder–Decoder Approach for Image Captioning in Hindi	(Mishra et al., 2022)	BLEU-1, BLEU-2, BLEU-3, dan BLEU-4
Meshed-memory transformer for image captioning	(Cornia et al., 2020b)	BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGEL, CIDEr, METEOR, dan SPICE
Boosted transformer for image captioning	(J. Li et al., 2019)	BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGEL, CIDEr, METEOR, dan SPICE
Engaging image captioning via personality	(Shuster et al., 2019)	BLE-1, BLEU-4, ROUGE-L, CIDEr, dan SPICE
Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning	(P. Sharma et al., 2018)	CIDEr, ROUGE-L, dan METEOR

Matriks berbeda digunakan oleh penelitian-penelitian sebelumnya untuk mengevaluasi performa dari model *image captioning* yang telah dibandung. Dapat dilihat pada

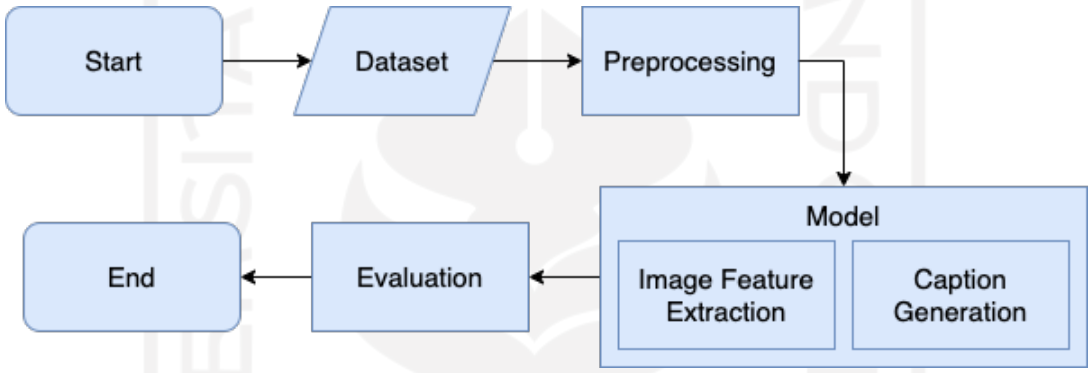
, dari perbandingan matriks evaluasi tersebut, dapat kita rangkum bahwa matriks evaluasi yang digunakan yakni BLEU-1, BLEU-2. BLEU-3, BLEU-4, CIDER, METEOR, ROUGE-L dan SPICE. BLEU-n (n-gram, biasanya n bernilai 1-4) mengevaluasi deskripsi gambar dari ukuran Panjang frase deskripsi yang digenerasi model dengan deskripsi asli gambar. METEOR mempertimbangkan *recall* dan presisi serta melihat korelasi semantik dalam level kata untuk menghitung skor evaluasi. ROUGE-L menghitung skor dengan melihat urutan kata terpanjang yang sama antara deskripsi referensi dan deskripsi yang digenerasi untuk menilai kecakapan bahasanya. CIDEr menilai dari makna setiap kata sedangkan SPICE menganalisa deskripsi yang digenerasi secara semantik.

Penelitian ini bertujuan untuk mengenerasi deskripsi teks dari gambar untuk mendapatkan pemahaman visual di dalam ruangan. Kontribusi utama penelitian ini yakni mempresentasikan evaluasi aristektur Transformer pada dataset gambar dengan *caption* berbahasa Indonesia yang berbeda dari dataset yang biasa digunakan seperti MSCOCO (X. Chen et al., 2015) atau Flickr30k (Plummer et al., 2015). *Caption* asli yang didapatkan dari MSCOCO selanjutnya dihapus dan diganti dengan keterangan gambar baru dengan deskripsi yang menggambarkan nama objek, posisi/lokasi (berdasarkan sudut pandang pengguna), karakteristik, dan objek di sekitarnya. Penelitian ini mengajukan sebuah arsitektur *deep learning* menggunakan model Transformer untuk melakukan *image captioning* untuk mendapatkan pemahaman visual di dalam ruangan. Penelitian ini juga bereksperimen dengan beberapa model fitur ekstraksi gambar serta mengubah dan menyesuaikan model (*fine tuning*) untuk mendapatkan model terbaik dalam mengenerasi deskripsi gambar.

BAB 3
Metodologi

3.1 Langkah Penelitian

Pada bab ini akan diuraikan metodologi yang digunakan dalam melakukan *image captioning* menggunakan Transformer. Terdapat 5 tahap yang digunakan yakni, pengumpulan dataset, preprocessing pada gambar dan *captions,* ekstraksi fitur, generasi *captions,* dan evaluasi model. *Flowchart* tahap *image captioning* Bahasa Indonesia dengan Transformer dapat dilihat pada Gambar 3.1.



Gambar 3.1 Tahap Indonesian Image Captioning (Nugraha et al., 2019).

3.2 Uraian Metodologi

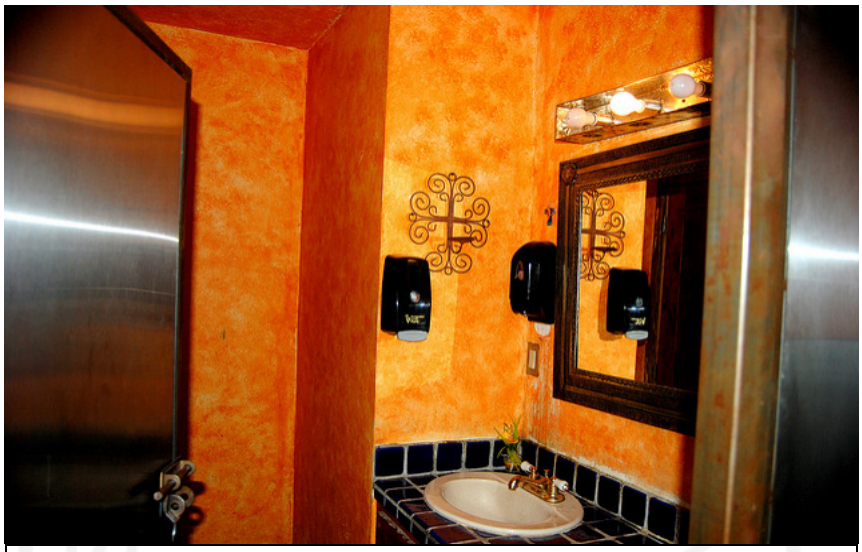
Pada bagian ini akan diuraikan langkah penelitian *image captioning* sesuai dari *flowchart* yang terdapat pada Gambar 3.1 dari pengumpulan data, *preprocessing* data, ekstraksi fitur, mengenerasi *captions* hingga evaluasi model. Pada tahap pertama, dilakukan pengoleksian data dari dataset yang berukuran besar MSCOCO dan melakukan *scraping* beberapa gambar dari Flickr. Kemudian dilakukan penambahkan *captions* baru yang telah dimodifikasi untuk setiap gambarnya. Tahap selanjutnya yakni *preprocessing* pada teks dan gambar sebelum memberikannya ke model image captioning. Tahap ke-tiga yakni ekstraksi fitur gambar, kemudian diikuti tahap ke-empat dengan memberikan set latih ke model transformer. Tahap terakhir yakni evaluasi dengan menggunakan beberapa matriks evaluasi untuk mengevaluasi model image captioning.

3.2.1 Pengumpulan Data

Data yang digunakan untuk penelitian *image captioning* didapatkan dari MSCOCO dan data yang dikumpulkan dari *website* Flickr. Digunakan 10 objek dalam ruangan dengan setiap objek memiliki 60-80 gambar. Objek yang digunakan antara lain tempat tidur, wastafel, kursi, sofa, meja, televisi, kulkas, tanaman rumah, oven, tangga, lampu gantung, rak gantung, pintu dan jendela, mesin cuci, dan telepon genggam. Objek yang digunakan merupakan objek yang diambil di dalam ruangan karena studi sebelumnya menunjukkan bahwa tunanetra memanfaatkan waktu mereka sebanyak 80%-90% di dalam ruangan (Jeamwatthanachai et al., 2019). Oleh karena itu, model *image captioning* yang dibangun dapat berkontribusi dalam membantu tunanetra untuk mendapatkan pemahaman visual tentang ruangan sekitar mereka secara lebih baik melalui prediksi deskripsi yang didapatkan dari *image captioning* yang kemudian dapat diterjemahkan ke dalam ucapan.

Dalam pemberian *captions*, *captions* yang didapatkan dari MSCOCO dihapus dan memberikan keterangan deskripsi dari gambar yang sesuai dengan tujuan penelitian ini, yakni memberikan pemahaman visual. Sehingga, alih-alih menggunakan teks MSCOCO yang tersedia, digunakan teks yang dibuat dalam Bahasa Indonesia yang meliputi nama objek, warna, posisi/lokasi (sudut pandang pengguna), karakteristik, dan objek sekitarnya. Setiap gambar diberikan 5 keterangan berbeda dengan meniru cara orang mendeskripsikan gambar secara berbeda pula. Contoh gambar dan teks deskripsi yang diberikan dapat dilihat pada Gambar 3.2.

Dalam pembuatan *captions* ini, penelitian ini terinspirasi dari penelitian sebelumnya untuk membuat beberapa aturan dalam mendeskripsikan gambar (X. Chen et al., 2015). (1) Karena tujuan penelitian ini adalah untuk menggambarkan ruangan untuk mendapatkan pemahaman visual, sehingga perlu menambahkan informasi lokasi setiap objek apakah objek terletak di sisi kiri/kanan/depan ruangan dan informasi benda-benda di sekitarnya. (2) Hanya mendeskripsikan bagian utama dari gambar dengan menggambarkan objek utama yang terlihat. (3) Dalam mendeskripsikan objek dalam tampilan, disebutkan pula warna objek dan karakteristik mereka karena dapat bermanfaat untuk membantu membedakan setiap objek (Rashid et al., 2017).



‘Di sebelah kanan ruangan terdapat wastafel berwarna putih dengan pot tanaman berukuran kecil di pojok wastafel.’ ‘Kaca cermin menggantung di dinding atas wastafel sebelah kanan.’ ‘Di sisi kiri ada pintu dari besi anti karat.’ ‘Di depan terdapat ruang toilet dengan wastafel dan cermin.’ ‘Di depan merupakan ruang toilet dengan pintu terbuka.’
‘a bathroom with some orange walls, and a black and white counter’ ‘a bathroom sink with mirror are shown in this picture.’ ‘a bathroom with a tiled counter, a sink and a soap dispenser.’ ‘a bathroom with orange walls and black tile features.’ ‘a bathroom area with a sink and tiled counter top.’

Gambar 3.2 Contoh dataset gambar yang digunakan beserta deskripsi gambar dengan Bahasa Indonesia (atas) dan deskripsi asli MSCOCO (bawah).

Data yang dikoleksi dari MSCOCO dan Flickr dibagi menjadi 15 objek *indoor* yang berbeda. 15 objek yang digunakan yakni tempat tidur, wastafel, kursi, sofa, meja, televisi, kulkas, tanaman rumah, oven, tangga, lampu gantung, rak gantung, pintu dan jendela, mesin cuci, dan telepon genggam di mana setiap objek terdiri dari 60 hingga 70 objek karena keterbatasan data gambar. Total gambar yang didapatkan berjumlah 1072 dan teks deskripsi berjumlah 5360. Dataset kemudian dibagi secara random menjadi dua dataset yakni data latih dan data uji. Berdasarkan studi sebelumnya, diperlukan untuk membagi dataset data latih sebesar 80% dan dataset uji sebesar 20% untuk mendapatkan hasil yang bagus (Gholamy et al., 2018). Data uji terdiri dari 214 gambar dan 1070 *captions* sedangkan data latih terdiri dari 858 gambar dan 4290 *captions*.

3.2.2 Preprocessing

Setelah didapatkan data yang digunakan untuk melakukan *image captioning*, perlu dilakukan *preprocessing* sebelum memberikannya sebagai masukan pada model yang digunakan. Pada tahap ini, dilakukan *preprocessing* pada data gambar dan teks deskripsinya. Pada data gambar hanya dilakukan pengubahan ukuran gambar dengan menyesuaikan model fitur ekstraktor yang digunakan, yakni DenseNet201, ResNet50, InceptionResNetV2, InceptionV3, dan Xception.

Kata yang ada pada *captions* akan disimpan sebagai kata unik yang kemudian akan digunakan sebagai indeks kata dalam kosa kata *image captioning* Bahasa Indonesia. Pada tahap ini, perlu dilakukan *preprocessing* pada teks deskripsi dengan menggunakan beberapa langkah berikut, yakni:

1) Mengubah kalimat menjadi *lower case,*

2) Menghapus *punctuation* atau tanda baca seperti !"#$%&()*+.,-/:;=?@[\]^_`{|}~ ',

3) Menghapus karakter yang memiliki kata kurang dari dua karakter,

4) Menghapus numerik,

5) Menambahkan tagar <start> dan <end> untuk menandai awal dan akhir kalimat,

6) Melakukan tokenisasi kalimat sehingga menjadi kata individual,

7) Menambahkan padding 0 pada vektor yang memiliki jumlah kurang dari jumlah kata yang ditetapkan, dan menghapus bagian akhir vektor jika jumlah kata kalimat lebih dari jumlah kata yang ditetapkan

3.2.3 Ekstraksi Fitur

Model *image captioning* dibagi menjadi dua bagian utama, yakni melakukan ekstaksi fitur gambar dan menghasilkan deskripsi dalam Bahasa alami menggunakan masukan dari fitur gambar. Pada tahap pertama, dilakukan ekstraksi fitur dari gambar menggunakan beberapa model pre-trained yakni DenseNet201, ResNet50, InceptionResNetV2, InceptionV3, dan Xception. Ekstraksi fitur didapatkan dari lapisan terakhir sebelum lapisan dense *fully connected,* mengikuti penelitian-penelitian *image captioning* sebelumnya. Hal ini dilakukan untuk mendapatkan informasi mengenai objek-objek yang terdapat dalam gambar dan hubungan di antara objeknya. Fitur yang sudah diekstraksi kemudian disimpan dalam .npy file.

3.2.4 Generasi Deskripsi Gambar

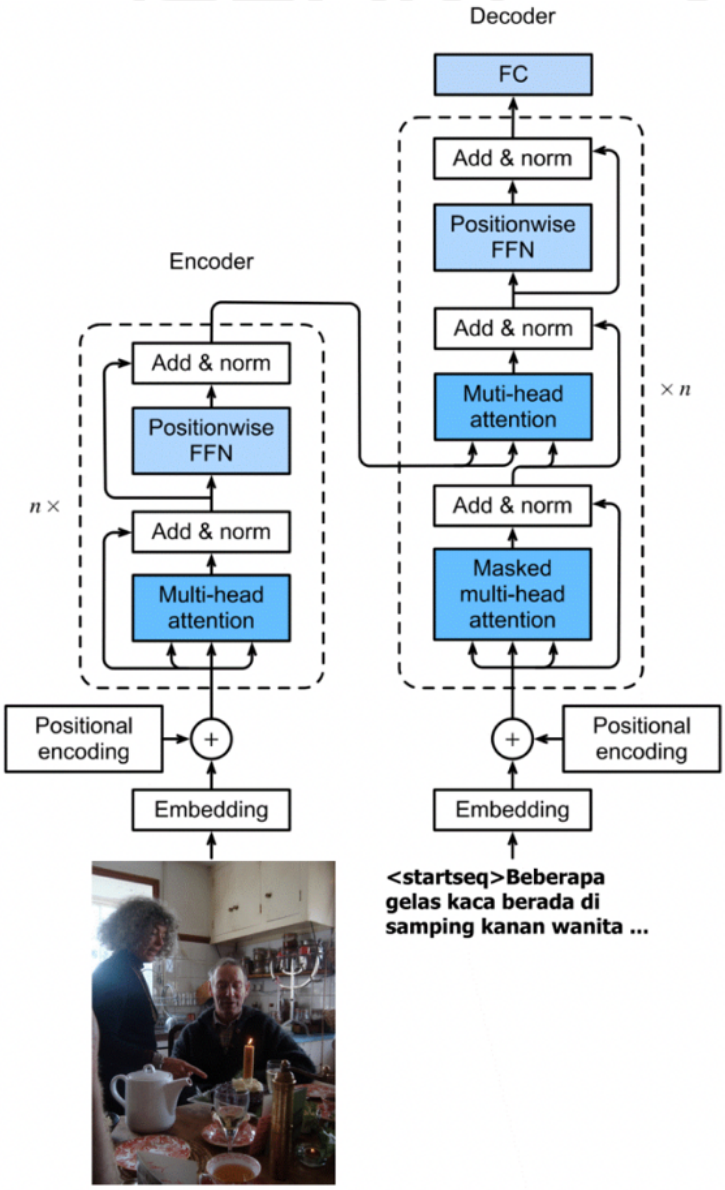
Dalam penelitian ini, arsitektur Transformer digunakan dengan mengikuti paper aslinya tanpa modifikasi model arsitektur yang signifikan. Penelitian ini menyempurnakan model dengan mengatur *hyper-parameter* pada model dengan nilai terbaik untuk mendapatkan hasil terbaik dalam memberikan teks deskripsi gambar. Untuk mendapatkan hasil terbaik, dilakukan eksperimen dengan mengubah ukuran *batch*, *attention heads*, dan *dropout*. *Dropout* dipilih sebagai salah satu *hyper-parameter* yang diubah karena pengaturan *dropout* mampu mengurangi *overfitting* model (Srivastava et al., 2014). Ukuran *batch* akan mengatur jumlah gambar yang digunakan ketika proses estimasi gradien dan mengontrol akurasi dari estimasi *gradient error* ketika pelatihan (Kandel & Castelli, 2020). Sedangkan *attention heads* berperan penting dalam memberikan pembobotan setiap kata. *Attention heads* memiliki fungsi masing-masing dalam memberikan pembobotan, diantaranya yakni mempertimbangkan kata yang berdekatan maupun kata yang memiliki relasi dependensi secara semantik (Voita et al., 2019). Selain itu, *attention heads* juga memperhatikan kata dilihat dari posisinya, kata yang jarang muncul, maupun hubungan sintaktis tertentu.

Untuk ukuran *batch*, Transformer biasanya menggunakan ukuran tipikal |Bk| {32,64,· · · ,512} (Keskar et al., 2017). *Dropout* juga diterapkan pada layer Transformer untuk mengurangi *over-fitting*. Nilai *dropout* yang digunakan berkisar antara 1,0 sampai 0,0, dimana 1,0 berarti tanpa *dropout*, dan sebaliknya, semakin rendah nilai *dropout* yang digunakan, semakin banyak *dropout* yang diaplikasikan (Srivastava et al., 2014). Penelitian Vaswani menggunakan *dropout* p= 0.1, dan juga bereksperimen menggunakan p=0.2 dan p=0.3 untuk model Transformer besar (Vaswani et al., 2017). Semua model Transformer berjalan dalam 40 epoch karena keterbatasan komputasi dan menggunakan *sparse categorical* sebagai fungsi loss. Model menggunakan *Adam optimizer* dengan β1=0.9 dan β2=0.98. *Learning rate* yang digunakan selama *training* model Transformer bervariasi mengikuti rumus yang digunakan pada paper aslinya dengan menaikkan dan menurunkan nilai *learning rate*.

Tabel 3.1 Rentang *hyperparameter* yang digunakan pada model Transformer.

Hyperparameter	Rentang
Attention Heads	8-16
Batch Size	32-128
Dropout	0.1-0.2

Modifikasi yang dilakukan pada model mengikuti rentang *hyperparameter* yang dapat dilihat pada Tabel 3.1. Selain tiga *hyperparameter* yang sudah disebutkan pada table, *hyperparameter* lain tidak dilakukan perubahan dan mengikuti penelitian aslinya. Untuk semua model, digunakan dimensionality d pada semua lapisan dengan nilai 512 dan dimensionalitas pada lapisan *feed-forward* bagian dalam df diberikan nilai 2048. Masukan kata pada model direpresentasikan menggunakan vektor *one-hot* dan diproyeksikan secara linear pada dimensionalitas masukan pada model d.



Gambar 3.3 Ilustrasi alur penggunaan Transformer untuk *image captioning.*

3.2.5 Pelatihan

Pada fase pelatihan, model menggunakan dua masukan, yakni masukan gambar dan masukan bahasa. Gambar akan digunakan sebagai masukan ke dalam model *pre-trained* dengan lapisan terakhir *softmax* yang dihapus (karena tidak diperlukan untuk mengklasifikasi) untuk mendapatkan fitur gambar yang sudah diekstraksi. Masukan Bahasa yang telah melalui *preprocessing* akan menjadi urutan indeks token di mana pada tahap pengujian, masukan berupa gambar dan indeks token sebelumnya akan menghasilkan indeks token berikutnya, begitu seterusnya hingga dihasilkan kata terakhir.

Pada penelitian ini, digunakan beberapa varian CNN yang dilatih menggunakan data ImageNet untuk mengambil fiturnya. Sedangkan model transformer digunakan untuk mengenerasi teks deskripsi gambar dari mendekode masukan fitur gambar dengan menggunakan mekanisme attention yang tersusun. Dalam tahap decoding fitur gambar, perlu melalui beberapa lapisan Transformer. Lapisan-lapisan dalam Transformer tersusun atas beberapa jumlah lapisan bertumpuk yang sama, dimana setiap lapisan dibagi menjadi sub lapisan (Lihat Gambar 3.3). Lapisan pertama yakni mekanisme *multi-headed self-attention* di mana didalamnya terdapat mekasime *mask attention* yang membuat model mengenerasi kata hanya bergantung pada kata berikutnya, karena *mask attention* berperan dalam membatasi model untuk melihat informasi yang akan datang. Lapisan ke-dua adalah *multi-headed attention* yang menghubungkan informasi Bahasa dan gambar menggunakan mekanisme *attention.* Lapisan ke-tiga adalah lapisan FC yang diikuti oleh lapisan normalisasi.

3.2.6 Matriks Evaluasi

Untuk evaluasi, digunakan 4 matriks evaluasi yang berbeda untuk mengevaluasi model *image captioning.* Matriks evaluasi yang digunakan yakni BLEU-n, ROUGE-L, METEOR dan CIDEr yang sudah biasa digunakan untuk mengevaluasi tugas *image captioning.* Dalam mengevaluasi deskripsi teks yang tergenerasi, istilah kandidat dan referensi digunakan. Kandidat yakni deskripsi teks yang digenerasi oleh model dan referensi merupakan deskripsi teks yang dianotasi oleh manusia. Matriks evaluasi bekerja dengan membandingkan kedekatan antara kandidat dengan referensi sesuai kalimat yang dibuat manusia atau benar secara semantik (H. Sharma et al., 2020). Semakin tinggi nilai skor, semakin prediksi deskripsi teks mendekati deskripsi teks yang asli.

BAB 4
Hasil dan Pembahasan

4.1 Eksperimen

Pada penelitian ini digunakan beberapa varian CNN untuk mengekstraksi fitur pada gambar. Varian-varian CNN yang digunakan yakni DenseNet201, ResNet50, IncepResNetV2, InceptionV3, dan Xception yang sebelumnya telah dilatih menggunakan data ImageNet (data besar yang dikembangkan untuk membantu menyelesaikan masalah visi komputer dengan data berjumlah 14 juta).

Dalam melatih model digunakan data berjumlah 1072 dan teks deskripsi berjumlah 5360 yang kemudian dibagi menjadi dua dataset; dataset latih dan dataset uji. Kedua dataset memiliki perbandingan 8:2 dengan data latih terdiri dari 858 gambar dan 4290 *captions* sedangkan dataset uji berjumlah 214 gambar dan 1070 *captions.* Semua dataset merupakan gambar yang diambil di dalam ruangan dengan memilih 15 objek yang diambil dari MSCOCO dan Flickr. Dataset ini akan digunakan sebagai data masukan pada model-model yang sudah disiapkan. Kami sebelumnya telah menunjukkan tabel dengan rentang hyperparameter yang diubah untuk mendapatkan hasil terbaik (Lihat Tabel 3.1). Pada Tabel 4.1 ditampilkan model-model beserta hyperparameter yang diubah untuk setiap modelnya.

Tabel 4.1 Pengaturan hyper-parameter untuk setiap model.

Number	Feature Extractor	Batch Size	Dropout	Number of Attention Heads
1	DenseNet201	128	0.2	4
2	DenseNet201	32	0.1	8
3	ResNet50	128	0.2	4
4	ResNet50	64	0.1	16
5	InceptionResNetV2	128	0.1	4
6	InceptionResNetV2	32	0.1	4
7	InceptionResNetV2	64	0.1	4
8	InceptionResNetV2	128	0.1	8
9	InceptionResNetV2	128	0.1	16
10	InceptionV3	128	0.2	4
11	InceptionV3	32	0.1	4
12	Xception	128	0.2	4
13	Xception	64	0.1	8

4.2 Skor Evaluasi Model

Pada Tabel Tabel 4.2 ditampilkan skor BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, CIDEr, dan ROUGE-L yang didapatkan setiap model. Dapat dilihat pada Tabel tersebut, bahwa skor yang didapat Model 1 dengan menggunakan pengekstraksi fitur InceptionResNetV2 lebih tinggi dibandingkan dari model-model yang lain dengan skor BLEU-1 0.6971, BLEU-2 0.5246, BLEU-3, 0.3921, BLEU-4 0.2831, METEOR 0.2468, CIDEr 0.4801, dan ROUGE-L 0.5114.

Selanjutnya, pengubahan nilai *hyperparameter* dilakukan pada *feature extractor* InceptionResNetV2. Dari hasil skor yang didapatkan, dapat dilihat bahwa nilai yang didapatkan Model 5 dengan *feature extractor* IncepResNetV2 mendapatkan nilai tertinggi dibandingkan Model 6, Model 7, Model 8, dan Model 9 yang juga menggunakan model ekstraksi fitur yang sama. Pada fitur ekstraktor ini, dilakukan eksperimen dengan mengubah nilai ukuran batch dengan nilai 32, 64, dan 128, atau nilai *attention heads* dengan nilai 4, 8, dan 16. Pada Model 5, Model 6, dan Model 7 hanya dilakukan pengubahan ukuran *batch* dengan pengaturan *dropout* 0.1 dan *attention heads* 4. Dari pengubahan ini, model dengan pengaturan *batch* semakin tinggi, memiliki skor matriks evaluasi yang semakin tinggi pula. Sedangkan pada Model 5, Model 8, dan Model 9 hanya dilakukan pengubahan nilai *attention heads* dengan *batch size* 128 dan *dropout* 0.1. Dari pengubahan ini, model dengan pengaturan *attention heads* yang semakin kecil, semakin tinggi skor yang didapatkan pada semua matriks evaluasi.

Tabel 4.2 Hasil skor matriks evaluasi untuk setiap model.

No	*Feature Extractor*	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L
1	DenseNet201	0.4611	0.2851	0.1775	0.0993	0.1667	0.1200	0.3450
2		0.4706	0.2831	0.1686	0.0756	0.1686	0.0806	0.3408
3	ResNet50	0.4715	0.2894	0.1703	0.0790	0.1680	0.0991	0.3475
4		0.4591	0.2635	0.1456	0.0673	0.1586	0.0838	0.3286
5	InceptionResNetV2	**0.6971**	**0.5246**	**0.3921**	**0.2831**	**0.2468**	**0.4801**	**0.5114**
6		0.6026	0.4280	0.3052	0.2067	0.2218	0.3114	0.4319
7		0.6530	0.4824	0.3579	0.2602	0.2433	0.4512	0.4650
8		0.6519	0.4678	0.3367	0.2347	0.2309	0.4293	0.4523
9		0.6424	0.4592	0.3281	0.2334	0.2365	0.4522	0.4269
10	InceptionV3	0.6503	0.4860	0.3634	0.2667	0.2425	0.4281	0.4654
11		0.6084	0.4245	0.2923	0.1971	0.2152	0.3241	0.4269
12	Xception	0.6431	0.4592	0.3243	0.2255	0.2246	0.3734	0.4412
13		0.6181	0.4370	0.3100	0.2174	0.2278	0.3275	0.4401

4.3 Hasil Prediksi *Image Captioning*

Hasil prediksi deskripsi gambar dapat dilihat pada Tabel 4.3. Dari tabel tersebut model-model IncepResNetV2 mampu mengenerasi lebih banyak deskripsi yang sesuai dengan konteks yang diberikan di gambar dibandingkan InceptionV3, Xception. Pada tabel Tabel 4.3 Gambar#1 model IncepResNetV2 mampu mengenerasi deskripsi “di atas meja”, “pizza”, “gelas kaca”, dan “botol minuman”. Hal ini menunjukkan bahwa model mampu mengenali konteks gambar yang diberikan dan mengenali objek-objek yang ada di ruangan tersebut. Pada prediksi ini terdapat kesalahan dalam mengenali tempat bumbu dari gelas yang terdapat di atas meja. Prediksi seperti “gelas kaca” dan “botol minuman” dapat terjadi dikarenakan kebanyakan dataset yang digunakan dalam penelitian ini, gambar dengan makanan biasanya juga terdapat minuman di dalamnya.

Pada Gambar#2 dan Gambar#3 model-model IncepResNetV2 juga dengan baik mengenali objek-objek yang terdapat di kamar mandi dan ruang seni. Model mampu mengenali beberapa objek seperti “rak buku”, “sofa”, dan “pintu” yang sesuai dengan gambar yang diberikan. Namun begitu, pada Gambar#3, Model IncepResNetV2 masih kesulitan untuk mengenali karakteristik objek dengan benar (model#5: rak buku rendah) dan lokasi (model#6 sofa di bagian kanan).

Dari gambar-gambar ini, dapat diketahui bahwa model IncepResNetV2 mampu mengenali beberapa karakteristik seperti pada Gambar#1 yakni (warna meja berwarna biru dan karakteristik gelas), Gambar#2 mampu mendapatkan (cermin berukuran besar dan handuk berwarna putih), dan Gambar#3 meskipun terdapat beberapa kesalahan dalam mendapatkan karakteristik objek yang benar, beberapa model IncepResNetV2 mampu mengenerasi karakteristik objek lain dengan benar (pintu berwarna coklat, sofa berwarna hitam, dan rak berukuran tinggi).

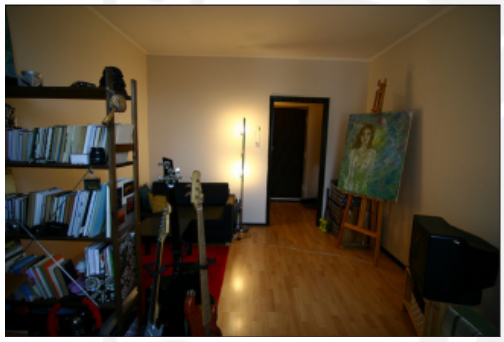
Model Inception dan juga Xception juga cukup baik dalam mengenerasi deskripsi gambar meskipun masih terdapat beberapa kesalahan prediksi dan tidak sebaik mengenali objek seperti IncepResNetV2. Pada Gambar#2 Inception dan Xception mampu mengenali ruangan dan objek-objeknya dengan baik, namun pada Gambar#1 dan Gambar#3 kedua model mengenali beberapa objek yang tidak ada di ruangan. Pada Gambar#1 model mengenerasi objek manusia di dalam ruangan yakni model#11 yang mengenerasi seorang wanita menggunakan ponsel lipat dan model#12 yang mengenerasi “banyak orang” yang tidak sesuai dengan gambar yang diberikan. Sedangkan pada Gambar#3, model#12 mengenali objek seorang anak yang tidak terdapat dalam ruangan.

Dari kedua tabel, Tabel 4.3 dan Tabel 4.4, model ResNet50 dan DenseNet201 tidak mampu mendeskripsikan gambar yang sesuai dengan konteks yang diberikan. Dapat dilihat dari tabel-tabel tersebut, semua deskripsi gambar yang diprediksi oleh Model#1, Model#2, Model#3, maupun Model#4 tidak memberikan hasil prediksi yang benar.

Tabel 4.3 Hasil Prediksi Deskripsi Gambar

No	Gambar	Deskripsi
1		model 1: di depan merupakan kamar tidur yang cukup luas model 2: di depan merupakan ruangan yang memiliki banyak perabotan di tengah dan sisi ruangan model 3: di sisi kanan terdapat tempat tidur berukuran besar dengan seprai berwarna putih model 4: di depan merupakan kamar tidur yang cukup luas model 5: di atas meja terdapat pizza dan botol minuman di atasnya model 6: di depan terdapat pizza dan meja bertaplak biru dengan banyak kursi model 7: di depan terdapat pizza dan gelas kaca di atas meja model 8: di atas meja kayu terdapat pizza dan gelas kaca model 9: di depan terdapat pizza dan gelas kaca di atas meja model 10: di depan terdapat pizza dan gelas kaca model 11: di depan terdapat seorang wanita yang sedang duduk di kursi dan menggunakan ponsel lipat

		model 12: di depan terdapat banyak orang yang sedang duduk di kursi dan menikmati pizza di atas meja model 13: di depan terdapat meja makan dengan banyak makanan dan minuman di atasnya
2		model 1: di depan terdapat meja bertaplak hijau yang panjang dengan banyak makanan dan minuman di atasnya model 2: di depan terdapat meja kursi restoran berbagai botol dan meja dengan beberapa kursi kanannya model 3: di depan terdapat meja kayu dengan banyak makanan di atasnya model 4: di depan terdapat pizza dan gelas kaca di atas meja model 5: di bagian kiri wastafel terdapat cermin berukuran besar model 6: di bagian kanan terdapat bak mandi dengan tirai terbuka model 7: di bagian kiri wastafel terdapat handuk berwarna putih yang menggantung model 8: di bagian kanan terdapat meja wastafel berwarna putih dengan cermin di atasnya model 9: terdapat wastafel berbentuk oval kecil berwarna putih di bagian kanan model 10: di bagian kiri terdapat wastafel dengan cermin besar di atasnya model 11: di depan terdapat wastafel dengan cermin di atasnya model 12: di bagian kiri terdapat meja wastafel dengan cermin besar di atasnya

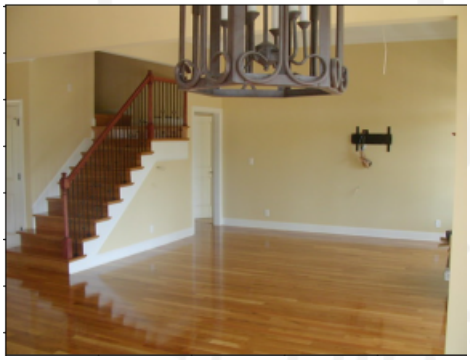
		model 13: di bagian kanan terdapat meja makan di dekat jendela kaca yang berukuran sedang
3		model 1: di depan merupakan kamar tidur yang cukup luas model 2: di ruangan terdapat banyak lukisan dan cermin yang digantung di dinding model 3: di sisi kiri ruangan terdapat tempat tidur berukuran besar dengan seprai berwarna putih model 4: di ujung depan terdapat pintu kecil dan jendela yang tertutup tirai berwarna putih yang tertutup model 5: di samping kiri terdapat rak buku rendah dengan banyak buku di dalamnya model 6: di depan merupakan ruangan dengan sofa berwarna yang berada di bagian kanan model 7: di depan merupakan ruang santai dengan dua pintu berwarna cokelat yang terbuka model 8: di bagian kiri terdapat sofa berwarna hitam model 9: di sisi kiri ruangan terdapat rak buku yang besar dengan banyak buku di dalamnya model 10: di samping kiri terdapat sofa panjang berwarna merah dengan meja kecil yang rendah dari kayu di depannya model 11: di ujung kiri ruangan terdapat rak buku tinggi dari kayu dengan beberapa ornamen di atasnya

		model 12: di depan ada seorang anak sedang duduk di kursi kayu berwarna cokelat model 13: di bagian kiri terdapat rak buku rendah dengan banyak buku di dalamnya

Meskipun model-model IncepResNetV2 mampu mengenerasi prediksi yang benar sesuai dari beberapa contoh yang diberikan sebelumnya, namun pada Tabel 4.4 dapat dilihat beberapa hasil prediksi lain di mana Model#5 IncepResNetV2 lebih banyak dalam mengenerasi deskripsi gambar yang baik dibandingkan model yang lain. Pada tabel ini, Model#5 mampu memberikan letak objek dengan benar, objek terdekatnya dan mengenali beberapa karakteristik seperti (Gambar#1: wastafel berwarna putih dan Gambar#3 pintu berwarna putih yang terbuka).

Tabel 4.4 Beberapa contoh hasil prediksi *caption* setiap model.

No	Gambar	Deskripsi
1		model 1: di depan merupakan lorong yang terletak di tengah ruangan model 2: di depan merupakan ruangan dengan banyak perabotan yang terletak di lantai model 3: di ujung depan terdapat tangga naik yang dapat digunakan dari sisi kiri model 4: di depan merupakan ruangan yang cukup luas dengan tangga naik berada di sisi kiri model 5: di sisi kiri ruangan terdapat wastafel berwarna putih dengan cermin yang menggantung di atasnya model 6: di depan merupakan kamar tidur yang cukup luas model 7: di kanan ruangan terdapat lemari penyimpanan yang terletak di atas mesin cuci

		model 8: di depan merupakan ruangan dengan mesin cuci berwarna putih di sisi kiri ruangan model 9: di depan merupakan kamar kecil yang cukup sempit model 10: di ujung kiri ruangan terdapat mesin cuci berwarna putih yang terletak di samping mesin cuci model 11: di ujung kiri ruangan terdapat pintu berwarna hijau yang tertutup model 12: di depan merupakan kamar kecil yang cukup luas model 13: di samping kiri terdapat mesin cuci berwarna putih dengan kipas yang menggantung di atasnya
2		model 1: di depan merupakan ruang dapur yang cukup sempit model 2: di depan terdapat tempat tidur yang dengan sprei berwarna putih yang menggantung model 3: di depan ada seorang wanita sedang membawa ponsel berwarna hitam model 4: di depan ada seorang pria mengenakan baju berwarna merah muda sedang berdiri model 5: di depan merupakan ruangan dengan tangga naik yang menempel di dinding model 6: di depan merupakan ruangan dengan tangga yang berada di sisi kiri model 7: di depan merupakan lorong yang dinding dan tangga naik

		model 8: di depan merupakan ruangan yang cukup luas dengan tangga di sisi kiri model 9: di samping kanan tangga terdapat pintu dari kayu yang terbuka menuju ruangan lain model 10: di ujung depan terdapat tangga naik yang berbentuk lurus model 11: di depan terdapat ruang santai dengan beberapa sofa di kanan dan kiri model 12: di bagian kiri terdapat pintu berwarna putih yang tertutup model 13: di samping kiri terdapat pintu berwarna putih yang tertutup
3		model 1: di sisi kanan ruangan terdapat meja konter kecil dari kayu dengan beberapa barang di atasnya model 2: di sisi kiri ruangan terdapat meja konter berwarna coklat konter berwarna coklat konter berwarna putih dengan mesin cuci di atasnya model 3: di ujung depan terdapat kompor oven dan oven di antara kompor model 4: di ujung depan terdapat jendela transparan yang terletak di samping kulkas kecil berwarna putih model 5: di sisi kanan depan terdapat pintu berwarna putih yang terbuka menuju ruangan lain model 6: di depan merupakan kamar tidur yang luas dengan sedikit perabotan di sisi kanan dan kiri ruangan

		model 7: di kanan ruangan terdapat lemari rendah dari kayu yang pintunya terbuka menuju ruangan lain model 8: di ujung depan terdapat pintu berwarna putih yang terbuka menuju ruangan lain model 9: di sisi kiri ruangan terdapat lemari penyimpanan dari kayu yang besar dengan cermin yang menggantung di atasnya model 10: di sisi kanan ruangan terdapat lemari penyimpanan rendah dari kayu dengan beberapa barang di atasnya model 11: di sisi kiri terdapat pijakan tangga naik ke lantai bawah model 12: di sisi kiri ruangan terdapat kulkas dua pintu berwarna putih yang terletak di bawah meja konter model 13: di depan merupakan kamar tidur yang luas dengan jendela dari kaca transparan yang tirainya digulung ke atas
4		model 1: di depan terdapat banyak orang yang sedang berdiri di ruangan model 2: di depan terdapat meja panjang dengan banyak makanan di atasnya model 3: di depan terdapat meja panjang dengan banyak makanan di atasnya model 4: di depan ada banyak makanan dan minuman di atas meja panjang model 5: di depan merupakan ruang santai dengan perapian di kanan ruangan model 6: di depan merupakan ruang dengan perapian yang menyala dengan lampu baca yang ada di atasnya

		model 7: di sisi kanan ruangan terdapat meja konter yang memanjang dengan wastafel yang terletak di tengahnya model 8: di ujung depan terdapat perapian yang tidak menyala dengan banyak perabotan di atasnya model 9: di ujung ruangan terdapat perapian yang tidak menyala dengan banyak pigura yang menggantung di atasnya model 10: di depan merupakan ruangan yang cukup luas dengan perapian di atasnya model 11: di ujung depan terdapat perapian dengan lampu gantung dari kayu di atasnya model 12: di ujung kiri ruangan terdapat perapian dengan pot bunga di atasnya model 13: di depan merupakan kamar tidur yang luas dengan jendela di ujung depan yang tirainya digulung ke atas

4.4 Prediksi *Caption* pada Gambar Berlatar di Indonesia

Karena data yang digunakan merupakan data dari MSCOCO dan Flickr. Dilakukan percobaan pada data berbeda dengan mengambil beberapa gambar dalam ruangan dari Google yang sesuai dengan perabotan yang biasa digunakan di Indonesia. Hasil prediksi dapat dilihat pada Gambar 4.1, Gambar 4.2, dan Gambar 4.3 dengan menggunakan data gambar dapur, tempat tidur, dan toilet. Hasil yang didapatkan merupakan hasil prediksi deskripsi gambar dari Model 5 yang memiliki skor evaluasi tertinggi dengan fitur ekstraksi IncepResNetV2.

Dari prediksi yang didapatkan, meskipun hasil *caption* belum sempurna dalam mendeskripsikan gambar, namun dapat dilihat bahwa hasil yang didapatkan cukup sesuai untuk mendeskripsikan konteks dari gambar yang diberikan. Pada Gambar 4.1 model mengerti bahwa gambar yang diberikan merupakan gambar dapur dengan mengenali "kompor oven" dari adanya kompor gas yang ada pada gambar. Model belum dapat mengenerasi deskripsi dengan baik karena kemunculan deskripsi "sungkup udara" pada

gambar ini. Hal ini terjadi karena kebanyakan data gambar dapur yang digunakan pada MSCOCO menggunakan kompor oven/*range stove* (kompor yang dikombinasi dengan oven) dan menggunakan sungkup udara di atasnya.

```
Predicted Caption: di depan terdapat kompor oven dengan sungkup udara berwarna putih di atasnya
<matplotlib.image.AxesImage at 0x7fb1940d6dd8>
```

Gambar 4.1 Hasil prediksi *caption* pada gambar dapur yang biasa digunakan di Indonesia.

Pada Gambar 4.2 model mampu mengenali bahwa gambar yang diberikan diambil di kamar tidur. Model mampu mengenerasi "tempat tidur" dan mengenali "nakas kecil berwarna putih" yang ada di dalam gambar. Pada gambar ini, model mampu mengenerasi deskripsi gambar dengan baik. Sedangkan pada Gambar 4.3 model mengenali bahwa gambar merupakan gambar dari toilet karena model mampu mengenerasi kata "toilet duduk berwarna putih". Sama halnya dengan yang terjadi pada Gambar 4.1, gambar yang dimiliki oleh MSCOCO hanya menyediakan gambar dengan toilet duduk, sehingga model belum bisa mengenali perbedaan bentuk toilet.

```
Predicted Caption: di samping kanan tempat tidur terdapat nakas kecil berwarna putih
<matplotlib.image.AxesImage at 0x7fb19414fb00>
```

Gambar 4.2 Hasil prediksi *caption* pada gambar tempat tidur yang bisasa digunakan di Indonesia.

```
Predicted Caption: di depan terdapat toilet duduk berwarna putih
<matplotlib.image.AxesImage at 0x7fb197fe9278>
```

Gambar 4.3 Hasil prediksi *caption* pada gambar toilet yang biasa digunakan di Indonesia.

4.5 Evaluasi *Image Captioning*

Dari eksperimen dengan kombinasi model fitur ekstraksi dengan transformer serta pengubahan *hyperparameter* yang telah kita lakukan, dapat dilihat bahwa perbedaan skor evaluasi yang didapatkan dipengaruhi oleh nilai *attention heads* dan *batch size* untuk model IncepResNetV2, InceptionV3, Xception, dan DenseNet. Pada fitur ekstraksi IncepResNetV2 dilakukan eksperimen dengan menggunakan nilai 4, 8, dan 16 untuk pengaturan nilai *attention heads*. Hasil yang didapatkan menunjukkan bahwa pengaturan dengan nilai terendah *attention heads* 4 membantu model untuk mendapatkan skor evaluasi yang tertinggi. Hal ini sesuai dengan penelitian sebelumnya yang membuktikan bahwa nilai *attention heads* dengan nilai kecil mampu memberikan hasil evaluasi skor yang tinggi (Al-Faruq, 2021). Sebaliknya di antara nilai 32, 64 dan 128 untuk ukuran *batch,* nilai *batch* tertinggi yakni 128 mampu memberikan skor evaluasi tertinggi pada model. Hasil dari kedua kombinasi penggunaan *attention heads* dengan nilai 4 dan *batch size* 128 pada Model 5 pada Tabel 4.2 berhasil memberikan nilai skor tertinggi untuk seluruh matriks evaluasi yakni BLEU-1 0.6971, BLEU-2 0.5246, BLEU-3, 0.3921, BLEU-4 0.2831, METEOR 0.2468, CIDEr 0.4801, dan ROUGE-L 0.5114. Model kombinasi dengan 4 *attention head* dan ukuran *batch* 128 memiliki performa yang baik karena model memiliki cukup parameter untuk mempelajari data yang digunakan dan penggunaan *batch* yang besar dapat menjangkau optimum minima (fungsi *loss* berada di titik minimum) dibandingkan *batch* yang kecil (Kandel & Castelli, 2020).

Kami juga telah melakukan analisis hasil prediksi yang didapatkan untuk setiap model fitur ekstraksi. Dari analisis yang dilakukan, model IncepResNetV2 dapat mengenali objek dan prediksi *caption* yang dihasilkan sesuai dengan konteks gambar yang diberikan. Prediksi deskripsi gambar yang didapatkan mampu dengan baik mengenerasi nama objek (Tabel 4.3 wastafel, pizza, rak buku), warna (meja biru, handuk putih), posisi/lokasi (Tabel 4.3 di depan, di bagian kiri, di atas meja), karakteristik (Tabel 4.4 tangga naik, pintu terbuka), dan objek sekitarnya (Gambar 4.2 mampu mengenali nakas di samping tempat tidur). Namun model ini juga memiliki kekurangan dalam mendapatkan karakteristik maupun letak objek dengan baik seperti pada Tabel 4.3 Gambar#3. Namun begitu, deskripsi yang dihasilkan oleh IncepResNetV2 masih cukup sesuai dan tidak keluar dari konteks gambar yang diberikan. Dari seluruh model eksperimen yang menggunakan IncepResNetV2, Model 5 dengan *attention heads* 4, *dropout* 0.1, dan ukuran *batch* 128 menghasilkan lebih banyak hasil prediksi yang sesuai dibandingkan model IncepResNetV2 yang lain (Lihat Tabel 4.4).

Pada fitur ekstraksi InceptionV3 dan Xception, model cukup baik dalam mengenali ruangan yang diberikan pada gambar. Namun model cenderung untuk mengenerasi objek yang tidak ada di dalam ruangan seperti pada Tabel 4.3 Gambar#1. Model InceptionV3 maupun Xception mengenerasi objek manusia (“seorang wanita” & “banyak orang”) yang tidak ada di dalam gambar yang diberikan. Sedangkan pada model DenseNet201 dan ResNet50 model gagal dalam mengenerasi *caption* yang benar dan sesuai dengan gambar yang diberikan. Banyaknya prediksi yang salah yang didapatkan oleh kedua model sesuai dengan evaluasi skor yang didapatkan karena DenseNet201 dan ResNet50 mendapatkan nilai yang rendah di seluruh matriks evaluasi.

Penelitian ini juga membandingkan model yang telah dibangun dengan beberapa penelitian-penelitian mengenai *image captioning* dengan Bahasa Indonesia yang sebelumnya pernah dilakukan. Telah ditampilkan penelitian-penelitian sebelumnya yang menggunakan dataset MSCOCO maupun Flickr yang diterjemahkan ke dalam Bahasa Indonesia pada Tabel 4.5. Pada tabel tersebut dapat dilihat bahwa model yang dibangun pada penelitian ini dengan menggunakan IncepResNetV2 dan Transformer dengan pengaturan *hyperparameter attention heads* 4, *dropout* 0.1, dan ukuran *batch* 128 meskipun menggunakan data yang sedikit namun mampu menghasilkan skor BLEU-1, BLEU-2, BLEU-3, dan BLEU-4 yang dapat dibandingkan dengan penelitian-penelitian sebelumnya.

Tabel 4.5 Perbandingan model dengan penelitian *image captioning* sebelumnya.

No	Dataset	Model	Total Images	Captions per image	BLEU Score (n-gram)			
					1	2	3	4
1	Flickr - FEEH – ID (Mulyanto et al., 2019)	CNN-LSTM	8099	5	50.0	31.4	23.9	13.1
2	Flickr30k-ID (Nugraha et al., 2019)	CNN-GRU	31783	5	36.7	17.8	6.7	2.0
3	MSCOCO & Flickr30k (Mahadi et al., 2020)	ResNet101-LSTM with adaptive attention	180k	5	67.8	51.2	37.5	27.4
4	**Our modified MSCOCO-Flickr**	IncepResNetV2-Transformer	1072	5	69.7	52.5	39.2	28.3

BAB 5
Kesimpulan dan Saran

5.1 Kesimpulan

Pada penelitian ini, dibangun model untuk mengenerasi teks dari gambar yang diambil di dalam ruangan untuk mendapatkan pemahaman visual di dalam ruangan. Penelitian ini berkontribusi untuk mempresentasikan evaluasi arsitektur Transformer dan mengidentifikasi objek untuk mendapatkan pemahaman visual di dalam ruangan pada dataset gambar MSCOCO dan Flickr dengan *caption* berbahasa Indonesia dengan deskripsi yang menggambarkan nama objek, posisi/lokasi (berdasarkan sudut pandang pengguna), karakteristik, dan objek di sekitarnya.

Dalam pengembangannya, digunakan model Transformer yang diubah dan disesuaikan modelnya dengan *hyper-parameter tuning* untuk mendapatkan model terbaik. Selain itu, dilakukan pula beberapa eksperimen dalam menggunakan varian *pre-trained* CNN untuk mendapatkan fitur gambar yang kemudian akan dilanjutkan ke dalam model Transformer. Dari penelitian yang dilakukan, didapatkan beberapa kesimpulan yang mengacu pada tujuan penelitian, yakni:

1. Telah dilakukan penelitian *image captioning* dengan menggunakan Transformer dan bereksperimen dengan beberapa varian model *pre-trained* untuk mengekstraksi fitur gambar dengan menggunakan data MSCOCO dan Flickr dengan keterangan gambar berbahasa Indonesia.
2. Didapatkan hasil terbaik pada seluruh matriks evaluasi (BLEU-1 0.6971, BLEU-2 0.5246, BLEU-3, 0.3921, BLEU-4 0.2831, METEOR 0.2468, CIDEr 0.4801, dan ROUGE-L 0.5114) dari penggunaan pengekstraksi fitur IncepResNetV2 dengan pengaturan *hyperparameter* Transformer ukuran *batch* 128, *dropout* 0.1, dan *attention heads* 4 untuk melakukan *image captioning*.
3. Dari pengaturan *hyperparameter* yang dilakukan pada model ekstraksi fitur IncepResNetV2, didapatkan nilai skor matriks evaluasi yang semakin baik dengan ukuran *batch* yang paling tinggi yakni 128 dan nilai *attention heads* yang paling kecil yakni 4.
4. Mengetahui hasil implementasi image captioning dari penggunaan IncepResNetV2 untuk mendapatkan pemahaman visual di dalam ruangan baik dalam data uji MSCOCO-Flickr serta pada data gambar Google yang berlatar di Indonesia.

5. Meskipun model yang dibangun memiliki dataset yang kecil dengan 1072 gambar, model dapat dibandingkan dengan penelitian-penelitian *image captioning* dengan Bahasa Indonesia sebelumnya.

5.2 Saran

Penelitian ini merupakan penelitian Transformer *image captioning* dengan Bahasa Indonesia untuk pemahaman visual di dalam ruangan dengan menggunakan dataset yang masih tergolong sedikit. Penelitian selanjutnya dapat dikembangkan kembali menggunakan dataset yang lebih besar, baik dari jumlah gambar maupun dari objek yang digunakan, sehingga tidak terbatas untuk penggunaan di dalam ruangan saja. Selain itu, model ini masih terbatas untuk menghasilkan satu deskripsi saja. Sehingga selanjutnya dapat dikembangkan *image captioning* untuk menghasilkan beberapa deskripsi untuk mendapatkan pemahaman visual yang lebih baik. Model juga dapat dikembangkan menggunakan model *state-of-the-art* untuk mendapatkan hasil terbaik.

Daftar Pustaka

Al-Faruq, U. A. A. (2021). *Implementasi Arsitektur Transformer pada Image Captioning dengan Bahasa Indonesia*.

Al-Malla, M. A., Al-Malla, M. A., Jafar, A., & Ghneim, N. (2022). Pre-trained CNNs as Feature-Extraction Modules for Image Captioning. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, *21*(1), 1–16. https://doi.org/10.5565/rev/elcvia.1436

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

Bhatia, Y., Bajpayee, A., Raghuvanshi, D., & Mittal, H. (2019). Image Captioning using Google's Inception-resnet-v2 and Recurrent Neural Network. *2019 Twelfth International Conference on Contemporary Computing (IC3)*, 1–6. https://doi.org/10.1109/IC3.2019.8844921

Chen, F., Li, X., Tang, J., Li, S., & Wang, T. (2021). A Survey on Recent Advances in Image Captioning. *Journal of Physics: Conference Series*, *1914*(1). https://doi.org/10.1088/1742-6596/1914/1/012053

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., & Zitnick, C. L. (2015). *Microsoft COCO Captions: Data Collection and Evaluation Server*. 1–7. http://arxiv.org/abs/1504.00325

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020a). Meshed-memory transformer for image captioning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 10575–10584. https://doi.org/10.1109/CVPR42600.2020.01059

Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020b). Meshed-memory transformer for image captioning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 10575–10584. https://doi.org/10.1109/CVPR42600.2020.01059

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*(Mlm), 4171–4186.

Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *Why 70 / 30 or 80 / 20 Relation Between Training and Testing Sets : A Pedagogical Explanation*. 1–6.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Huang, G., Liu, Z., van der Maaten, L., & Weinberger, K. Q. (2016). *Densely Connected Convolutional Networks*.

Jaiswal, A., Gianchandani, N., Singh, D., Kumar, V., & Kaur, M. (2021). Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning. *Journal of Biomolecular Structure and Dynamics*, *39*(15), 5682–5689. https://doi.org/10.1080/07391102.2020.1788642

Jeamwatthanachai, W., Wald, M., & Wills, G. (2019). Indoor navigation by blind people: Behaviors and challenges in unfamiliar spaces and buildings. *British Journal of Visual Impairment*, *37*(2), 140–153. https://doi.org/10.1177/0264619619833723

Jia, X., Gavves, E., Fernando, B., & Tuytelaars, T. (2015). Guiding the long-short term memory model for image caption generation. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 2407–2415. https://doi.org/10.1109/ICCV.2015.277

Kandel, I., & Castelli, M. (2020). The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*, *6*(4), 312–315. https://doi.org/10.1016/j.icte.2020.04.010

Katiyar, S., & Borgohain, S. K. (2021). *Image Captioning using Deep Stacked LSTMs, Contextual Word Embeddings and Data Augmentation*.

Keskar, N. S., Nocedal, J., Tang, P. T. P., Mudigere, D., & Smelyanskiy, M. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1–16.

Lavie, A., & Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. *Proceedings of the Second Workshop on Statistical Machine Translation*, *0*(June), 228–231.

Li, G., Zhu, L., Liu, P., & Yang, Y. (2019). Entangled transformer for image captioning. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-Octob*(c), 8927–8936. https://doi.org/10.1109/ICCV.2019.00902

Li, J., Yao, P., Guo, L., & Zhang, W. (2019). Boosted transformer for image captioning. *Applied Sciences (Switzerland)*, *9*(16), 1–15. https://doi.org/10.3390/app9163260

Lin, C. Y. (2005). *ROUGE: A Package for Automatic Evaluation of Summaries*.

Lu, H., Yang, R., Deng, Z., Zhang, Y., Gao, G., & Lan, R. (2021). Chinese Image Captioning via Fuzzy Attention-based DenseNet-BiLSTM. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *17*(1s), 1–18. https://doi.org/10.1145/3422668

Maeda-Gutiérrez, V., Galván-Tejada, C. E., Zanella-Calzada, L. A., Celaya-Padilla, J. M., Galván-Tejada, J. I., Gamboa-Rosales, H., Luna-García, H., Magallanes-Quintanar, R., Guerrero Méndez, C. A., & Olvera-Olvera, C. A. (2020). Comparison of convolutional neural network architectures for classification of tomato plant diseases. *Applied Sciences (Switzerland)*, *10*(4). https://doi.org/10.3390/app10041245

Mahadi, M. R. S., Arifianto, A., & Ramadhani, K. N. (2020). Adaptive Attention Generation for Indonesian Image Captioning. *2020 8th International Conference on Information and Communication Technology, ICoICT 2020*. https://doi.org/10.1109/ICoICT49345.2020.9166244

Mishra, S. K., Rai, G., Saha, S., & Bhattacharyya, P. (2022). Efficient Channel Attention Based Encoder–Decoder Approach for Image Captioning in Hindi. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(3), 1–17. https://doi.org/10.1145/3483597

Mulyanto, E., Setiawan, E. I., Yuniarno, E. M., & Purnomo, M. H. (2019). Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset. *2019 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, CIVEMSA 2019 - Proceedings*. https://doi.org/10.1109/CIVEMSA45640.2019.9071632

Nugraha, A. A., Arifianto, A., & Suyanto. (2019). Generating image description on Indonesian language using convolutional neural network and gated recurrent unit. *2019*

7th International Conference on Information and Communication Technology, ICoICT 2019, 1–6. https://doi.org/10.1109/ICoICT.2019.8835370

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, *371*(23), 311. https://doi.org/10.3115/1073083.1073135

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. *International Journal of Computer Vision*, *123*(1), 74–93. https://doi.org/10.1007/s11263-016-0965-7

Rashid, H., Al-Mamun, A. S. M. R., Robin, M. S. R., Ahasan, M., & Reza, S. M. T. (2017). Bilingual wearable assistive technology for visually impaired persons. *1st International Conference on Medical Engineering, Health Informatics and Technology, MediTec 2016*. https://doi.org/10.1109/MEDITEC.2016.7835386

Sharma, H., Agrahari, M., Singh, S. K., Firoj, M., & Mishra, R. K. (2020). Image Captioning: A Comprehensive Survey. *2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and Its Control, PARC 2020*, 325–328. https://doi.org/10.1109/PARC49193.2020.236619

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, *1*, 2556–2565. https://doi.org/10.18653/v1/p18-1238

Shuster, K., Humeau, S., Hu, H., Bordes, A., & Weston, J. (2019). Engaging image captioning via personality. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 12508–12518. https://doi.org/10.1109/CVPR.2019.01280

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*. https://doi.org/10.1016/0370-2693(93)90272-J

Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2021). *From Show to Tell: A Survey on Image Captioning*. 1–22. http://arxiv.org/abs/2107.06912

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2016). *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.

Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 4566–4575. https://doi.org/10.1109/CVPR.2015.7299087

Vellakani, S., & Pushbam, I. (2020). An enhanced OCT image captioning system to assist ophthalmologists in detecting and classifying eye diseases. *Journal of X-Ray Science and Technology*, *28*(5), 975–988. https://doi.org/10.3233/XST-200697

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2014). *Show and Tell: A Neural Image Caption Generator*.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., & Titov, I. (2019). *Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned*.

Wang, C., Zhou, Z., & Xu, L. (2021). An Integrative Review of Image Captioning Research. *Journal of Physics: Conference Series*, *1748*(4). https://doi.org/10.1088/1742-6596/1748/4/042060

Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., Wang, Y., & Wang, Y. (2017). *AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding*. https://doi.org/10.1109/ICME.2019.00256

Zakir Hossain, M. D., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, *51*(6). https://doi.org/10.1145/3295748