Indonesian Journal of Computer Science
ISSN 2302-4364 (*print*) dan 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam, No. 1, Padang, Indonesia, Telp. (0751) 7056199, 7058325
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Image Captioning untuk Gambar Rambu Lalu Lintas Indonesia Menggunakan *Pretrained* CNN dan *Transformer*

Novia Pramesti Aprilia[1], Theresia Herlina Rochadiani[2]
Novia.pramesti@student.pradita.ac.id, theresia.herlina@pradita.ac.id
Universitas Pradita

Informasi Artikel

Diterima : 12 Mei 2024
Direview : 19 Mei 2024
Disetujui : 15 Jun 2024

Kata Kunci

Image Captioning, Inception V3, Indonesia, Rambu Lalu Lintas, *Transformer*

Abstrak

Penelitian ini bertujuan untuk mengatasi kurangnya pemahaman terhadap rambu lalu lintas di Indonesia melalui pengembangan model *image captioning* menggunakan Inception V3 dan *Transformer*. Dengan menggunakan pendekatan ini, dataset gambar rambu lalu lintas yang terdiri dari 9.594 gambar dengan 31 kelas dikumpulkan dan dimodifikasi. Evaluasi model dilakukan menggunakan metrik BLEU, ROUGE-L, METEOR, dan CIDEr. Hasil penelitian menunjukkan kinerja yang baik dengan skor BLEU-1=0.89, BLEU-2 = 0.82, BLEU-3 = 0.75, BLEU-4 = 0.68, CIDEr = 0.57, ROUGE-L = 0.25, dan METEOR = 0.26. Dari hasil tersebut, dapat mengindikasikan bahwa model ini dapat meningkatkan pemahaman tentang rambu lalu lintas Indonesia. Pendekatan ini dapat membantu pengguna jalan memahami rambu lalu lintas dengan lebih baik, serta memiliki potensi untuk diterapkan dalam aplikasi praktis untuk meningkatkan keselamatan lalu lintas.

Keywords

Image Captioning, Inception V3, Indonesia, Traffic Sign, Transformer

Abstract

This research aims to address the lack of understanding of traffic signs in Indonesia through the development of an image captioning model using Inception V3 and Transformer. With this approach, a dataset of traffic sign images consisting of 9,594 images with 31 classes was collected and modified. Model evaluation was conducted using BLEU, ROUGE-L, METEOR, and CIDEr metrics. The research results show good performance with BLEU-1 score of 0.89, BLEU-2 = 0.82, BLEU-3 = 0.75, BLEU-4 = 0.68, CIDEr = 0.57, ROUGE-L = 0.25, and METEOR = 0.26. From these results, it can be indicated that this model can enhance understanding of Indonesian traffic signs. This approach can assist road users in better understanding traffic signs and has the potential to be applied in practical applications to improve traffic safety.

A. Pendahuluan

Lalu lintas merupakan bagian integral dari kehidupan perkotaan yang *modern*, dan keselamatan pengguna jalan sangat bergantung pada pemahaman yang tepat terhadap rambu lalu lintas. Kemampuan untuk memahami rambu-rambu lalu lintas dianggap sebagai kontributor utama dalam keselamatan jalan raya. Jika tidak dilakukan, hal ini dapat menimbulkan masalah yang lebih serius [1]. Oleh karena itu, diperlukan peningkatan keselamatan pengguna jalan melalui komunikasi pesan rambu lalu lintas untuk meningkatkan pemahaman terhadap rambu lalu lintas.

Undang-undang Nomor 22 Tahun 2009 tentang Lalu Lintas dan Angkutan Jalan menyatakan bahwa Rambu Lalu Lintas adalah lambang, huruf, angka, kalimat, atau kombinasi di antaranya yang berfungsi sebagai peringatan, larangan, perintah, atau petunjuk bagi pengguna jalan [2]. Kurangnya pemahaman terhadap lalu lintas telah mengakibatkan beberapa risiko yang menjadi perhatian utama di banyak negara. Di Manila, kurangnya pemahaman terhadap rambu lalu lintas telah terbukti menjadi penyebab utama kecelakaan dan pelanggaran lalu lintas berdasarkan laporan tahun 2017 dari *Metropolitan* Manila *Development Authority* [3].

Kesalahan dalam pemahaman untuk mencari makna dari sebuah gambar dapat menciptakan kerancuan untuk mengartikan apa arti dari gambar tersebut. Terkadang, dengan hanya melihat sebuah objek gambar secara sekilas, manusia mampu untuk mengartikan atau mendeskripsikan objek tersebut, karena hal tersebut adalah pekerjaan yang mudah. Dengan adanya *image captioning*, bertujuan untuk memberikan *caption* atau deskripsi singkat pada gambar untuk menghasilkan kalimat yang benar secara linguistik dan semantik sesuai dengan isi gambar tersebut [4] .

Pada dasarnya *image captioning* adalah membuat kalimat deskriptif berdasarkan gambar secara otomatis [5]. Teknologi tersebut didasari oleh *computer vision* yang menggabungkan identifikasi objek dalam gambar melalui penggunaan *Convolutional Neural Network* (CNN) sebagai *encoder* dan *Natural Language Processing* (NLP) yang digunakan untuk menghasilkan deskripsi atau keterangan dari gambar tersebut dalam bentuk bahasa alami sebagai *decoder* [6]. *Encoder* dalam model *image captioning* berfungsi untuk mengekstraksi fitur dari gambar, menghasilkan kumpulan data yang merepresentasikan objek-objek yang terdapat dalam gambar tersebut. Selanjutnya, *decoder* digunakan untuk merekonstruksi kalimat berdasarkan data yang telah diidentifikasi sebelumnya, sehingga dapat menghasilkan kalimat deskripsi yang sesuai dengan gambar masukan [7].

Penelitian oleh P. Chun, T. Yamane, dan Y. Maemura pada tahun 2021, berjudul "*A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage*", menggunakan model Inception V3 dan arsitektur *attention mechanism*. Studi ini menunjukkan skor yang lebih tinggi dibandingkan dengan tidak menggunakan *attention mechanism*, dengan skor BLEU-1 sebesar 0.782, BLEU-2 sebesar 0.749, BLEU-3 sebesar 0.711, dan BLEU-4 sebesar 0.693 [8].

Penelitian yang menggunakan *Transformer*, dilakukan oleh F.M. Shah, M. Humaira, Md. A.R.K. Jim, A.S. Ami, dan S. Paul pada tahun 2021 dengan judul "Bornon: Bengali *Image captioning With Transformer-Based Deep Learning Approach*" menunjukkan bahwa metode berbasis *Transformer* mengungguli metode *attention mechanism*. Selain memberikan hasil yang lebih baik, model ini juga

mampu meningkatkan kinerja dan kecepatan pelatihan dengan memungkinkan mekanisme paralel. Di dalam *Transformer* juga dilengkapi dengan *multi-head attention mechanism* [5] *.*

Penelitian sebelumnya yang dilakukan oleh O. Abbas dan J. Dang pada tahun 2023, berjudul "*Using image captioning for automatic post-disaster damage detection and identification*", menggunakan dataset Image-Hub yang terdiri dari 2000 data. Data tersebut dibagi secara acak menjadi set pelatihan, *validation*, dan uji dengan proporsi 60-20-20. Dalam penelitian ini, model CNN seperti VGG16, ResNet50, Inception V3, dan EfficientNetB0 digunakan sebagai *encoder*, sedangkan LSTM digunakan sebagai *decoder*. Hasil penelitian menunjukkan bahwa Inception V3 lebih unggul dibandingkan dengan model lainnya, dengan skor BLEU-1 sebesar 0.8731, BLEU-2 sebesar 0.8088, BLEU-3 sebesar 0.769, BLEU-4 sebesar 0.7372, ROUGE-1 sebesar 0.8788, ROUGE-L sebesar 0.875, dan SPICE sebesar 0.7338 [9].

Pada tahun yang sama, yaitu 2023, penelitian yang dilakukan oleh B. Das, R. Pal, M. Ajumder, S. Phadikar, dan A. A. Sekh, yang berjudul "*A Visual Attention Based Model for Bengali Image captioning*", menggunakan dua dataset, yaitu BLIC dan BNLIT yang keduanya berbahasa Bengali. Dataset tersebut dibagi menjadi 80% untuk pelatihan, 10% untuk *validation*, dan 10% untuk pengujian. Dalam penelitiannya, Inception V3 dengan *attention mechanism* mengungguli model ResNet50 dan VGG16, dengan skor BLEU1 sebesar 0.67 dan BLEU4 sebesar 0.25 pada dataset BLIC, serta BLEU-1 sebesar 0.65 dan BLEU-4 sebesar 0.24 pada dataset BNLIT [10].

Walaupun sudah banyak penelitian mengenai *image captioning*, belum ada penelitian yang khusus melakukan *image captioning* terhadap rambu lalu lintas Indonesia. Oleh karena itu, penelitian ini bertujuan untuk membangun model *image captioning* rambu lalu lintas Indonesia menggunakan *pretrained* CNN yaitu Inception V3 dan arsitektur *Transformer*. Diharapkan penelitian ini dapat memberikan kontribusi bagi pengembangan dan penelitian lebih lanjut serta meningkatkan pemahaman tentang pengembangan model *image captioning* menggunakan Inception V3 dan *Transformer* untuk rambu lalu lintas Indonesia.

B. Metode Penelitian

1. Data Collection

Penelitian ini bertujuan untuk menerapkan *image captioning* pada gambar rambu lalu lintas yang ada di Indonesia, sehingga membutuhkan modifikasi dataset yang memanfaatkan dataset gabungan dari GTSRB (German *Traffic Sign Recognition Benchmark*) [11] dan *Traffic Sign in* Indonesia yang diperoleh dari *platform* Roboflow dibuat oleh UMY [12].

Sebelum diproses, dilakukan analisis awal pada dataset GTSRB untuk diambil rambu-rambu yang hanya berlaku di Indonesia. Selanjutnya, menggabungkan kedua dataset tersebut dengan mengelompokan rambu tersebut secara manual, dengan memisahkannya antar *folder* yang berbeda ke masing-masing kelasnya (rambu lalu lintas). Terdapat 31 kelas atau 31 jenis rambu lalu lintas yang berlaku di Indonesia. Dalam satu kelas, terdapat antara 250 hingga 310 gambar.

Tujuan utama dari penelitian ini adalah menghasilkan *caption* atau deskripsi gambar secara otomatis. Setiap kelas akan diberikan deskripsi umum

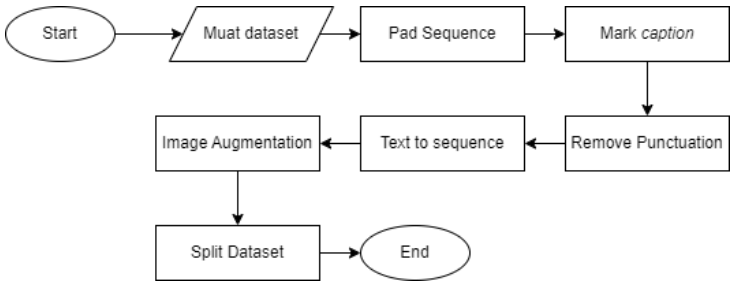
yang sesuai dengan "Buku Petunjuk Tata Cara Berlalu Lintas (*Highway Code*) di Indonesia" [13] yang diterbitkan oleh Direktorat Jenderal Perhubungan Darat pada tahun 2005.

Penelitian ini memerlukan *caption* atau deskripsi singkat untuk setiap gambar, karena deskripsi umum hanya mewakili satu arti untuk satu kelas, maka diperlukan pengembangan dari deskripsi umum tersebut untuk memperkaya kalimat dalam dataset. Untuk mencapai hal tersebut, peneliti menggunakan bantuan ChatGPT (*Generative Pretrained Transformer*) dari Open AI untuk menghasilkan pengembangan deskripsi umum dari masing-masing kelas menjadi 200 hingga 300 *caption*. Setelah berhasil mengembangkan *caption* dari deskripsi umum, setiap gambar dalam satu kelas akan dipasangkan dengan lima *caption* yang dipilih secara acak berdasarkan pengembangan *caption* dari deskripsi umum tersebut. Selanjutnya, *caption* tersebut akan digabungkan dalam satu file atau dokumen.

Dalam penelitian ini, terdapat dua berkas. Pertama adalah *folder* yang berisikan gambar-gambar dari gabungan seluruh kelas, totalnya mencapai 9.594 gambar. Kedua adalah *file* teks yang berisikan nama-nama gambar beserta lima *caption* untuk setiap gambar dari gabungan seluruh kelas, jumlah keseluruhannya mencapai 47.970 *caption.*

2. Data Preprocessing

Tahapan *preprocessing* merupakan tahapan awal yang dilakukan mempersiapkan data input atau masukan sebelum diproses pada *model captioning* [14].



Gambar 1. Alur data *preprocessing*

Data gambar dan *caption* akan dimuat terlebih dahulu sebelum melakukan proses ini. Setiap gambar memiliki lima *caption* yang berbeda, dan setiap *caption* memiliki format: <nama_gambar>#<nomor_*caption*>\t<*caption*>. Akan dilakukan *pad sequence* atau penghapusan *caption* yang terlalu pendek (kurang dari 5 *token*) atau terlalu panjang (lebih dari panjang maksimum yang ditentukan yaitu 40 *token*). Selanjutnya, akan ditambahkan *token* <start> dan <end> pada awal dan akhir setiap *caption* proses ini dapat disebut sebagai *mark caption*. Hasilnya adalah sebuah *dictionary* yang memetakan nama gambar ke daftar *caption* yang sesuai, serta daftar semua *caption* yang ada.

Dalam tahapan ini, standarisasi teks dilakukan dengan mengubah semua karakter teks menjadi huruf kecil dan menghapus karakter-karakter tertentu

seperti tanda baca dan karakter khusus (*remove punctuation*). Layer '`TextVectorization`' digunakan untuk melakukan *text to sequence* atau mengonversi data teks menjadi urutan bilangan bulat, di mana setiap bilangan bulat mewakili indeks kata dalam kosakata. Skema standarisasi kustom diterapkan di mana tanda baca kecuali < dan > dihapus.

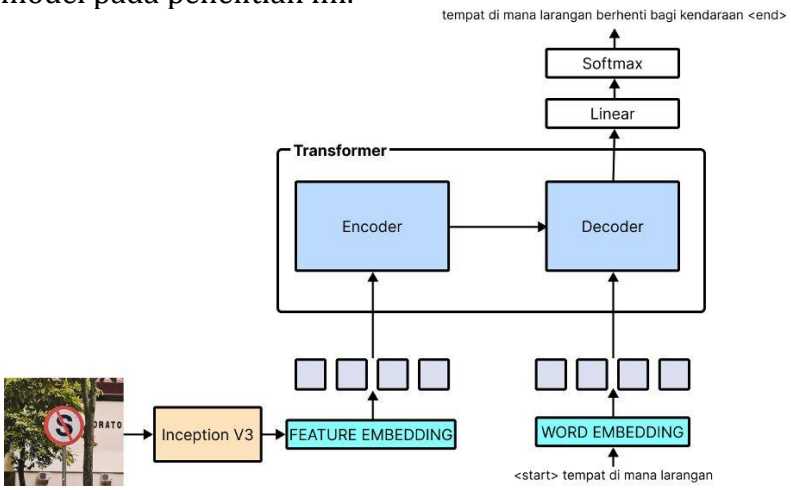
Pada dataset gambar, dilakukan proses augmentasi. Augmentasi melibatkan tiga jenis transformasi. Pertama, `RandomFlip("horizontal")` untuk secara acak melakukan *horizontal flipping* pada gambar-gambar. *Horizontal Flipping* adalah transformasi yang membalik gambar secara *horizontal*. Kedua, `RandomRotation(0.2)` secara acak melakukan rotasi pada gambar-gambar sebesar 0.2 radian. Rotasi gambar adalah transformasi yang memutar gambar searah jarum jam atau berlawanan arah jarum jam. Ketiga, `RandomContrast(0.3)` untuk secara acak meningkatkan kontras gambar sebesar 0.3. Peningkatan kontras merupakan transformasi yang meningkatkan perbedaan antara nilai piksel yang berdekatan dalam gambar. Selain itu, ukuran gambar yang ada akan diubah menjadi 299 x 299 untuk mengikuti ketentuan dari Inception V3.

Terakhir adalah *split* dataset atau memisahkan dataset *caption* menjadi data latih dan data *validation*. Proses ini melibatkan mengambil daftar semua nama gambar, mengacak urutan data tersebut secara *random*, dan yang terakhir adalah membaginya berdasarkan proporsi yang ditentukan yaitu 80% atau sama dengan 7480 gambar untuk data latih dan 20% atau sebesar 1870 gambar untuk data *validation*. Hasilnya adalah dua *dictionary* terpisah untuk data *training* dan *validation*.

3. Image captioning Model

Dalam pengembangan model untuk *image captioning*, digunakan arsitektur *Transformer*. Pertama, fitur gambar diekstraksi menggunakan model CNN yang telah dilatih sebelumnya, yaitu Inception V3. Setelah ekstraksi fitur, hasilnya diteruskan ke lapisan *dense* dengan fungsi aktivasi ReLU. Kemudian, hasil tersebut dimasukkan ke dalam lapisan *encoder* pada arsitektur *Transformer*. *Encoder* bertugas untuk mengolah fitur-fitur gambar tersebut. Hasil keluaran dari *encoder* kemudian diteruskan ke dalam lapisan *decoder*. Lapisan *decoder* menghasilkan kalimat-kalimat keterangan atau *caption* yang

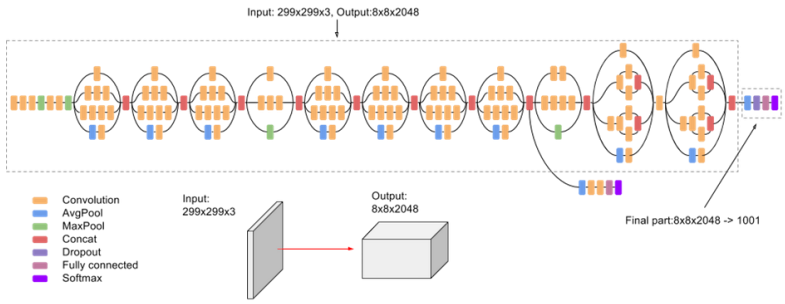
sesuai dengan gambar yang diberikan. Gambar 2 menunjukkan ilustrasi dari arsitektur model pada penelitian ini.



Gambar 2. Ilustrasi Arsitektur Model

3.1 Inception V3

Inception V3 adalah arsitektur *Convolutional Neural Network* (CNN) yang merupakan pengembangan dari versi sebelumnya, yaitu GoogleNet (Inception) dan Inception V2. Arsitektur ini dapat digunakan untuk melakukan klasifikasi gambar[15]. Pada dasarnya Inception V3 adalah sebuah model berbasis CNN yang telah dilatih sebelumnya dengan dataset 'ImageNet' dan merupakan salah satu *state-of-the-art* dari *pretrained* model [16]. Berikut arsitektur Inception V3 yang ditunjukan pada Gambar 1.



Gambar 3. Arsitektur Inception V3 [17]

Arsitektur ini memodifikasi konvolusi dengan mengganti *filter* yang ada menjadi *filter* 1-D. Fitur utama dari arsitektur ini adalah mengkombinasikan beberapa *filter* konvolusi yang berukuran berbeda menjadi satu *filter* tunggal. Pendekatan ini mengurangi kompleksitas komputasi dengan mengurangi jumlah parameter yang perlu dilatih [18].

Pada peneltian ini, model Inception V3 dimanfaatkan untuk mengekstraksi fitur dari gambar. Untuk tujuan tersebut, lapisan *softmax* dihilangkan karena hanya diperlukan ekstraksi vektor gambar. Sebelum dimasukkan ke dalam model, semua gambar disesuaikan ukurannya menjadi 299x299 piksel [19].

3.2 *Transformer*

Transformer adalah sebuah arsitektur inovatif yang mengadopsi *attention mechanism* untuk meningkatkan kecepatan, dimana arsitektur ini salah satu arsitektur Seq2Seq yang memiliki dua bagian, yaitu *encoder* dan *decoder* [20]. *Attention* pada awalnya diperkenalkan untuk meniru pikiran manusia, yaitu secara selektif fokus pada hal yang relevan dan mengabaikan yang lain [19]. Model ini berbeda dari arsitektur Seq2Seq biasa karena *Transformer* tidak memerlukan *Recurrent Neural Networks* (RNN). Dalam struktur RNN, terdapat kekurangan dimana sebagian informasi dari urutan *input* hilang saat *encoder* mengkompresi urutan tersebut menjadi vektor konteks. Untuk mengatasi hal ini, *attention mechanism* diperkenalkan [21].

Seperti yang ditunjukan pada Gambar 2, *Transformer* terdiri dari blok *encoder* dan *decoder*. Pada bagian blok *encoder* terdiri dari lapisan *Multi-Head Attention* dan laposan *Feed Forward*. Blok *decoder* memiliki lapisan yang kurang lebih mirip dengan *encoder* namun memiliki lapisan tambahan, yaitu *Masked Multi-Head Attention*. Hal ini memungkinkan model melihat posisi lain di *input* untuk menghasilkan *encoding* yang lebih baik [20].

Feature embedding dan *word embedding* digunakan untuk mengubah fitur visual dari gambar dan kata-kata dalam *caption* menjadi vektor. Pertama, gambar diekstrak dan diubah menjadi vektor menggunakan InceptionV3. Kemudian, kata-kata dalam *caption* diubah menjadi vektor menggunakan dengan penambahan positional encoding untuk memperhitungkan posisi relatif kata-kata dalam kalim. Proses ini memungkinkan model untuk memahami konteks dan urutan kata-kata dalam kalimat. Selanjutnya, model memilih kata-kata berikutnya dalam caption sampai caption lengkap dihasilkan.

Setelah melewati serangkaian lapisan decoder, output terakhir dikirimkan ke linear layer untuk diubah menjadi representasi probabilitas. Dengan menggunakan prediksi probabilitas ini, model secara iteratif memutuskan kata-kata berikutnya dalam caption, satu demi satu, sampai caption lengkap telah dihasilkan.

Beberapa *hyperparameter* diatur sesuai dengan nilai *default* yang direkomendasikan oleh penelitian sebelumnya, seperti EMBED_DIM, FF_DIM, dan NUM_HEADS, yang diambil dari penelitian "*Attention is All You Need*" [20]. Sedangkan *hyperparameter* lainnya, seperti VOCAB_SIZE dan SEQ_LENGTH, disesuaikan dengan ukuran dataset, sementara EPOCHS dan BATCH_SIZE disesuaikan dengan kemampuan komputasi yang tersedia [22].

Tabel 1. *Hyperparameter* dalam penelitian

Hyperparameters	***Value***
VOCAB_SIZE	10000
SEQ_LENGTH	25
EMBED_DIM	512
NUM_HEADS	8
FF_DIM	512
BATCH_SIZE	64
EPOCHS	50

Optimizer yang digunakan dalam penelitian ini adalah Adam *optimizer* dengan *learning rate scheduler* yang menggunakan fase pemanasan (*warmup*). Ini memungkinkan *learning rate* untuk meningkat secara bertahap dari nilai awal hingga mencapai nilai yang ditetapkan selama fase *warmup*, dan tetap konstan setelah fase pemanasan selesai, hal tersebut dinyatakan pada rumus 2.

$$lrate = d_{model}^{-0.5} \cdot (step_{num}{}^{-0.5}, step_{num} \cdot warmup_{steps}{}^{-1.5}) \quad (1)$$

Dalam rumus tersebut, $step\_num$ merupakan total *training step* dan $warmup\_steps$ merupakan *total step warmup* yang diatur pada awal pelatihan untuk mengurangi dampak ketidakseimbangan awal model. Nilai $warmup\_steps$ diambil dari sebagian $step\_num$ yang mengalikan jumlah panjang dataset pelatihan dengan *epochs*. Setelah fase *warmup* selesai, digunakan *post_warmup_learning_rate* sebesar 1e-4 untuk mengontrol seberapa cepat atau lambat model akan belajar. Kriteria *early stopping* digunakan untuk menghentikan pelatihan jika tidak terjadi peningkatan pada nilai *loss* dalam tiga *epoch* berturut-turut.

4. Model Evaluation Metrics

Penelitian ini menggunakan empat *evaluation metrics* yang berbeda untuk mengevaluasi kinerja dari model *image captioning*. *Evaluation Metrics* yang digunakan adalah BLEU-n, ROUGE-L, METEOR, dan CIDEr. *Metric* tersebut adalah penilaian yang umum digunakan [19]. Dalam mengevaluasi teks yang dihasilkan, digunakan dua jenis teks yaitu kandidat dan referensi. Kandidat merujuk pada teks yang dihasilkan oleh model, sedangkan referensi mengacu pada teks yang telah dianotasi oleh manusia atau yang berada di dalam dataset. Metrik evaluasi berfungsi dengan membandingkan kandidat dengan referensi dalam hal kesamaan teks dengan kalimat manusia atau kebenaran semantik[23]. Semakin tinggi skornya, semakin terkait teks prediksi tersebut dengan teks aslinya.

1. BLEU

   BLEU (*Bilingual Evaluation Understudy*) merupakan metrik evaluasi yang mencocokan kesamaan antara teks yang diprediksi dengan referensi dari teks tersebut [24]. Skor BLEU berkisar antara 0 dan 1, di mana skor 1 menunjukkan kesamaan sempurna dengan teks referensi, sementara skor 0 menunjukkan ketidakmiripan total. Metrik BLEU menggunakan n-grams, yaitu urutan kata yang terdiri dari beberapa kata berturut-turut. Untuk BLEU, panjang maksimal n-grams yang dihitung adalah 4, karena hal ini ditemukan memiliki korelasi yang tinggi dengan preferensi manusia. Persamaan BLEU dihitung dengan membandingkan jumlah n-grams yang cocok antara teks hasil prediksi dan teks referensi. Kemudian, hasilnya dinormalisasi agar skor BLEU tetap berada dalam rentang 0 hingga 1. Skor ini dihitung menggunakan rumus:

$$= \min\left(1, \frac{output-length}{reference-length}\right) \left(\prod_{i-1}^{4} precision_i\right)^{\frac{1}{4}} \quad (2)$$

2. METEOR

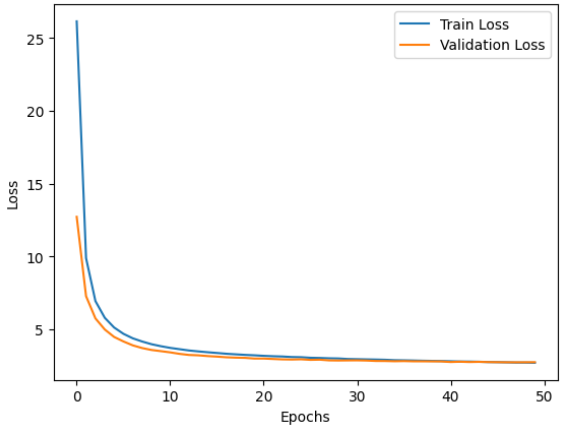
   METEOR [25] atau *Metric for Evaluation for Translation with Explicit Ordering* adalah metrik evaluasi yang dirancang untuk mengevaluasi kualitas terjemahan teks, baik dari model mesin terjemahan maupun model *caption*ing gambar. Tujuan METEOR adalah untuk mengatasi kekurangan metrik BLEU dengan fokus pada presisi tunggal (*single-precision*) dan recall kata. Perhitungan ini membutuhkan METEOR untuk menggunakan kumpulan penyesuaian yang telah ditentukan sebelumnya, khususnya, tesaurus WordNet, untuk mempertimbangkan kata, akar kata, dan sinonim.
3. ROUGE-L

   ROUGE [26] (*Recall-Oriented Understudy for Gisting Evaluation – Longest Common Subsequence*) adalah sebuah metrik yang digunakan untuk membandingkan unit-unit dasar seperti n-gram, urutan kata, dan pasangan kata antara *caption* yang diprediksi dan referensi dalam proses evaluasi. Salah satu metode evaluasi dari seri ROUGE adalah ROUGE-L, yang berfokus pada *Longest Common Subsequence* (LCS) pada tingkat kalimat. Metode ini tidak memerlukan pencocokan kata secara berkelanjutan, sehingga memudahkan dalam mengevaluasi kecocokan antara teks yang dihasilkan dan referensi.
4. CIDEr

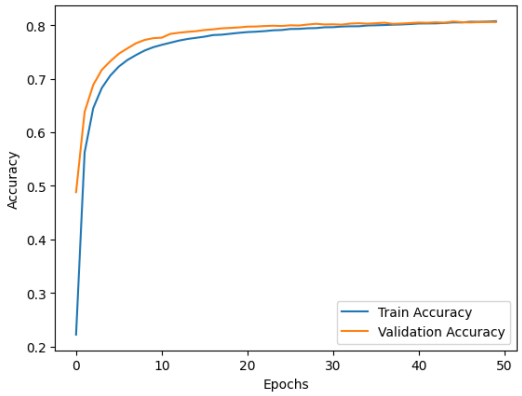
   CIDEr [27] (*Consensus-based Image Description Evaluation*) adalah salah satu metrik evaluasi yang digunakan dalam penilaian deskripsi gambar. Metrik ini mempertimbangkan setiap kalimat sebagai kumpulan n-gram, yaitu urutan kata-kata dalam teks. N-gram ini kemudian diolah, dan bobot untuk masing-masing n-gram dihitung menggunakan teknik *term frequency-inverse document frequencies* (TF-IDF). Pendekatan ini memungkinkan CIDEr untuk fokus pada kata-kata yang penting dan signifikan dalam mengevaluasi kemiripan antara teks yang diprediksi oleh model dengan teks referensi manusia. Berbeda dengan metode lain seperti BLEU, yang memberikan bobot yang sama pada setiap kata dalam kalimat, CIDEr menggunakan TF dan IDF untuk mengevaluasi kata-kata yang lebih relevan dengan konteks.

C. Hasil dan Pembahasan

Pelatihan ini dilakukan di Google Colab Pro dengan menggunakan GPU A100 yang memakan waktu pelatihan selama 45 menit dan berdasarkan *setup* yang telah diatur sebelumnya pada Tabel 1. Setelah proses pelatihan, berikut merupakan hasil dari riwayat pelatihan untuk tingkat *accuracy* dan tingkat *loss* pada model untuk data *training* dan *validation*.



Gambar 4. *Training* dan *Validation Loss*



Gambar 5. *Training* dan *Validation Accuracy*

Gambar 4 adalah kurva yang menunjukkan tingkat *loss* dari model *image captioning* pada data *training* dan *validation* dalam penelitian ini. Pada awal pelatihan, tingkat *loss* yang tinggi pada *data training*, mencapai 26%, menunjukkan prediksi yang belum tepat dan tingkat kesalahan yang tinggi. Namun, seiring berjalannya waktu, terjadi penurunan signifikan pada *epoch* ke-3, mencapai 6%. Hingga mencapai *epoch* ke-50, tingkat *loss* pada *data training* turun menjadi 2%. Pada sisi *validation*, terjadi penurunan signifikan pada *epoch* ke-3, dengan tingkat *loss* awal sebesar 12%, turun menjadi 2% pada akhir *epoch* ke-50. Meskipun penurunan *loss* dari *epoch* ke-3 hingga ke-50 berlangsung lambat, penurunan yang stabil menandakan pendekatan konvergensi model dan tingkat kinerja yang memadai.

Gambar 5 menampilkan kurva yang memperlihatkan tingkat akurasi dari model *image captioning* pada *data training* dan *validation* dalam penelitian ini. Pada awal pelatihan, tingkat akurasi pada *data training* rendah, hanya mencapai 10%. Namun, pada *epoch* kedua, terjadi peningkatan signifikan hingga 52%. Seiring dengan berjalannya iterasi, kurva *accuracy* mulai meningkat secara bertahap. Pada *epoch* ke-50, tingkat *accuracy* meningkat hingga mencapai 80%. Hal ini menunjukkan peningkatan dalam kemampuan model untuk memahami hubungan antara gambar dan teks. Pada sisi *validation*, tingkat akurasi awal sebesar 48%, namun meningkat secara bertahap hingga mencapai 80% pada akhir *epoch* ke-50.

Peningkatan yang stabil menandakan kemajuan model dalam memperbaiki prediksi dan pemahaman atas hubungan antara gambar dan teks.

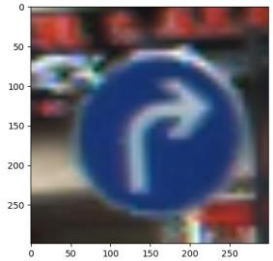
Berdasarkan Gambar 4 dan Gambar 5 di atas, kurva *accuracy* dan *loss* pada *data validation* memiliki kurva yang sejajar dengan data *training* dan dapat dikatakan bahwa model ini mencapai tingkat *goodfitting*. Hal ini menandakan bahwa model dalam penelitian ini tidak hanya mampu mempelajari pola-pola spesifik dalam *data training*, tetapi juga mampu melakukan generalisasi dengan baik pada data yang tidak pernah dilihat sebelumnya, sehingga memungkinkan untuk memberikan prediksi yang akurat pada data baru. Ini dapat dijadikan indikasi yang kuat bahwa model dalam penelitian ini dapat diandalkan dalam tugas *image captioning* khususnya pada dataset rambu lalu lintas di Indonesia.

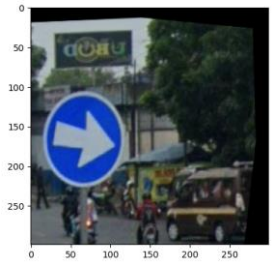
Tabel 2. Hasil Evaluasi Model *Image Captioning* untuk rambu lalu lintas di Indonesia menggunakan *Pretrained* CNN (Inception V3) dan *Transformer*

Metrik Evaluasi	Skor
BLEU-1	0.89
BLEU-2	0.82
BLEU-3	0.75
BLEU-4	0.69
CIDEr	0.57
ROUGE-L	0.25
METEOR	0.26

Model dinilai menggunakan *data validation* berjumlah 1870 gambar. Data ini tidak digunakan dalam tahap pelatihan model, sehingga dapat digunakan untuk mengevaluasi kemampuan model secara independen. Sesuai dengan yang tertulis pada Tabel 2, model yang dikembangkan mampu untuk meraih skor BLEU-1 sebesar 0.89, skor BLEU-2 sebesar 0.82, skor BLEU-3 sebesar 0.75, skor BLEU-4 sebesar 0.69, skor CIDEr sebesar 0.57, skor ROUGE-L sebesar 0.25, dan skor METEOR meraih angka 0.26.

Tabel 3. Hasil *caption* dari model

No	Gambar	Generated *Caption*
1		instruksi untuk berhenti sebentar pastikan jalan bebas risiko sebelum meneruskan
2		pastikan kendaraan bergerak sesuai dengan arah yang tertera

3		rambu yang menunjukkan lajur yang harus dilewati untuk memfasilitasi evakuasi dalam keadaan darurat

Tabel 3 menunjukkan beberapa hasil *caption* dari model pada penelitian ini. Seperti yang dapat dilihat pada tabel bahwa model yang dibangun telah berhasil membuat *caption* gambar yang cukup baik. Model ini mampu menghasilkan *caption* yang sesuai konteks dari gambar yang diberikan.

D. Simpulan

Penelitian ini bertujuan untuk membangun model *image captioning* untuk rambu lalu lintas Indonesia menggunakan Inception V3 dan *Transformer*. Tujuannya adalah untuk meningkatkan pemahaman tentang pengembangan model *image captioning* untuk rambu lalu lintas Indonesia. Data yang digunakan adalah dataset gabungan dari GTSRB dan *Traffic Sign in Indonesia* dari *platform* Roboflow, dengan total 9.594. Model dilatih dengan menggunakan Inception V3 sebagai *encoder* dan *Transformer* sebagai *decoder*.

Hasil evaluasi model menunjukkan bahwa model yang dikembangkan mampu menghasilkan deskripsi gambar dengan baik dengan skor BLEU-1 sebesar 0.89, BLEU-2 sebesar 0.82, BLEU-3 sebesar 0.75, dan BLEU-4 sebesar 0.69, CIDEr sebesar 0.57, ROUGE-L sebesar 0.25, dan METEOR sebesar 0.26. Hasil ini menunjukkan kemampuan model untuk menghasilkan deskripsi yang semakin mirip dengan deskripsi manusia khususnya pada gambar-gambar lalu lintas di Indonesia.

Selain itu, melalui proses pelatihan dan evaluasi model, terlihat bahwa model mampu belajar pola-pola spesifik dalam *data training* dan mampu melakukan generalisasi dengan baik pada *data validation* yang tidak pernah dilihat sebelumnya. Hal ini menunjukkan kemampuan model untuk memberikan prediksi yang akurat pada data baru.

Dengan demikian, penelitian ini memberikan kontribusi penting dalam pengembangan model *image captioning* khususnya untuk rambu lalu lintas Indonesia. Hasilnya dapat digunakan sebagai dasar untuk pengembangan dan penelitian lebih lanjut dalam meningkatkan pemahaman dan keselamatan pengguna jalan melalui komunikasi pesan rambu lalu lintas.

E. Referensi

[1] E. Kirmizioglu and H. Tuydes-Yaman, “Comprehensibility of traffic signs among urban drivers in Turkey,” *Accid Anal Prev*, vol. 45, pp. 131–141, Mar. 2012, doi: 10.1016/j.aap.2011.11.014.

[2] *Undang-undang (UU) Nomor 22 Tahun 2009 tentang Lalu Lintas Dan Angkutan Jalan*. 2009. Accessed: Apr. 24, 2024. [Online]. Available: https://peraturan.bpk.go.id/Details/38654/uu-no-22-tahun-2009

[3] MMDA: Metro Manila, “Accomplishment Report of 2017,” 2017.

[4] J. Sarkar, R. Ramachandran, and S. Kizhiyedath, "*Caption* Generation for Traffic Signs Using a Deep Neural Scheme," *SAE Int J Adv Curr Pract Mobil*, vol. 5, no. 4, Oct. 2022, doi: 10.4271/2022-28-0006.

[5] F. M. Shah, M. Humaira, M. A. R. K. Jim, A. S. Ami, and S. Paul, "Bornon: Bengali Image *Caption*ing with *Transformer*-based Deep learning approach," *ArXiv*, vol. 3, pp. 1–16, 2021.

[6] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on Attention for Image *Caption*ing," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2019, pp. 4633–4642. doi: 10.1109/ICCV.2019.00473.

[7] A. Alsayed, M. Arif, T. M. Qadah, and S. Alotaibi, "A Systematic Literature Review on Using the Encoder-Decoder Models for Image *Caption*ing in English and Arabic Languages," *Applied Sciences*, vol. 13, no. 19, Sep. 2023, doi: 10.3390/app131910894.

[8] P. Chun, T. Yamane, and Y. Maemura, "A deep learning-based image *caption*ing method to automatically generate comprehensive explanations of bridge damage," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 11, pp. 1387–1401, Sep. 2022, doi: 10.1111/mice.12793.

[9] O. Abbas and J. Dang, "Using image *caption*ing for automatic post-disaster damage detection and identification," *Intelligence Informatics and Infrastructure*, vol. 4, no. 2, pp. 66–74, 2023.

[10] B. Das, R. Pal, M. Majumder, S. Phadikar, and A. A. Sekh, "A Visual Attention-Based Model for Bengali Image *Caption*ing," *SN Comput Sci*, vol. 4, no. 2, Feb. 2023, doi: 10.1007/s42979-023-01671-x.

[11] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The {G}erman {T}raffic {S}ign {R}ecognition {B}enchmark: A multi-class classification competition," in *IEEE International Joint Conference on Neural Networks*, 2011, pp. 1453–1460.

[12] UMY, "Traffic sign in indonesia Dataset," \url{ https://universe.roboflow.com/umy-35d0e/traffic-sign-in-indonesia-1j13y }.

[13] T. Hidayat, A. Yani, and J. A. Barata, *Buku Petunjuk Tata Cara Berlalu Lintas (Highway Code) di Indonesia*. Direktorat Jenderal Perhubungan Darat Departemen Perhubungan, 2005.

[14] D. Setiawan, M. A. C. Saffachrissa, S. Tamara, and D. Suhartono, "Image *Caption*ing with Style Using Generative Adversarial Networks," *JOIV : International Journal on Informatics Visualization*, vol. 6, no. 1, Mar. 2022, doi: 10.30630/joiv.6.1.709.

[15] Maryamah, N. A. Alya, M. H. Sudibyo, E. Liviani, and R. I. Thirafi, "Image Classification on Fashion Dataset Using Inception V3," *Journal of Advanced Technology and Multidiscipline*, vol. 2, no. 1, pp. 10–15, May 2023, doi: 10.20473/jatm.v2i1.44131.

[16] V. Maeda-Gutiérrez *et al.*, "Comparison of Convolutional Neural Network Architectures for Classification of Tomato Plant Diseases," *Applied Sciences*, vol. 10, no. 4, Feb. 2020, doi: 10.3390/app10041245.

[17] Google Cloud, "Advanced Guide to InceptionV3," Cloud TPU.

[18] P. Dandwate, C. Shahane, V. Jagtap, and S. C. Karande, "Comparative study of *Transformer* and LSTM Network with attention mechanism on Image *Caption*ing," *ArXiv*, Mar. 2023.

[19] D. H. Fudholi and R. A. N. Nayoan, "The Role of *Transformer*-based Image *Caption*ing for Indoor Environment Visual Understanding," *International Journal of Computing and Digital Systems*, vol. 12, no. 3, pp. 479–488, Aug. 2022, doi: 10.12785/ijcds/120138.

[20] A. Vaswani *et al.*, "Attention Is All You Need," Jun. 2017.

[21] D. I. Lee, J. H. Lee, S. H. Jang, S. J. Oh, and I. C. Doo, "Crop Disease Diagnosis with Deep Learning-Based Image *Caption*ing and Object Detection," *Applied Sciences*, vol. 13, no. 5, p. 3148, Feb. 2023, doi: 10.3390/app13053148.

[22] U. A. A. Al-faruq, "Implementasi Arsitektur *Transformer* Pada Image *Caption*ing Dengan Bahasa Indonesia," Universitas Islam Indonesia, 2021.

[23] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj, and R. K. Mishra, "Image *Caption*ing: A Comprehensive Survey," in *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, IEEE, Feb. 2020, pp. 325–328. doi: 10.1109/PARC49193.2020.236619.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Morristown, NJ, USA: Association for Computational Linguistics, 2001, pp. 311–318. doi: 10.3115/1073083.1073135.

[25] M. Denkowski and A. Lavie, "Meteor Universal: Language Specific Translation Evaluation for Any Target Language," in *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 376–380. doi: 10.3115/v1/W14-3348.

[26] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *In Text Summarization Branches Out*, Barcelona: Association for Computational Linguistics, 2004, pp. 74–81.

[27] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," Nov. 2014.