TESIS

KETERANGAN GAMBAR OTOMATIS BAHASA INDONESIA MENGGUNAKAN CNN DENGAN PENDEKATAN MODEL DEEP LEARNING BERBASIS TRANSFORMER

Disusun oleh:

Nama	**: Rifqi Mulyawan**
NIM	**: 21.55.1029**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2023

TESIS

KETERANGAN GAMBAR OTOMATIS BAHASA INDONESIA MENGGUNAKAN CNN DENGAN PENDEKATAN MODEL DEEP LEARNING BERBASIS TRANSFORMER

AUTOMATIC INDONESIAN IMAGE CAPTIONING USING CNN WITH TRANSFORMER-BASED DEEP LEARNING MODEL APPROACH

Diajukan melalui Jalur Jurnal Bereputasi
untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama	**: Rifqi Mulyawan**
NIM	**: 21.55.1029**
Konsentrasi	**: Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA

2017

HALAMAN PENGESAHAN

KETERANGAN GAMBAR OTOMATIS BAHASA INDONESIA MENGGUNAKAN CNN DENGAN PENDEKATAN MODEL DEEP LEARNING BERBASIS TRANSFORMER

AUTOMATIC INDONESIAN IMAGE CAPTIONING USING CNN WITH TRANSFORMER-BASED DEEP LEARNING MODEL APPROACH

Dipersiapkan dan Disusun oleh

Rifqi Mulyawan

21.55.1029

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Kamis, 02 Februari 2023

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 02 Februari 2023
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

KETERANGAN GAMBAR OTOMATIS BAHASA INDONESIA MENGGUNAKAN CNN DENGAN PENDEKATAN MODEL DEEP LEARNING BERBASIS TRANSFORMER

AUTOMATIC INDONESIAN IMAGE CAPTIONING USING CNN WITH TRANSFORMER-BASED DEEP LEARNING MODEL APPROACH

Dipersiapkan dan Disusun oleh

Rifqi Mulyawan

21.55.1029

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Kamis, 02 Februari 2023

Pembimbing Utama

Dr. Andi Sunyoto, M. Kom.
NIK. 190302052

Pembimbing Pendamping

Alva Hendi Muhammad, S.T., M. Eng., Ph.D.
NIK. 190302493

Anggota Tim Penguji

Dr. Arief Setyanto, S.Si., M.T.
NIK. 190302036

Dr. Kumara Ari Yuana, S.T., M.T.
NIDN. 0515067101

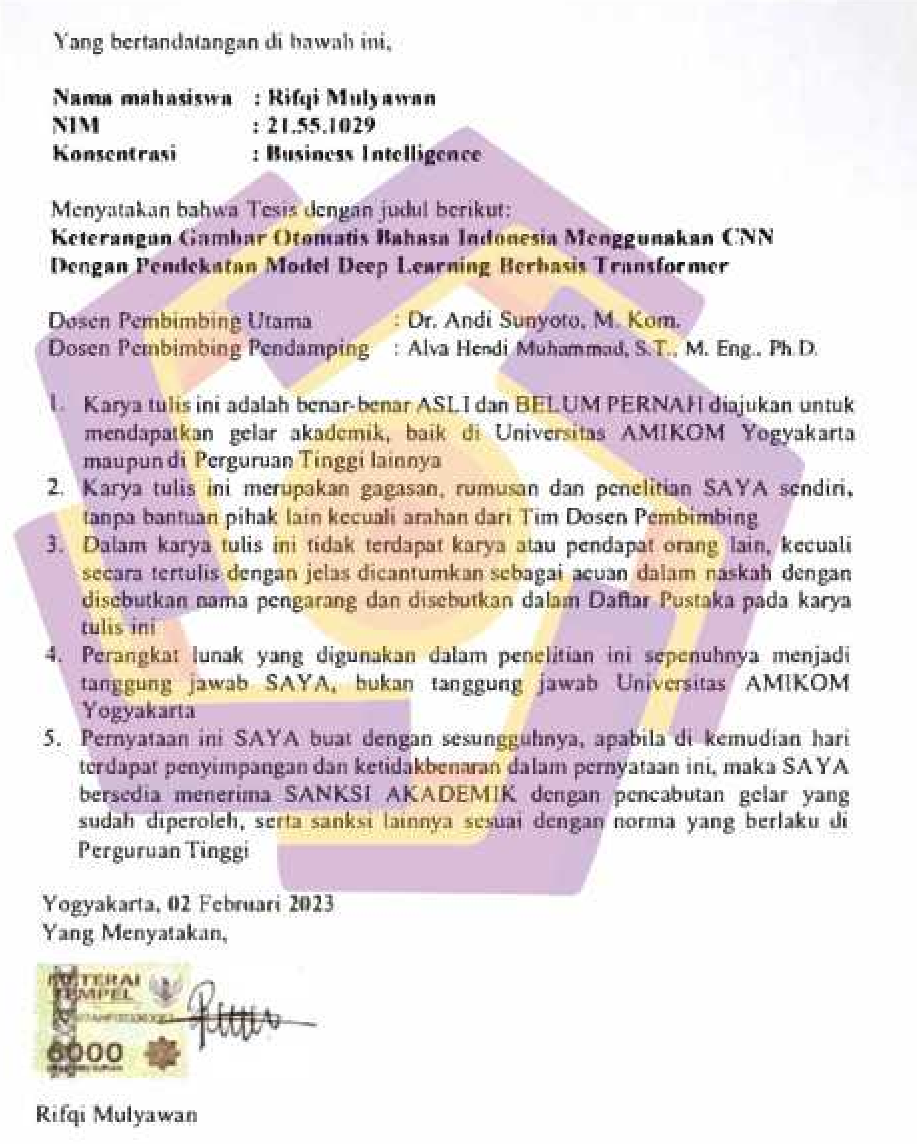
Dr. Andi Sunyoto, M.Kom.
NIK. 190302052

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 02 Februari 2023
Direktur Program Pascasarjana

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS



Yang bertandatangan di bawah ini,

Nama mahasiswa : Rifqi Mulyawan
NIM : 21.55.1029
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Keterangan Gambar Otomatis Bahasa Indonesia Menggunakan CNN Dengan Pendekatan Model Deep Learning Berbasis Transformer

Dosen Pembimbing Utama : Dr. Andi Sunyoto, M. Kom.
Dosen Pembimbing Pendamping : Alva Hendi Muhammad, S.T., M. Eng., Ph.D.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 02 Februari 2023
Yang Menyatakan,

TERAI TEMPEL
6000

Rifqi Mulyawan

HALAMAN PERSEMBAHAN

Dengan segala puji dan syukur kepada Allah Subhanahu Wa Ta'ala dan atas dukungan serta doa dari orang-orang tercinta, akhirnya tesis ini dapat terselesaikan dengan baik. Oleh karena itu, dengan rasa bangga dan bahagia penulis persembahkan tesis ini kepada:

1. Orang tua penulis, Bpk. Akhyadi dan Ibu Ani Noor Asifah yang telah merawat, mendidik, men-*support*, dan selalu mendoakan atas kebahagiaan serta kesuksesan penulis.
2. Istri dan anak penulis, Noor Hasanah Ananda dan Ahmad Adlan Malik yang telah memberikan dukungan moril maupun materi serta doa yang tiada henti untuk kesuksesan penulis.
3. Mertua penulis, Bpk. M. Yanto dan Ibu Rahmaniah yang selalu mendukung dan mendoakan penulis.
4. Para Paman-Paman dan Acil-Acil penulis yang selalu memberikan doa dan semangat kepada penulis.
5. Bapak Dr. Andi Sunyoto, M.Kom. dan Bpk. Alva Hendi Muhammad, S.T. M.Eng., Ph.D. selaku dosen pembimbing dan dosen-dosen penguji penulis yang telah memberikan arahan serta saran sehingga tesis ini dapat terselesaikan dengan baik.
6. Keluarga besar kelas PJJ Magister Teknik Informatika Angkatan 2021, terima kasih atas bantuan dan kebersamaan yang sangat berarti bagi penulis.

7. Grup Proyek *Data Science*, terima kasih atas bantuan dan kebersamaan yang sangat berarti bagi penulis.
8. Terima kasih atas semangat dan dukungan dari semua pihak, terutama pihak keluarga penulis. Semoga tesis ini dapat bermanfaat dan berguna di masa yang akan datang. Aamiin.

HALAMAN MOTTO

"Yooo! Be Grateful."

Yooo! Bersyukurlah.

KATA PENGANTAR

Puji syukur penulis panjatkan kehadirat Allah Subhanallau Wata'ala yang telah memberikan kesehatan jasmani dan rohani serta petunjuk dan kekuatan kepada penulis sehingga tesis yang berjudul "Keterangan Gambar Otomatis Bahasa Indonesia Menggunakan CNN Dengan Pendekatan Model Deep Learning Berbasis Transformer" dapat terselesaikan dengan baik. Kritik dan saran sangat diharapkan penulis agar dapat lebih baik lagi di kemudian hari. Dalam penyusunan dan penulisan tesis ini tidak terlepas dari bantuan dan bimbingan serta dukungan dari berbagai pihak. Oleh karena itu dalam kesempatan ini penulis dengan senang hati menyampaikan terima kasih kepada:

1. Prof. Dr. M. Suyanto, M.M. selaku rektor Universitas AMIKOM Yogyakarta.
2. Prof. Dr. Kusrini, M.Kom. selaku Direktur Program Pascasarjana Universitas AMIKOM Yogyakarta yang telah menunjuk dosen pembimbing sehingga memperlancar penulis dalam menyelesaikan tesis.
3. Bapak Dr. Andi Sunyoto, M.Kom. dan Bpk. Alva Hendi Muhammad, S.T. M.Eng., Ph.D. selaku pembimbing utama dan pendamping yang telah mencurahkan perhatian, bimbingan, nasihat, doa, dan kepercayaan yang sangat berarti bagi penulis serta telah meluangkan waktu dalam memberikan arahan dan masukan selama penelitian dan penyusunan tesis.

4. Dosen-dosen penguji yang telah memberikan saran dan masukan yang membangun.
5. Orang tua, istri dan anak, adik, paman-paman, tante-tante, mertua, adik yang telah memberikan doa dan motivasi sehingga menjadi penyemangat bagi penulis dalam mengerjakan tesis.
6. Teman-teman yang telah memberikan semangat dan dukungan sehingga tesis ini dapat terselesaikan dengan baik.
7. Semua pihak yang telah membantu, baik secara langsung maupun tidak langsung yang tidak dapat penulis sebutkan satu per satu.
8. Akhir kata, semoga tesis ini bermanfaat, khususnya bagi penulis dan umumnya bagi masyarakat dalam rangka menambah wawasan pengetahuan.

Yogyakarta, 20 Januari 2023

Penulis

BAB I

PENDAHULUAN

Computer Vision (CV) telah membuat kemajuan yang signifikan dalam gambar pengolahan, seperti kategorisasi gambar dan objek pengakuan. Keterangan gambar yang dikenal dengan *state-of-the-art problem* "Image Captioning" menggunakan klasifikasi dan pengenalan objek untuk membuat satu atau lebih teks teks yang menggambarkan konten secara visual dan otomatis sebuah gambar.

Image Captioning (IC) atau yang dikenal dengan keterangan gambar dalam bahasa Indonesia ini memiliki beberapa manfaat di berbagai bidang untuk dunia nyata. Salah satu manfaatnya adalah dapat membantu membuat informasi visual lebih mudah diakses oleh orang-orang dengan gangguan penglihatan, dengan membuat deskripsi teks dari gambar secara otomatis. IC juga dapat membantu meningkatkan aksesibilitas web dan menjadikannya lebih inklusif bagi pengguna penyandang disabilitas. Selain itu, keterangan gambar dapat digunakan dalam berbagai aplikasi, seperti pada mobil *self-driving*, untuk memberikan deskripsi tekstual tentang pemandangan di depan kendaraan. IC juga dapat digunakan dalam sistem pengawasan untuk secara otomatis menghasilkan keterangan tentang apa yang terjadi di tempat kejadian yang sedang dipantau. Aplikasi potensial lain dari keterangan gambar termasuk di mesin pencari, untuk membantu pengguna menemukan gambar dengan lebih mudah, dan di media sosial, untuk membantu pengguna memahami konten gambar yang dibagikan oleh orang lain.

Banyak algoritme yang tersedia saat ini untuk menyampaikan esensi gambar dengan kata-kata didasarkan pada arsitektur *encoder-decoder*, di mana infrastruktur *decoder* dapat memprediksi kata-kata dengan memanfaatkan fungsi yang diterima dari jaringan *encoder*. Studi tentang pemberian keterangan pada gambar ini secara keseluruhan mengambil konsep dari *Machine Translation*.

Membuat keterangan gambar dapat digunakan untuk berbagai tujuan, termasuk mengotomatiskan mengemudi mobil, mengembangkan sistem pengenalan wajah, mencirikan individu dengan gangguan penglihatan, meningkatkan kualitas kueri foto, dan banyak lagi. Tugas yang menantang dalam mengembangkan deskripsi bahasa alami dari informasi dalam gambar berada dalam antarmuka CV untuk ekstraksi fitur gambar dan menghasilkan urutan menggunakan teknologi *Natural Language Processing* (NLP). Tugas tersebut telah memberikan dampak yang signifikan di beberapa bidang, seperti pencarian gambar juga berbagai disiplin ilmu, seperti pengembangan perangkat lunak untuk penyandang disabilitas, pengawasan dan keamanan video, dan antarmuka antara manusia dan komputer.

Sebagai salah satu masalah populer yang melibatkan pemodelan urutan teks, SOTA IC untuk pembuatan keterangan foto menggunakan berbagai pendekatan jaringan saraf tiruan. Sebagai contoh, *Convolutional Neural Network*, ConvNet, atau yang dikenal sebagai CNN ini diterapkan dengan arsitektur bahasa lain, seperti *Recurrent Neural Network* (RNN), sebagai pendekatan kerangka kerja berbasis CNN-RNN. Pekerjaan ini menggunakan arsitektur standar menggunakan model CNN yang sudah dilatih sebelumnya (*pre-trained*) untuk membangun vektor fitur.

Mereka kemudian dimasukkan ke dalam RNN sebagai *decoder* yang bertanggung jawab untuk menghasilkan deskripsi bahasa.

Struktur berulang dari RNN, juga tipe RNN yang ditingkatkan, seperti *Long Short Term Memory* (LSTM), membuatnya lebih sulit untuk dilatih karena sifatnya yang berurutan, di mana pada akhirnya, implementasinya dalam IC dapat menghasilkan skor evaluasi yang lebih rendah pada model berbasis RNN standar. Namun, masalah paralelisme yang menjadi kekurangan pada RNN-*based* akhirnya diatasi oleh model SOTA Transformer *neural network*. Karena arsitektur tersebut dibangun di atas mekanisme perhatian (*attention mechanism*) yang dapat membaca sebuah konteks, model tersebut juga dapat beroperasi secara paralel selama fase pelatihan dan tidak memerlukan urutan tertentu.

Dalam masalah keterangan gambar dalam bahasa Indonesia, pendekatan GRU pada penelitian sebelumnya untuk pembuatan keterangan gambar berbahasa Indonesia digunakan untuk mengatasi beberapa masalah yang ada pada RNN. Namun, karena modelnya masih berbasis RNN, temuan mereka menunjukkan bahwa model tersebut kurang memahami konteks dan menyatakan bahwa perlunya implementasi penelitian SOTA untuk pembuatan *sequence* dalam bahasa Indonesia. Penelitian lainnya untuk pembuatan teks bahasa Indonesia pada gambar juga menggunakan CNN dengan arsitektur pra-pelatihan VGG-16 untuk *encoder* model dengan tipe RNN lain yaitu LSTM sebagai *decoder* pada modelnya, tetapi tanpa menyelidiki dampak ekstraksi fitur pada pengukuran kualitas teks gambar untuk kinerja model. Temuan mereka menunjukkan bahwa hasil model memiliki skor evaluasi yang lebih baik dengan BLEU 1-4 (masing-masing 50.00, 31.40, 23.90, 13.10).

Penelitian sebelumnya dalam menghasilkan keterangan gambar dalam bahasa Indonesia menyisakan ruang untuk mengeksplorasi efek menggunakan lapisan CNN pra-terlatih lainnya dengan pendekatan SOTA berbasis Transformer yang dapat mengatasi kelemahan model berbasis RNN dengan sifat paralelnya dan dapat memahami konteks (*context-aware*) untuk mendapatkan model yang lebih baik.

BAB II

PUBLIKASI PERTAMA

1. Jurnal

   2022 5th International Conference on Information and Communications Technology (ICOIACT), Status: Published

2. Judul

   Automatic Indonesian Image Captioning using CNN and Transformer-Based Model Approach

3. Abstrak

   Image captioning melibatkan pembuatan teks frase atau lebih untuk deskripsi konten visual dari gambar. *Caption generation* untuk sebuah gambar dianggap penting untuk membantu aktivitas manusia dalam memahami materi visual, seperti *caption* pada gambar medis, kontak manusia dengan robot, dan membantu penyandang tunanetra menjelaskan visual. Studi kami bertujuan untuk membuat deskripsi gambar bahasa Indonesia dan menilai seberapa sukses pendekatan teks tersebut. Kumpulan data Flickr8k yang diterjemahkan disediakan untuk investigasi ini. Penelitian ini menggunakan dua metode untuk pembuatan teks dari sebuah foto: CNN dengan ResNet sebagai *encoder*, dan Transformer, mekanisme berbasis perhatian diri sebagai *decoder*. Dengan menggunakan kumpulan data khas Indonesia, bernama Flickr8k Bahasa, kami menggunakan pendekatan berbasis Transformer untuk membuat model pembuatan teks bahasa Indonesia dari foto. Kami menunjukkan bahwa strategi berbasis Transformer Indonesia mengungguli yang lama, di mana hasil terbaik diperoleh dengan BLEU-1 hingga 4, METEOR, ROUGE_L, CIDEr masing-masing 56.00, 41.17, 29.42, 20.57, 19.50, 44.16, 57.26. Selain membandingkan kinerja model menggunakan model pra-pelatihan CNN yang berbeda, model CNN yang lebih besar tidak menjamin peningkatan akurasi selama lima puluh perulangan proses pelatihan.

4. Masalah Penelitian

   - Model apa yang performanya lebih baik dari *Image Captioning* berbasis CNN dan LSTM?
   - Arsitektur *decoder Image Captioning* apa yang performanya lebih baik dalam masalah sekuens pada LSTM?
   - Bagaimana pengaruh model *pre-trained* CNN sebagai *encoder* dalam ekstraksi fitur gambar?
   - Bagaimana pengaruh penyesuaian *hyperparameter* pada *decoder* pada hasil evaluasi model?

5. Tujuan Penelitian

   - Mengembangkan model yang memiliki performa lebih baik untuk masalah dalam *Image Captioning* yang berbasis CNN dan LSTM
   - Mengidentifikasi arsitektur *decoder* apa yang lebih efektif dari model *Image Captioning* yang berbasis CNN dan LSTM

- Mengidentifikasi *pre-trained* model CNN apa yang memiliki hasil yang lebih baik sebagai *encoder* berbasis ResNet dalam model Image Captioning untuk ekstraksi fitur gambar
- Mengidentifikasi *hyperparameter* terbaik dalam masalah *Image Captioning* berbasis Transformer pada *dataset* Flickr8k Bahasa Indonesia.

6. Hasil Penelitian
   - Dihasilkan model *Image Captioning* Bahasa Indonesia dengan menggunakan arsitektur CNN sebagai *encoder* dan Transformer sebagai *decoder*-nya.
   - Arsitektur Transformer sebagai *decoder* lebih efektif karena dapat memanfaatkan paralelisasi saat pelatihan.
   - Diidentifikasi bahwa model *pre-trained* CNN berbasis ResNet50 memiliki skor evaluasi yang lebih baik dibandingkan ResNet18 atau ResNet101.
   - Diidentifikasi bahwa 1 Transformer *Head* dan 3 Transformer *Layer* mendapatkan hasil evaluasi yang lebih baik sebagai *hyperparameter* pada *decoder* model.
7. Kesimpulan Penelitian

   Beberapa hal yang dilaporkan dalam makalah ini dirangkum oleh (1) Riset dan eksperimen dengan model *Image Captioning* bahasa Indonesia, (2) menggunakan *encoder-decoder* sebagai model pembuatan keterangan gambar dalam bahasa Indonesia, (3) menggunakan arsitektur Transformer dengan ResNet-family yang telah dilatih sebelumnya sebagai *decoder*-nya. Implementasi model berbasis Transformer dapat mengungguli desain model sebelumnya menggunakan *dataset* skala kecil hingga menengah karena mereka sepenuhnya menghindari rekursi dengan menganalisis seluruh kalimat dan mempelajari asosiasi antara kata-kata menggunakan teknik *multi-head attention* dan *positional encoding*. Dengan beberapa langkah eksperimental, ditemukan bahwa menggunakan model CNN terlatih yang berbeda seperti ResNet18 dan ResNet50 sebagai dekoder telah meningkatkan skor hasil sementara menggunakan ResNet101 yang lebih besar tidak, karena masalah *overfitting* model adalah masalah yang menyebabkan hasil ini. Adapun untuk penelitian di masa depan, *finetuning* dan menerapkan pendekatan berbasis Transformer ini ke kumpulan data yang berbeda seperti Flickr30k, MS COCO 14, 17, atau kumpulan data terkait pembuatan teks lainnya, juga ada banyak metode berbasis Transformers yang lebih hebat, seperti model berbasis XL, *Entangled*, dan *Meshed Memory Transformer*, yang dapat digunakan untuk mendapatkan hasil yang lebih baik lagi dalam aplikasi pembuat teks otomatis Bahasa Indonesia.

BAB III

PUBLIKASI KEDUA

1. Jurnal

   International Journal on Informatics Visualization (JOIV), Status: Accepted

2. Judul

   Pre-Trained CNN Architecture Analysis for Transformer-Based Indonesian Image Caption Generation Model

3. Abstrak

   Klasifikasi dan pengenalan objek dalam pemrosesan gambar telah mengalami peningkatan yang signifikan dalam bidang visi komputer. Metode ini sering digunakan untuk masalah-masalah yang berkaitan dengan visual, terutama dalam klasifikasi gambar dengan memanfaatkan *Convolutional Neural Network* (CNN). Dalam tugas *state-of-the-art* (SOTA) populer untuk menghasilkan keterangan pada gambar, implementasinya sering digunakan untuk ekstraksi fitur gambar sebagai *encoder*. Alih-alih melakukan klasifikasi langsung, fitur yang diekstraksi ini dikirim dari encoder ke bagian decoder untuk menghasilkan urutan. Jadi, beberapa lapisan CNN yang terkait dengan tugas klasifikasi tidak diperlukan. Penelitian ini bertujuan untuk menentukan arsitektur atau model pra-pelatihan CNN mana yang dapat melakukan yang terbaik dalam mengekstraksi fitur gambar menggunakan model Transformer sebagai dekodernya. Tidak seperti arsitektur Transformer asli, kami menerapkan cara *vector-to-sequence* alih-alih *sequence-to-sequence* untuk modelnya. *Dataset* Flickr8k Bahasa Indonesia digunakan dalam penelitian ini. Evaluasi dilakukan menggunakan beberapa arsitektur terlatih, termasuk ResNet18, ResNet34, ResNet50, ResNet101, VGG16, Efficientnet_b0, Efficientnet_b1, dan Googlenet. Hasil inferensi model kualitatif dan skor evaluasi kuantitatif dianalisis dalam penelitian ini. Hasil pengujian menunjukkan bahwa arsitektur ResNet50 dapat menghasilkan pembuatan *sequence* yang stabil dengan nilai akurasi tertinggi. Dengan beberapa eksperimen, melakukan *finetuning* pada *encoder* dapat meningkatkan skor evaluasi model secara signifikan. Untuk pekerjaan di masa depan, menjelajahi *dataset* yang lebih besar seperti Flickr30k, MS COCO 14, MS COCO 17, dan *dataset* teks gambar lainnya dalam bahasa Indonesia, juga menerapkan metode SOTA berbasis Transformer dapat digunakan untuk mendapatkan model teks gambar otomatis Indonesia yang lebih baik.

4. Masalah Penelitian
    - Model apa yang performanya lebih baik dari *Image Captioning* berbahasa Indonesia berbasis CNN + GRU dan CNN + LSTM?
    - Arsitektur decoder *Image Captioning* apa yang lebih efektif dan dapat memahami konteks urutan dalam masalah sekuens pada arsitektur berbasis RNN (GRU atau LSTM)?
    - Bagaimana pengaruh model *pre-trained* CNN yang berbeda sebagai *encoder* pada kualitas hasil teks yang dihasilkan model?
    - Bagaimana pengaruh CNN dengan *channel size* yang berbeda?

- Bagaimana efek *finetuning* pada *encoder* model CNN dengan Transformer sebagai *decoder*-nya?

5. Tujuan Penelitian

- Mengembangkan model yang memiliki performa lebih baik untuk masalah dalam *Image Captioning* yang berbasis berbasis CNN + GRU dan CNN + LSTM
- Mengidentifikasi arsitektur *decoder* apa yang lebih efektif dan dapat memahami konteks urutan dalam masalah sekuens pada arsitektur berbasis RNN (GRU atau LSTM)
- Mengidentifikasi *pre-trained* model CNN apa yang memiliki hasil yang lebih baik sebagai *encoder* model pada kualitas hasil teks yang dihasilkan model
- Mengidentifikasi pengaruh CNN dengan *channel size* yang berbeda
- Mengidentifikasi efek *finetuning* pada *encoder* model CNN dengan Transformer sebagai *decoder*-nya

6. Hasil Penelitian

- Dihasilkan model *Image Captioning* Bahasa Indonesia dengan menggunakan arsitektur CNN sebagai dan Transformer sebagai *decoder*-nya.
- Arsitektur Transformer sebagai *decoder* lebih efektif karena dapat memanfaatkan paralelisasi saat pelatihan dan dapat memahami konteks urutan dengan mekanisme berbasis perhatian (*attention mechanism*).

- Diidentifikasi bahwa model *pre-trained* CNN berbasis ResNet50 memiliki hasil evaluasi yang lebih baik dibandingkan ResNet18, ResNet34, ResNet101, VGG16, Efficientnet_b0, Efficientnet_b1, dan Googlenet.
- Diidentifikasi bahwa pengaruh CNN dengan *channel size* yang berbeda dapat meningkatkan hasil evaluasi model kecuali untuk ukuran kanal 2048 pada ResNet101 karena masalah terkait *overfitting*.
- Diidentifikasi bahwa efek *finetuning* pada *encoder* model CNN dengan Transformer sebagai *decoder*-nya dapat meningkatkan hasil evaluasi model.

7. Kesimpulan Penelitian

Penelitian ini menggunakan beberapa model CNN yaitu ResNet18, ResNet34, ResNet50, ResNet101, VGG16, Efficientnet_b0, Efficientnet_b1, dan Googlenet, untuk mendapatkan model CNN yang dapat menghasilkan performa terbaik sebagai ekstraktor fitur untuk memprediksi urutan teks yang dilakukan oleh dekoder Transformer. Pengujian dilakukan dengan menggunakan berbagai ukuran CNN Channel, dimana model terbaik diperoleh dengan menggunakan ResNet50 dan terbukti model tersebut dapat menghasilkan teks bahasa Indonesia yang benar secara tata bahasa. Eksperimen menunjukkan bahwa menyempurnakan model pembuat enkode hampir selalu meningkatkan keluaran model dekoder, menghasilkan skor evaluasi yang lebih baik sekitar 2% daripada model CNN lainnya pada validasi. Model ResNet50 direkomendasikan untuk menggunakan sistem berbasis CNN sebagai *backbone* dan Transformer sebagai dekoder, dimana hasil kuantitatifnya lebih baik dibandingkan pendekatan pembuatan *caption* sebelumnya dengan menggunakan *dataset* berbahasa Indonesia. Analisis kepekaan pada berbagai arsitektur pra-terlatih CNN dan penerapan *finetuning* pada *encoder* model dapat meningkatkan hasil evaluasi model dekoder berbasis Transformer untuk setiap arsitektur pembuat enkode pra-terlatih yang berbeda dengan BLEU 1-4, METEOR, ROUGE_L, CIDer dari 58.10, 42.91, 30.40, 21.13, 20.12, 45.32, 60.80, masing-masing. Untuk penelitian mendatang, menjelajahi kumpulan data yang lebih besar, seperti Flickr30k, MS COCO 14, MS COCO 17, dan kumpulan data lain yang terkait dengan keterangan gambar yang lebih besar dari Flickr8k karena sumber daya komputasi kami terbatas pada *platform* pasti akan meningkatkan kinerja model. Semoga, karena temuan ini hanya berfokus pada bagian *encoder* model, akan sangat menarik untuk menguji dampak dari penggunaan penyematan kata (*word embedding*) yang telah dilatih sebelumnya untuk bagian *decoder*, terutama dalam bahasa Indonesia, serta arsitektur berbasis Transformers yang lebih kompleks..

LAMPIRAN

Tangkapan layar e-mail *acceptance* Publikasi I (Pertama)

[8th ICOIACT 2023] Your paper #1570[illegible] ("Automatic Indonesian Image Captioning using CNN and Transformer-Based Model Approach")

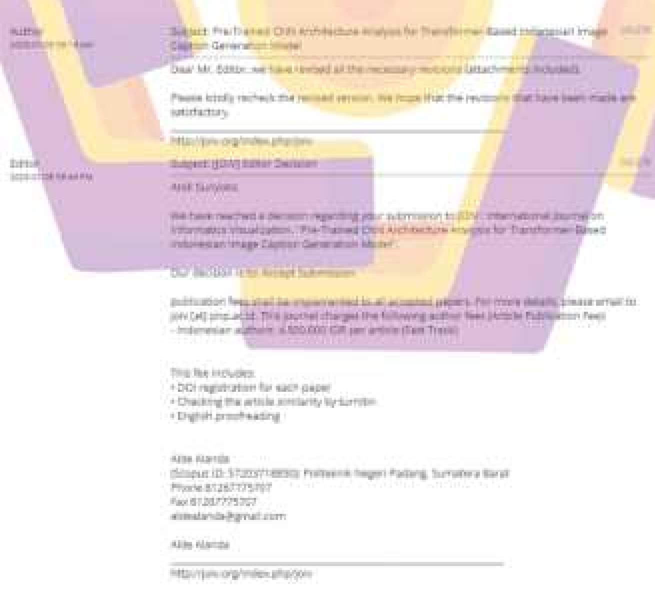
[illegible]

Review 1

[illegible]

Review 2

Tangkapan layar e-mail *acceptance* Publikasi II (Kedua)



Author
[illegible]

Subject: Pre-Trained CNN Architecture Analysis for Transformer-Based Indonesian Image Caption Generation Model

Delete

Dear Mr. Editor, we have revised all the necessary revisions (attachments included).

Please kindly recheck the revised version. We hope that the revisions that have been made are satisfactory.

http://joiv.org/index.php/joiv

Editor
[illegible]

Subject: [JOIV] Editor Decision

Delete

[illegible]

We have reached a decision regarding your submission to JOIV : International Journal on Informatics Visualization, "Pre-Trained CNN Architecture Analysis for Transformer-Based Indonesian Image Caption Generation Model".

Our decision is to: Accept Submission

publication fees shall be implemented to all accepted papers. For more details, please email to joiv [at] pnp.ac.id. This journal charges the following author fees (Article Publication Fees):
- Indonesian authors: 6.500.000 IDR per article (Fast Track)

This fee includes:
• DOI registration for each paper
• Checking the article similarity by turnitin
• English proofreading

Alde Alanda
(Scopus ID: 57203718850) Politeknik Negeri Padang, Sumatera Barat
Phone 6126777570?
Fax 61267775707
aldealanda@gmail.com

Alde Alanda

http://joiv.org/index.php/joiv

AUTOMATIC INDONESIAN IMAGE CAPTIONING USING CNN AND TRANSFORMER-BASED MODEL APPROACH

EKSEMPLAR PUBLIKASI I

Disusun oleh:

Rifqi Mulyawan 21.55.1029

PROGRAM STUDI MAGISTER TEKNIK INFORMATIKA
UNIVERSITAS AMIKOM YOGYAKARTA
2022

Automatic Indonesian Image Captioning using CNN and Transformer-Based Model Approach

Rifqi Mulyawan
Post Graduate Program
Universitas Amikom Yogyakarta, Indonesia
rifqi.mulyawan@students.amikom.ac.id

Andi Sunyoto*
Post Graduate Program
Universitas Amikom Yogyakarta, Indonesia
andi@amikom.ac.id

Alva Hendi Muhammad
Post Graduate Program
Universitas Amikom Yogyakarta, Indonesia
alva@amikom.ac.id

Abstract — **Image captioning involves a phrase text generation or more for the visual content descriptions from images. Caption generation for an image is considered important to aid human activities in comprehending visual material, such as captions on medical images, human contact with robots, and helping visually impaired people explain visuals. Our study aims to create an Indonesian image description and assess how successful the caption's approach is. Translated Flickr8k datasets are provided for this investigation. This study employs two methods for captions generation from a photo: CNN with ResNet as the encoder, and Transformer, a self-attention-based mechanism as the decoder. Using distinct Indonesian datasets, named Flickr8k Bahasa, we used a Transformer-based approach to create an Indonesian captions generation model from photos. We demonstrated that our Indonesian Transformer-based strategy outperformed the old one, where the best results were obtained with BLEU-1 to 4, METEOR, ROUGE_L, CIDEr of 56.00, 41.17, 29.42, 20.57, 19.50, 44.16, 57.26, respectively. Besides comparing the model's performance using different ResNet-based CNN pre-trained models, a larger CNN model did not guarantee any accuracy improvement through fifty epochs of training processes.**

Keywords* — *Deep Neural Network, Convolutional Neural Network, CNN, Indonesian Image Captioning, Transformer, Attention Mechanism

I. INTRODUCTION

Computer vision has made significant progress in image processing, such as image categorization and object recognition. Image captioning uses advances in image classification and object recognition to create one or more text captions that visually and automatically describe the content of an image [1].

The automatic generation of full and natural descriptions of photos benefits titles attached to news photographs, descriptions related to the medic, image capture based on language, information accessed by visually impaired users, as well as the interaction of human and robot development [2]. Written descriptions of visual content, such as photographs, make the information more digestible.

A more sophisticated understanding of image categorization and detection is required to generate a proper natural language text for a photograph. This issue is intriguing because it merges two of Artificial Intelligence's key fields [3]: (1) Computer Vision or CV. (2) Natural Language Processing (NLP), alternatively referred to as Natural Language Understanding (NLU). The conventional way of solving this problem is to use a CNN or Convolutional Neural Network (ConvNet) and an LSTM decoder architecture, which stands for Long Short Term Memory. The input image is encoded using neural networks, and the data captions are created using an iterative neural network called LSTM [4]. However, this strategy has the disadvantage of requiring that the sequence be handled sequentially. Numerous researchers recently used the Transformer model for LSTM replacement to automatically generate captions for photos using a range of language-based datasets to overcome this constraint.

CNN and attention-based mechanism discoveries [5] should be used in this work to offer images with descriptions in Indonesian. Along with providing descriptions in Indonesian, the technique should strive for as natural syntax as possible, close to that of human language [1]. Additionally, our experiment compares our suggested Indonesian caption-generating model to a more substantial layer of the CNN model that has been pre-trained.

II. LITERATURE REVIEW

A. Automatic Image Caption

A recent image captioning technique uses the widely utilized encoder-decoder framework types. The model is constructed using neural-machine translation [6], such as the end-to-end framework utilized in conjunction with the ConvNet for picture encoding and the RNN-LSTM for sentence generation.

ConvNet [7] is utilized to extract features in this work. This type of feed-forward neural network investigates the hierarchical structure of an image by analyzing the representation of internal features and generalizing those features in an image in general, such as object recognition and other computer vision tasks [8], where in practice, the implementation of the architecture can be applied in other fields that specifically include such as the problem of classifying rice leaf diseases for agriculture [9] and detection of CT-Scan lung disease images for the health sector [10]. Fig.1 illustrates the general CNN architecture to recognize the computer vision image classification.

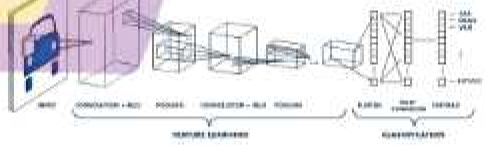


Fig. 1. CNN Architecture Image [7]

B. Transformer-Based Architecture Model

Research "All You Need is Attention" [11] debuted a new Transformer architecture. According to the title of the research, the design makes use of the previously reported attention mechanism, a self-only attention-based architecture that converts one sequence to another using two encoder and decoder components.

*Corresponding author: Andi Sunyoto (andi@amikom.ac.id)

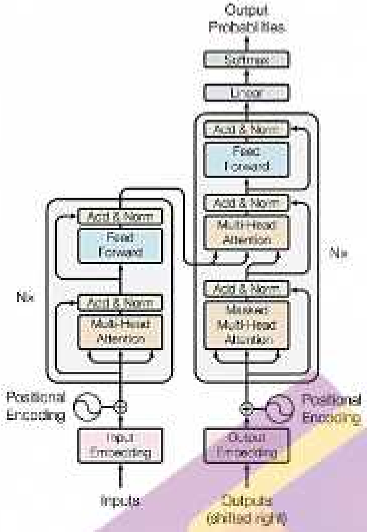


Fig. 2. Original Transformer Model Architecture Image [11]

The attention-based mechanism is the first transduction model that computes input and output representations purely through self-attention tools rather than RNN. The Transformer-model-based image captioning is stacked with self-attention and point-wise layers that completely connect and utilize the dot-product attention mechanism to implicitly link the informative region. Fig. 2 on the left and right are the encoder-decoder part of the original Transformer, where we replaced the encoder portion of the architecture with a ResNet-based CNN.

C. Related Work

Many studies were previously done in English since the relevant datasets were all English. The attention mechanism was altered to create captions by Jyoti Aneja [12].

Most articles employed VGG-16 for the visual aspect of picture captioning in [2], [13], [14], although other's research also implemented the pre-trained AlexNet [2], [14], or ResNet [2] for the feature image. Even so, BiLSTM [2] was used by some researchers. For other languages, researchers created subtitles in Chinese [15], [16], Japanese Yoshikawa [17], Arabic [18], Bahasa Indonesia's adaptive attention [19], Gated Recurrent Unit (GRU) [20], and using the FEEH-ID Flickr8k's dataset [21] in addition to English.

Some studies also use Transformer architecture as the basic model, although each study uses its method, where research obtained the highest BLEU score with 81.7, 66.8, 52.4, and 40.4 (BLEU 1 to 4). These methods include [22] using Image Transformers, [23] using Meshed-Memory Transformers, [24] using Boosted Transformers, and [25] using Multimodal Transformers with multi-view visual representation approaches.

A Transformer-based model was also utilized to caption images using an English dataset. Li and his colleagues [26] looked at a Transformer-architect-based sequence modelling framework for picture text generation that was solely made up of attention and feed-forward layers. Furthermore, in their study [27], object spatial connection modelling was used for image captioning, notably inside the architecture, using the Transformer as a model encoder to incorporate the module's object relation types. In addition, Atliha and Eok [28] proposed and employed image caption augmentation in a dataset to find a solution to the picture captioning challenge, incorporating BERT augmentation.

Zhang also uses a ConvNet as the encoder for the pre-trained CNN model for extracting picture features. The Transformer is used to construct captions using the encoder's output vector, which contains critical information from the image. In contrast, in their study [22], the image Transformer for picture captioning was presented, with each layer implementing numerous sub-Transformers for spatial connection encoding.

III. Proposed Method

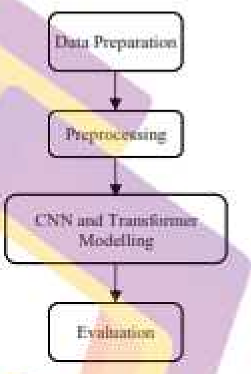


Fig. 3. Transformer-Based Model Image Caption Generation Flowchart

Data Preparation, Data Preprocessing (pictures and phrases), Modelling (image feature extraction and caption generation), and evaluation are the four primary procedures in this study technique. Fig. 3 depicts the flowchart of our study for caption generation in Bahasa Indonesia utilizing a Transformer.

A. Data Preparation

The dataset was retrieved using common Flickr8k and then translated into a total of 40460 captions using Google Translate into Indonesian, cross-checked, also manually confirmed before being used in this study. Named Flickr8k Bahasa, this Indonesian Flickr8k dataset contains 8091 photographs and a caption file with five types of words, 6000 train, 1000 testing, and validation.

B. Preprocessing

Preprocessing consists of two major steps: image resizing and text preprocessing. The preprocessing input approach was employed to normalize the image to adapt the picture format utilized in the ResNet pre-trained model for the image extraction process.

We also change all phrases to lowercase to make counting the number of unique words in the dataset easier. The "<start>" (called SOS) token is added to the first sentence, and the "<end>" token is added to the last section of the sentence

(EOS), so the model can generate photo sentences when it encounters the beginning token and stops when it encounters the end token.

Additionally, we divided the text into individual words for sentences to ensure that our Bahasa Indonesia dataset had a unique vocabulary. By turning these words into sequential word indexes, a vector can be used to represent the order of words in a sentence. A "Pad" in the framework returns zero if the vector's length is less than the supplied value and deletes the end of the vector if the length exceeds the specified value before then, shifting the word order vector by one step to study the model for the following phrase.

C. Model Definition

Three critical components contribute to the definition of our Transformer-based model: ResNet50, Positional Encoding, and Attention in Multiple Heads. Residual Networks are also known as ResNets. Like the encoder, we employed a CNN architecture from the ResNet family to generate one-dimensional vector representations of the input images, including Label Smoothing and Factorized 7x7 convolution. The use of additional classifiers to transmit the label of information down the network [27] to execute feature extraction on the photo in our Flickr8k's Bahasa dataset. Since it is only required to extract the picture vector, the softmax layer in this model must be deleted. Prior to providing the model, all pictures must be preprocessed by equalizing the image size to 224px x 224px.

The following element is positional encoding using sin and cos functions with varying frequencies. If the index position in the input vector is odd, the first function is constructed; if the index position in the input vector is even, the second function is created. By including such vectors in the proper input embedding, network information about each vector's position is obtained.

The weight of attention is calculated using multi-head attention. The principal dimensions of Q (query), K (key), and V (value) must be identical. K and V must have penultimate dimensions that are identical. The covers vary based on the kind (Padding or Look Ahead). Four parts exist inside this multi-head attention. (1) Linear layer, (2) Scaled dot-product attention, (3) Concatenation, (4) End linear layer.

To maximize our decoder, we also include a beam search method to select the words with the highest score at each decoding step, as it identifies the most optimum sequence.

D. Evaluation

The evaluation includes training and testing/validation implemented to run simultaneously. The standard model results evaluation is done with the Bilingual Evaluation Understudy, BLEU [29]. This metric evaluates the resulting sentence to the target sentence. The score means how close the generated caption text with the expected text or caption. The resulting sentence similar to the target sentence will be given a score of 1.0; if it is not similar, it will be given a score of 0.0. This evaluation method calculates precision using the Ngrams metric, where the maximum length of n-grams is four because it has the highest correlation with people's judgment. The score is measured using the below formula:

$$BLEU\ = BP.exp\left(\textstyle\sum_{n-1}^{N} w_n \, log \, p_n\right) \qquad (1)$$

In Equation 1, where the p_n is the modified precision for n-gram, the base log is the natural base e, w_n for the weight from 0 to 1 of $log \, p_n$, $\sum_{n-1}^{N} w_n$ equals to 1. The brevity penalty (BP) for machine translation penalization can be calculated using the below formula:

$$BP = \begin{cases} 1 & if \; c > r \\ exp\left(1 - \frac{r}{c}\right) & if \; c \leq r \end{cases} \qquad (2)$$

In Equation 2, the c is for the unigrams (length) numbers in every candidate sentence and r for the best match of each candidate sentence's length in the dataset.

In addition to the sparse cross entropy function, which is generally used in caption generation problems as a loss function, these caption findings acquired in this study were also obtained with other popular metrics [29] like METEOR, ROUGE_L, CIDEr that unconventionally used by any researchers used before for evaluating Transformer-based model to create Bahasa Indonesia automatic captions generation.

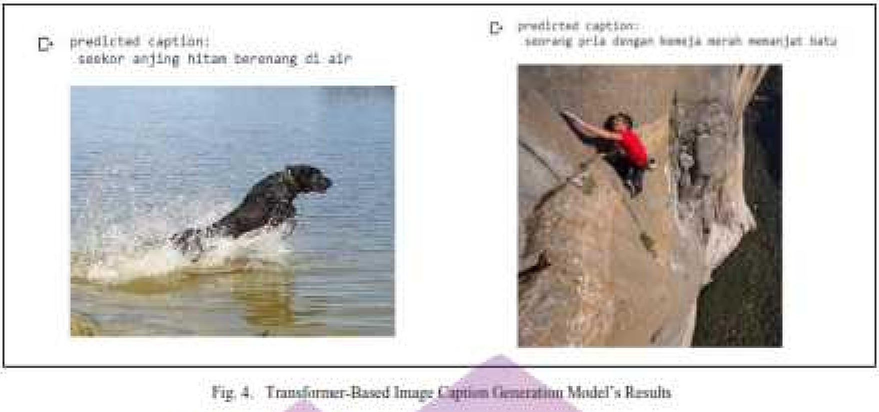
IV. Results and Discussion

To generate captions in Bahasa Indonesia, we used Adam's optimization to train a Transformer-based model with a 50-epoch with cross-entropy loss on our Indonesian Flickr8k's Bahasa after the transfer learning process in which they use pieces of a pre-trained model to our new model. This method is nearly always preferable to building a new model from the start (i.e., knowing nothing) as we utilize a pre-trained Resnet family's encoder for our picture captioning.

The "test" results were analyzed using our evaluation method, where each sentence generated by the model in the given image was compared to the five target sentences used as sentence references to calculate the BLEU score plus the METEOR, ROUGE L, and CIDEr's metrics of 56.00, 41.17, 29.42, 20.57, 19.50, 44.16, 57.26, respectively.

The qualitative results can be seen in Fig. 4, which is the result of the direct prediction of the Transformer-based model's inference. The image does not exist in the previous training dataset, where certain images and sentences from our Transformer-based model generate grammatically correct captions.

Table I shows the quantitative score for the Indonesian Flickr8k Bahasa dataset using Transformer in comparison to Nugraha's GRU-based model [20], Mulyanto's Flickr8k FEEH-ID dataset's LSTM-based model approach [21], and Mahadi's LSTM-adaptive-attention model [19]



```
predicted caption:
 seekor anjing hitam berenang di air
```

```
predicted caption:
 seorang pria dengan kemeja merah memanjat batu
```

Fig. 4. Transformer-Based Image Caption Generation Model's Results

TABLE I. INDONESIAN IMAGE CAPTIONING METHOD COMPARISON RESULTS

Method	**Dataset**	**BLEU**				**METEOR**	**ROUGE_L**	**CIDER**
		1	**2**	**3**	**4**			
CNN + Adaptive Attention LSTM [19]	MSCOCO + FLICKR30K	67.80	51.20	37.50	27.40	-	-	99.00
CNN + LSTM [21]	FLICKR8K FEEH-ID	50.00	31.40	23.90	13.10	-	-	-
CNN + GRU [20]	FLICKR30K	36.70	17.80	06.70	02.00	-	-	-
Proposed (TRANSFORMER)	FLICKR8K BAHASA	56.00	41.17	29.42	20.57	19.50	44.16	57.26

As we can see in Table I, our Transformer-based model approach clearly beats all other approaches for the Indonesian photo captioning task, with the exception of adaptive-attention-based considering the Flickr8k dataset, the language used is very small compared to what they use, where they combine two datasets that are quite large in terms of image captioning and the implementation of attention-based mechanisms in their methods with LSTM-based models.

As a result of this finding and our available computational resources, we may compare our Transformer-based model approach for the Indonesian language to the English caption generation model for small-scale Indonesian datasets such as Flickr8k.

Experiments were also carried out to implement other pre-trained architectures using our Indonesian Flickr8k Bahasa dataset because the scores evaluated using this Transformer-based model were unsatisfactory. In contrast, we suspect the larger pre-trained CNN model can improve each mentioned metrics score and vice-versa.

The results are obtained through some of the experiments using other ResNet's pre-trained CNN models for the encoder with standard one Transformer's multi-heads attention. The channel size when applying different ResNet-based CNN pre-training models is adjusted to 512 for ResNet18, 2048 for ResNet50, and ResNet101 on the model flow-through 50 epochs with the criteria for stopping training if there is no improvement in BLEU-4 in the last ten epochs, as noted in Table II.

TABLE II. PRE-TRAINED MODEL EXPERIMENTAL RESULTS

RESNET	**BLEU**				**METEOR**	**ROUGE_L**	**CIDER**
	1	**2**	**3**	**4**			
RESNET18	53.34	37.93	26.27	17.97	18.71	41.82	48.69
RESNET50	**56.00**	**41.17**	**29.42**	**20.57**	19.50	**44.16**	**57.26**
RESNET101	54.04	39.06	27.39	18.95	**20.12**	44.05	56.64

As noted in Table II, ResNet50 proved to be the best model for our dataset as it gives the highest score compared to the ResNet18 and ResNet101 pre-trained CNN models.

Using these ResNet50 pre-trained CNN models, we also made changes to the decoder heads number, as we expect to increase the number of Transformer-based multi-head attention while training our dataset. The results should increase the score as the Transformer's architecture can attend to information from different subspaces of representation.

TABLE III. HEADS NUMBER VARIATION EXPERIMENTAL RESULTS

HEADS	**BLEU**				**METEOR**	**ROUGE_L**	**CIDER**
	1	**2**	**3**	**4**			
1	**56.00**	**41.17**	**29.42**	**20.57**	**19.50**	**44.16**	**57.26**
2	53.49	38.43	27.16	18.99	19.18	43.19	56.32
3	55.43	40.13	28.50	19.70	19.21	43.22	52.30

As we can see in Table III, there is no model score improvement to it as we suspect that increasing the number of heads also causes our Transformer-based model to overfit for the small dataset like Flickr8k compared to bigger ones like Flickr30k, MSCOCO 14, or 17.

V. Conclusion

In this paper, our research attempts to find the most effective approach and sensitivity for the Indonesian image captioning problem using Indonesian Flickr8k's Bahasa datasets.

Several things reported in this paper are summarized by (1) Research and experiment with the Bahasa Indonesia image captioning model, (2) using encoder-decoder as an image caption generation model in Indonesian, (3) using Transformer architecture with pre-trained Resnet's family as the decoder.

Transformers-based model implementation can outperform previous designs using small to medium-scale datasets because they entirely avoid recursion by analyzing whole sentences and learning associations between words using multi-head attention techniques and positional embeddings. It's also worth noting that Pytorch's Transformers can only capture dependencies within the limited input size used to train them.

With some experimental work, we found that using different pre-trained CNN models like ResNet18 and ResNet50 as decoders has improved results scores while using the larger ResNet101 has not, as we believe the model-overfitting problem is the problem that causes these results.

As for future work, fine-tuning and applying this Transformer-based approach to a different dataset like Flickr30k, MS COCO 14, 17, or other caption-generation-related datasets, also there are many new powerful Transformers-based methods, such as the XL, Entangled, and Meshed Memory Transformer-based model, that can be used to get even better outcomes in Bahasa Indonesia's automatic caption generation applications.

References

[1] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, "A short review on image caption generation with deep learning," *Proc. Int. Conf. Image Process. Comput. Vision, Pattern Recognit.*, pp. 10–18, 2019.

[2] H. Shi, P. Li, B. Wang, and Z. Wang, "Image captioning based on deep reinforcement learning," *ACM Int. Conf. Proceeding Ser.*, vol. 01052, pp. 1–7, 2018, doi: 10.1145/3240876.3240900.

[3] G. Zhang *et al.*, "A reversible fluorescent pH-sensing system based on the one-pot synthesis of natural silk fibroin-capped copper nanoclusters," *J. Mater. Chem. C*, vol. 4, no. 16, pp. 3540–3545, 2016, doi: 10.1039/c6tc00314a.

[4] S. Pa Pa Aung, W. Pa Pa, and T. L. Nwe, "Automatic Myanmar Image Captioning using CNN and LSTM-Based Language Model," *Proc. 1st Jt. Work. Spok. Lang. Technol. Under-resourced Lang. Collab. Comput. Under-Resourced Lang.*, no. May, pp. 139–143, 2020, [Online]. Available: https://www.aclweb.org/anthology/2020.sltu-1.19.

[5] M.-H. Guo *et al.*, "Attention Mechanisms in Computer Vision: A Survey," vol. 14, no. 8, pp. 1–27, 2021, [Online]. Available: http://arxiv.org/abs/2111.07624.

[6] F. Stahlberg, "Neural machine translation: A review," *J. Artif. Intell. Res.*, vol. 69, pp. 343–418, 2020, doi: 10.1613/JAIR.1.12007.

[7] A. Ajit, K. Acharya, and A. Samanta, "A Review of Convolutional Neural Networks," *Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020*, pp. 1–5, 2020, doi: 10.1109/ic-ETITE47903.2020.049.

[8] A. F. Gad, *Practical Computer Vision Applications Using Deep Learning with CNNs*. 2018.

[9] A. Julianto and A. Sunyoto, "A performance evaluation of convolutional neural network architecture for classification of rice leaf disease," *IAES Int. J. Artif. Intell.*, vol. 10, no. 4, pp. 1069–1078, 2021, doi: 10.11591/IJAI.V10.I4.PP1069-1078.

[10] U. Vwxgbqwv and D. Df, "'hwhfwlrq 2i &7 ± 6fdq /xqjv &29,' gdwd 8vlqj &rqyroxwlrqdo 1hxudo 1hwzrun $qg &/$+(," vol. 0, pp. 302–307, 2021.

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

[12] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 5561–5570, 2018, doi: 10.1109/CVPR.2018.00583.

[13] R. Subash, R. Jebakumar, Y. Kamdar, and N. Bhatt, "Automatic image captioning using convolution neural networks and LSTM," *J. Phys. Conf. Ser.*, vol. 1362, no. 1, 2019, doi: 10.1088/1742-6596/1362/1/012096.

[14] C. Wang, H. Yang, and C. Meinel, "Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 14, no. 2s, 2018, doi: 10.1145/3115432.

[15] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," *MM 2017 - Proc. 2017 ACM Multimed. Conf.*, pp. 1549–1557, 2017, doi: 10.1145/3123266.3123366.

[16] X. Li, W. Lan, J. Dong, and H. Liu, "Adding Chinese captions to images," *ICMR 2016 - Proc. 2016 ACM Int. Conf. Multimed. Retr.*, pp. 271–275, 2016, doi: 10.1145/2911996.2912049.

[17] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR captions: Constructing a large-scale Japanese image caption dataset," *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, pp. 417–421, 2017, doi: 10.18653/v1/P17-2066.

[18] H. A. Al-muzaini, T. N. Al-yahya, and H. Benhidour, "Automatic Arabic image captioning using RNN-LSTM-based language model and CNN," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, pp. 67–73, 2018, doi: 10.14569/IJACSA.2018.090610.

[19] M. R. S. Mahadi, A. Arifianto, and K. N. Ramadhani, "Adaptive Attention Generation for Indonesian Image Captioning," *2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020*, 2020, doi: 10.1109/ICoICT49345.2020.9166244.

[20] A. A. Nugraha, A. Arifianto, and Suyanto, "Generating image description on Indonesian language using convolutional neural network and gated recurrent unit," *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, pp. 1–6, 2019, doi: 10.1109/ICoICT.2019.8835370.

[21] E. Mulyanto, E. I. Setiawan, E. M. Yuniarno, and M. H. Purnomo, "Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset," *2019 IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. CIVEMSA 2019 - Proc.*, 2019, doi: 10.1109/CIVEMSA45640.2019.9071632.

[22] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image Captioning Through Image Transformer," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12625 LNCS, pp. 153–169, 2021, doi: 10.1007/978-3-030-69538-5_10.

[23] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 10575–10584, 2020, doi: 10.1109/CVPR42600.2020.01059.

[24] J. Li, P. Yao, L. Guo, and W. Zhang, "Boosted transformer for image captioning," *Appl. Sci.*, vol. 9, no. 16, pp. 1–15, 2019, doi: 10.3390/app9163260.

[25] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal Transformer with Multi-View Visual Representation for Image Captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4467–4480, 2020, doi: 10.1109/TCSVT.2019.2947482.

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

[27] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 1–11, 2019.

[28] V. Atliha and D. Šešok, "Text augmentation using BERT for image captioning," *Appl. Sci.*, vol. 10, no. 17, 2020, doi: 10.3390/app10175978.

[29] Y. Cui, G. Yang, A. Veit, X. Huang, S. Belongie, and C. Tech, "Learning to Evaluate Image Captioning - Supplementary Material," pp. 3–5, 2015.

PRE-TRAINED CNN ARCHITECTURE ANALYSIS FOR TRANSFORMER-BASED INDONESIAN IMAGE CAPTION GENERATION MODEL

EKSEMPLAR PUBLIKASI II

Disusun oleh:

Rifqi Mulyawan 21.55.1029

PROGRAM STUDI MAGISTER TEKNIK INFORMATIKA
UNIVERSITAS AMIKOM YOGYAKARTA
2023

Pre-Trained CNN Architecture Analysis for Transformer-Based Indonesian Image Caption Generation Model

Rifqi Mulyawan[a], Andi Sunyoto[a,*], Alva Hendi Muhammad[a]

[a] *Post Graduate Program, Universitas Amikom Yogyakarta, Daerah Istimewa Yogyakarta 55283, Indonesia*
*Corresponding author: *andi@amikom.ac.id*

Abstract— **Classification and object recognition in image processing has seen significant improvement in computer vision tasks. The method is often used for these visual-related problems, especially in picture classification utilizing the Convolutional Neural Network (CNN). In the popular state-of-the-art (SOTA) task of generating a caption on an image, the implementation is often used for feature extraction of an image as an encoder. Instead of performing direct classification, these extracted features are sent from the encoder to the decoder section to generate the sequence. So, some CNN layers related to the classification task are not required. This study aims to determine which CNN pre-trained architecture or model can perform the best in extracting image features using a state-of-the-art Transformer model as its decoder. Unlike the original Transformer's architecture, we implemented a vector-to-sequence way instead of sequence-to-sequence for the model. Indonesian Flickr8k and Flick30k datasets were used in this research. Evaluations were carried out using several pre-trained architectures, including ResNet18, ResNet34, ResNet50, ResNet101, VGG16, Efficientnet_b0, Efficientnet_b1, and Googlenet. The qualitative model inference results and quantitative evaluation scores were analyzed in this study. The test results show that the ResNet50 architecture can produce stable sequence generation with the highest accuracy value. With some experimentation, finetuning the encoder can significantly increase the model evaluation score. As for future work, further exploration with larger datasets like Flickr30k, MS COCO 14, MS COCO 17, and other image captioning datasets in Indonesian also implementing a new Transformers-based method can be used to get a better Indonesian automatic image captioning model.**

Keywords— **Artificial intelligence; deep learning; convolutional neural network; indonesian image caption generation; transformer.**

Manuscript received 23 Nov. 2022; accepted 08 Jan. 2023.

I. Introduction

Many currently available captioning algorithms to convey in words an essence of an image are based on the architecture of an encoder-decoder, in which a decoder infrastructure may anticipate words by making use of a function received from an encoder network through an attention approach. Studies on image subtitling have mainly concentrated on a translation approach consisting of a visual encoder and a language decoder [1].

Creating image captions may be utilized for various purposes, including automating the driving of autos, developing face recognition systems, characterizing individuals with visual impairments, enhancing the quality of photo queries, and many more. The difficult task of developing the natural language descriptions of the information in a picture resides within the computer vision (CV) interface for image feature extraction and generating the sequence using the natural language processing (NLP) technology.

The task of photo caption generation has already had a significant impact in several fields, such as image search also various disciplines, such as software development for people with disabilities, video surveillance and security, and the interface between humans and computers [2].

As a popular challenge involving sequence modeling, the state-of-the-art (SOTA) problem of photo caption generation uses various approaches. For example, the Convolutional Neural Network, ConvNet, known as the CNN, is applied with other language architecture, like the Recurrent Neural Network (RNN), as a CNN-RNN-based framework approach [3]. This work uses the standard encoder-decoder architecture using a pre-trained CNN model to build feature vectors. They are then fed into an RNN as the decoder responsible for generating the language description.

The standard encoder-decoder model was also utilized in [4] and [5] to make subtitles out of photographs. However, the recurrent structure of the enhanced RNN type, like the Long Short Term Memory (LSTM), makes it harder to train because of its sequential nature, resulting in a lower evaluation score on the standard RNN-based model. However, in [6], the parallelism problem was finally overcome by the SOTA model, the Transformer. Due to the fact that the architecture is built on a context-aware attention mechanism, it can operate in parallel throughout the training phase and does not require a certain order.

For image captioning in Indonesian, the GRU approach in [7] for generating Indonesian captions on an image is used to overcome some problems that exist in the RNN. However, as the model is still RNN-based, their finding shows that it lacks context understanding and stated that the need for SOTA research implementation for sequence generation in Indonesian is a must. Earlier research for Indonesian caption generation in [8] also uses CNN with the pre-trained architecture of VGG-16 for the model's encoder with another RNN type, the LSTM, as the decoder, but without investigating feature extraction impact on measuring image text quality for the model's performance. Their finding shows that the model's result has a better evaluation score with BLEU 1-4 (50.00, 31.40, 23.90, 13.10, respectively). Previous research in [7] and [8] for generating image captions in Indonesian leaves a space for exploring the effect of using another pre-trained CNN layer with the SOTA approach using Transformer-based that are context-aware to get better model evaluation results.

Our contributions to this research are: (1) Created an Indonesian image captioning dataset based on the rules of the standard benchmark of Flickr8k and Flickr30k to train the model. (2) Proposed a Transformer-based model using CNN as the encoder to generate photo captions in Indonesian. (3) Employing a context-aware using an attention-mechanism-based decoder. (4) Compared 8 different pre-trained CNN as photo feature extraction to the Transformer-based model. (5) Compared the model performance to the previous approach in Indonesian image captioning.

In this study, we used Indonesian Flickr8k [9] and translated Flickr30k [10] to test our model's performance in the Indonesian language to produce an image captioning model in Indonesian, proposing SOTA Transformer-based architecture. Fig. 2 depicts the proposed Transformer-based model's approach to caption images in Indonesian.

This research aims to explore which CNN architecture is the most effective at generating high-accuracy results by comparing and contrasting their respective performances on eight different pre-trained CNNs. This study also investigates the effect of varying CNN channel size (depth) on the Transformer-based model performance for image feature extraction.

A. Image Caption Generation in Another Language

Since most datasets are written in English, most of the study for caption generation was done in that language, whereas in [11], the attention-based mechanism is adapted for caption generation.

Most studies implement the VGG-16 for the encoder part of the captioning model, like the ConvNet in [12]. However, several researchers also employed the pre-trained AlexNet in [13], [4], or Residual Network (ResNet) for the visual feature and BiLSTM in [13] is also used.

For other languages, other datasets like Chinese [14], [15], Japanese Yoshikawa [16], Arabic [17], Bahasa Indonesia in [18] (custom dataset that combines MS COCO and Flickr30k), Indonesian Flickr30k [7], and the FEEH-ID Flickr8k's dataset [8] also created besides English.

B. Image Caption Generation using Attention-Mechanism

A significant number of researchers in the past have made use of visual attention to English datasets. Encoder-decoder research has used two primary kinds of attention, namely for the purpose of captioning images or videos. The first sort of attention is called semantic attention, which refers to attention to words. The second one of attention is known as spatial attention, which relates to the focus placed on images. Research by Xu et al. [19] on photo captioning saw the introduction of a model for visual attention for the first time. They either applied "hard" pooling, which finds the region that is more likely to be attended, or "soft" pooling, which takes the average of the spatial qualities and assigns attentive weights to each of those variables.

Moreover, in [20], CNN's Channel-wise Attention and Spatial Attention were both put to use when watching the network. Research by Chen et al. [21] also used visual attention when creating captions for the pictures. Also, in [22], a semantic attention model was used in RNNs to link the visual feature with the visual ideas to create the picture description.

C. Image Captioning using Transformer-Based Approach

Image captioning with Transformer as the model's decoder using an English dataset was used in previous research. Li et al. [23] studied a Transformer-based framework for sequence modeling in picture captioning. When it was initially developed, it included simply the attention and feed-forward layers.

In addition, the study presented in [24] makes use of spatial object relationship modeling for picture caption generation. It is explicitly done inside the encoder-decoder architecture using the SOTA Transformer. It is done by implementing the object relation module to the encoder as the first step in developing image captions. Research in [25] suggested that augmenting the photo captions in a dataset with additional information, such as employing BERT, might be an effective method for enhancing an existing solution to the problem of image captioning.

Research in [26] used two different streams of architecture based on Transformers-one for the graphical component and another for the linguistic component. Paper in [26] additionally utilized a CNN model for the encoding component, while a Transformer model was used for the decoding section of the model. Both the encoder and decoder models were utilized. The architecture was constructed using a Transformer, which consists of a model for both an encoder and a decoder. In addition, it employs a system for stacking its attention on top of itself. When CNN is employed as an encoder, as explored in [27], image features may be obtained, and the encoder's output is a context vector containing the most significant picture information. After that, this vector is sent into Transformer, which creates the captions for the pictures based on those captions.

To put it into perspective, research in [28] presented the image Transformer as a tool for image captioning. Each layer of the Transformer implements several sub-Transformers that enable the encoding of spatial relationships between picture portions and decoding of the different forms of information contained within the image regions.

II. The Materials and Method

A. Dataset

The dataset used for this analysis is the standard English Flickr8K [28]. We translated it to Indonesian using Google Translate and manually cross-checked the annotation. Named Flickr8k Bahasa [9], like the original Flickr8k, our dataset features 8,091 photos. There are 6,000 training photos, 1,000 validation, and 1,000 for testing.

In addition, five human-created reference captions are linked to each image, meaning that for every image in our training set, there are 40,460 corresponding caption samples.

We also prepared Indonesian Flickr30k's Bahasa, which consists of 158,915 captions to test our final model performance. This translated dataset contains 31,783 photos, including a caption file comprising five types of sentences, 29,000 used for training, 1,000 used for testing also validations.

B. System Design

Fig. 1, which can be seen further down this page, is a flow or process that describes in detail the experiments carried out to determine how the different ConvNet or CNN's pre-trained model methods perform in generating and evaluating image caption problems in Indonesian. It provides an easy-to-follow visual representation of the entire procedure.

The first step is to preprocess the caption text and the input image. The caption text from the dataset is tokenized to ensure we have a unique vocabulary. At this stage, each image in our Indonesian dataset changed to less than the original size. Then the dataset was prepared for training, validation, and testing, resulting in the input data for the training process using the CNN method with transfer learning techniques.

Eight different CNN pre-trained architectures are used at this stage, namely ResNet18, ResNet34, ResNet50, ResNet101, VGG16, Efficientnet_b0, Efficientnet_b1, and Googlenet. Another system output is the prediction or the inferences of the CNN-Transformer model.

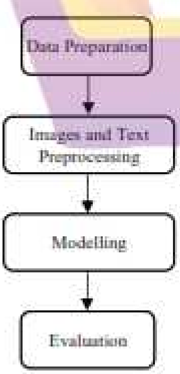


Fig. 1. Indonesian Image Captioning Model Analysis Flowchart

C. CNN-Transformer

The ResNet CNN model was utilized as our choice for the encoding algorithm baseline. Vectors of fixed-length feature representation that CNN extracts are called encoder's hidden states, which are then used as the basis for the attention mechanism alongside the annotation vectors.

Various networks, including ResNet18, ResNet34, ResNet50, ResNet101, VGG16, Efficientnet b0, Efficientnet b1, and Googlenet, were used in our tests. Since we are not interested in classifying the input, the last pooling and softmax layer are unneeded, and retrieved annotation vectors from the last convolutional layer instead. Here, the output is of the size that can be expressed with $x * y$, n, where n is the CNN feature channels that vary with the particular encoder employed and x, y represents the shape of the feature map.

Afterward, n number of decoder layers was applied to the summed-up output. Each decoder layer comprised three further layers: (1) a sub-layer of masked multi-head attention that includes both a padding mask and a look-ahead mask. (2) an attention sub-layer with many heads with a padding mask that accepts the encoder output as inputs (with two inputs). (3) a masked multi-head attention sub-layer that has an output query.

Look-ahead and the padding mask of the Transformer were multi-head attention sub-layers that were disguised. Within this specific architectural design context, the third layer was made up of feed-forward networks. Then, the information that the Transformer decoder had produced was sent to the linear layer so that it could be utilized as input there.

In the end, probabilistic softmax predictions are constructed in a serial way, and the output that has been generated up to this point is employed to determine the subsequent step that has to be done to complete the process. Fig. 2, which can be seen below, is the image for our proposed architecture.

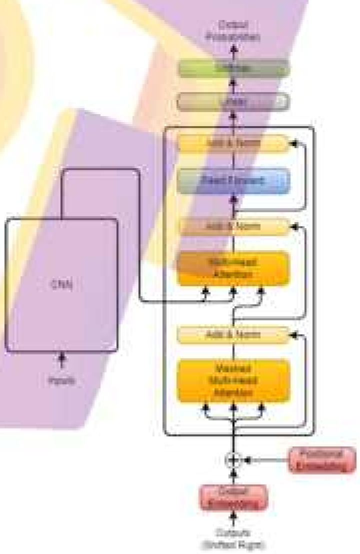


Fig. 2. CNN and Transformer-Based Decoder Model Architecture

Unlike RNN, where we send the words of a sentence one by one into the model, we send the whole sentence to the decoder simultaneously. This parallelization is the main benefit of why the architecture is faster to train compared to the previous one, like RNN/LSTM and GRU.

D. Model Evaluation Metrics

When assessing the quality of automatically generated captions, we make use of BLEU [29], METEOR [30], ROUGE [31], and CIDEr [32]. Utilizing n-grams, BLEU [29] determines the degree of similarity between a collection of reference texts and the text created by a computer. The word-to-word matching algorithm METEOR [30] uses equivalent word stems and synonyms to find straight matches between words. ROUGE [31] measures sentence similarity using word pairings, n-grams, and word sequences, whereas extant research on picture captioning makes considerable use of different metrics like BLEU, METEOR, and ROUGE. In addition, CIDEr [32] is also utilized to quantify the similarity between reference texts and predicted text for every n-gram. On the other hand, it has been discovered that CIDEr has a stronger correlation with human evaluation [33]. As a result, we concluded that including CIDEr would provide a more accurate depiction of the caption quality.

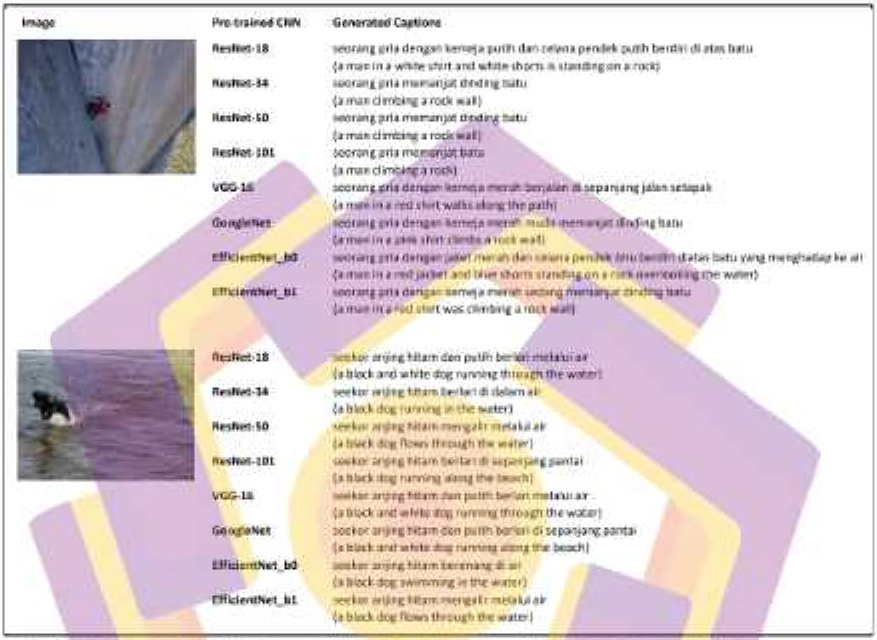


Image	Pre-trained CNN	Generated Captions
	ResNet-18	seorang pria dengan kemeja putih dan celana pendek putih berdiri di atas batu (a man in a white shirt and white shorts is standing on a rock)
	ResNet-34	seorang pria memanjat dinding batu (a man climbing a rock wall)
	ResNet-50	seorang pria memanjat dinding batu (a man climbing a rock wall)
	ResNet-101	seorang pria memanjat batu (a man climbing a rock)
	VGG-16	seorang pria dengan kemeja merah berjalan di sepanjang jalan setapak (a man in a red shirt walks along the path)
	GoogleNet	seorang pria dengan kemeja merah muda memanjat dinding batu (a man in a pink shirt climbs a rock wall)
	EfficientNet_b0	seorang pria dengan jaket merah dan celana pendek biru berdiri diatas batu yang menghadap ke air (a man in a red jacket and blue shorts standing on a rock overlooking the water)
	EfficientNet_b1	seorang pria dengan kemeja merah sedang memanjat dinding batu (a man in a red shirt was climbing a rock wall)
	ResNet-18	seekor anjing hitam dan putih berlari melalui air (a black and white dog running through the water)
	ResNet-34	seekor anjing hitam berlari di dalam air (a black dog running in the water)
	ResNet-50	seekor anjing hitam mengalir melalui air (a black dog flows through the water)
	ResNet-101	seekor anjing hitam berlari di sepanjang pantai (a black dog running along the beach)
	VGG-16	seekor anjing hitam dan putih berlari melalui air (a black and white dog running through the water)
	GoogleNet	seekor anjing hitam dan putih berlari di sepanjang pantai (a black and white dog running along the beach)
	EfficientNet_b0	seekor anjing hitam berenang di air (a black dog swimming in the water)
	EfficientNet_b1	seekor anjing hitam mengalir melalui air (a black dog flows through the water)

Fig. 3. CNN-Transformer Model Inference Qualitative Results Comparison With Different Pre-trained CNN Architecture

TABLE I
PRE-TRAINED CNN MODEL RESULTS

Encoder	BLEU 1-4				METEOR	ROUGE L	CIDER
ResNet18	54.21	39.40	28.26	19.98	19.68	43.39	54.78
ResNet34	55.96	41.24	29.41	20.85	20.24	45.05	59.09
ResNet50	**56.57**	**42.16**	**30.57**	**21.82**	20.32	45.21	60.72
ResNet101	55.55	40.45	28.69	20.16	**21.11**	**45.81**	**62.63**
VGG16	55.68	40.61	28.77	20.19	20.29	44.57	59.40
GoogleNet	52.18	37.54	26.69	18.90	18.79	41.89	52.46
EfficientNetb0	52.98	37.55	26.15	18.31	19.09	42.82	53.52
EfficientNetb1	52.80	37.76	26.31	18.20	19.36	43.05	55.06

III. RESULTS AND DISCUSSION

Three Transformer layers using A ResNet50 model as the encoder was our basic configuration for the ResNet-Transformer architecture, where one head is used for each SOTA Transformer layer. Here, we carried out some experiments: one in which we varied the encoder pre-trained model type; another in which we used the inference.

Fig. 3 shows the qualitative model inference comparison, where the ResNet50 generates the Indonesian caption with a stably generated prediction (translated caption can be seen below each generated caption) and the detail of the experiment's quantitative test results in Table I.

On the graphics processing unit (GPU) of a Google Colab Pro, each experiment was trained at a constant learning rate of 0.00004 using the Adam optimizer. It is done within fifty epochs and stopped if there has been no improvement in BLEU-4 throughout the most recent 10 epochs (the halting training criteria), where the overall training process is done in 5-12 hours on each pre-trained CNN architecture.

Python with PyTorch's library is the performance analysis environment for each CNN model that includes three phases: (1) Training phase. (2) Validation. (3) Testing. In other words, we implement the parallelization to it, as the Transformer's architecture supports the simultaneous process.

As seen in Fig. 2, here we changed the model's encoder part of the Transformer with a CNN. Instead of modeling sequence-to-sequence, like in the original Transformers, the modeling is done in a vector-to-sequence way. The input is the image we send into the CNN as the backbone. A Transformer decoder can handle the sequences generation part, which can generate the next word of a sentence. The decoder accepts these input features that extract input images from the CNN backbone as the visual backbone, where they predict the caption generation token by token. The generated captions are formulated as $Caps = (C_0, C_1, C_2, C_3, \dots, C_{token}, C_{token+1})$. The first generated caption $C_0 = < SOS >$ where the "SOS" stands for the start of a sentence, and the $C_{token+1} = < EOS >$ where "EOS" is the unique token meant as the last of the sentence. In short, this model architecture has two different sources of input: (1) The image we want to caption. (2) The very sentence we want it to generate but shifted one word to the left.

To begin, we use trained tokens and positional embeddings to transform the tokens that make up the caption into vectors. After that, we perform the vector's element-wise sum, layer normalization, and drop out. Next, these vectors are processed into a series of transformation layers.

As seen in the proposed model architecture, the model uses the decoder component from the original Transformer. In addition to conducting masked multi-head self-attention on the token vectors, image vectors in each layer implement a two-layer fully-connected network for every vector in turn.

The third step, layer normalization, comes after these three operations and is preceded by a dropout wrapped in a residual connection. Through their attention, token vectors interact with one another token. The masking that occurs throughout this procedure keeps the final predictions' causal structure intact. After applying the last Transformer layer, the unnormalized log probabilities throughout the token vocabulary are predicted by applying a linear layer to each vector that occurs after the application of the end of the Transformer layer. The pre-trained ResNet50 network, after the last convolutional layer, takes an image with dimensions of 224 by 224. It generates a 7 by 7 grid of features with a total of 2048 dimensions.

Because of the unique nature of the pre-training architecture, the CNN channel must be changed to each different model. 512 CNN channel for ResNet18 and ResNet34, 1024 for GoogleNet, 1280 for Efficientnet, 2048 for ResNet50, and ResNet101. The learning rate and epoch values implemented during the training phase were also consistent throughout the experiments.

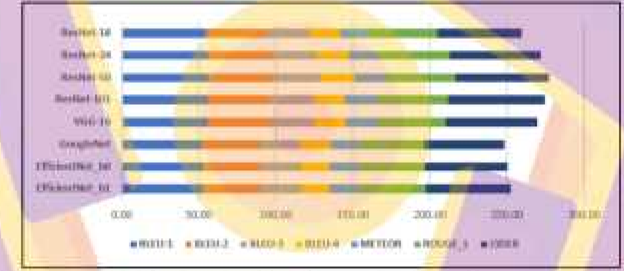


Fig. 4. Pre-trained CNN Architecture Results Comparison

Using the eight's different pre-trained CNN architecture in Fig. 4 shows that the ResNet50-based also has the best overall evaluation result.

As shown in Fig. 4 above, the difference in CNN's channel size or depth affects the prediction results produced. The larger the size, the higher the accuracy value obtained. Based on the visualization of the test results, this increased accuracy value applies to all tested CNN models except for the ResNet101 CNN pre-trained model type with a 2048 channel size. We expect this to occur because our small Flickr8k's Bahasa dataset is underfitting.

We also examined the effect of finetuning the encoder and the model's performance after finetuning. It is accomplished by prohibiting gradient computation for the encoder's second blocks through the fourth convolutional as if we used zero learning rate for these parts. The validation results can be seen in Table II below.

TABLE II
FINETUNED PRE-TRAINED CNN MODEL RESULTS

Encoder	BLEU 1-4				METEOR	ROUGE_L	CIDER
ResNet18	52.91	37.53	25.70	17.39	18.79	42.31	49.85
ResNet34	55.73	40.38	28.41	19.57	19.29	43.53	53.17
ResNet50	**58.10**	**42.91**	**30.40**	21.13	20.12	45.32	60.80
ResNet101	56.24	41.46	29.44	20.28	**20.45**	**45.88**	**61.15**
VGG16	53.70	38.96	26.86	18.00	18.88	42.83	50.97
GoogleNet	52.39	37.12	25.54	17.20	18.49	41.16	48.43
EfficientNetb0	57.73	42.54	30.33	21.10	20.62	45.73	62.57
EfficientNetb1	56.84	42.07	30.24	**21.24**	20.40	45.59	61.09

TABLE III
STATE-OF-THE-ART RESULTS COMPARISONS ON INDONESIAN IMAGE CAPTIONING

Model	Dataset	BLEU 1-4				METEOR	ROUGE_L	CIDER
CNN + GRU [7]	FLICKR30K INDONESIAN	36.70	17.80	06.70	02.00	-	-	-
CNN + LSTM [8]	FLICKR8K FEEH-ID	50.00	31.40	23.90	13.10	-	-	-
CNN + LSTM with Adaptive Attention [18]	MS COCO + FLICKR30K	67.80	51.20	37.50	27.40	-	-	99.00
Ours (CNN + Transformer)	FLICKR8K BAHASA	58.10	42.91	30.40	21.13	20.12	45.32	60.80
Ours (CNN + Transformer)	**FLICKR30K BAHASA**	**75.34**	**62.84**	**50.58**	**40.04**	**27.52**	**58.52**	**110.28**

The finetuned model's results in Table II effectively increase the overall model's result score evaluation except for ResNet18 as it seems other parameters like learning rate or Transformer's layer for the ResNet18-based model need to be readjusted.

With some experimentation, we tested our Transformer-based finetuned model with the larger Flickr30k Bahasa dataset that has been prepared for experimental work. As we expected, the validation results were outstanding, as the Transformer-based model works better with larger training data. Based on the results, we can now compare the model with other previous approaches in Indonesian image captioning. Here, Transformer's context-aware attention mechanism as the model's decoder proved to be better than the previous types that used an RNN-type approach like GRU or LSTM as the model's decoder resulting better evaluation score, as shown in Table III.

IV. CONCLUSION

This study uses several CNN models, namely ResNet18, ResNet34, ResNet50, ResNet101, VGG16, Efficientnet_b0, Efficientnet_b1, and Googlenet, to obtain a CNN model that can produce the best performance as a feature extractor for predicting text sequences performed by the Transformer decoder.

The test is carried out using different sizes of the CNN Channel, where the best model was acquired using ResNet50 and proved that the model could generate grammatically correct Indonesian captions. Experiments indicate that finetuning the encoder model nearly always enhances the decoder model's output, producing a better evaluation score of about 2% than other CNN models.

The ResNet50 model is recommended for using CNN-based systems as the backbone and Transformer as the decoder, where the quantitative results are slightly better than earlier caption generation approaches using the Indonesian dataset. A sensibility analysis on a variety of CNN pre-trained architectures and implementing finetuning to the encoder improve the output of the Transformer-based decoder model for every different pre-trained encoder architecture with BLEU 1-4, METEOR, ROUGE_L, CIDEr of 58.10, 42.91, 30.40, 21.13, 20.12, 45.32, 60.80 respectively for Flickr8k Bahasa and BLEU 1-4, METEOR, ROUGE_L, CIDEr of 75.34, 62.84, 50.58, 40.04, 27.52, 58.52, 110.28 for the final validated model on Flickr30k Bahasa dataset.

As for future work, as our computational resources are platform-limited, further exploration of larger datasets such as Flickr30k, MS COCO 14, MS COCO 17, and other datasets related to image captioning undoubtedly improves the model's performance. Hopefully, as this finding only focuses on the encoder part of the model, it would be fascinating to test the impact of employing pre-trained word embeddings for the decoder part, mainly in Indonesian, as well as a more complex Transformers-based model.

REFERENCES

[1] R. Subash, R. Jebakumar, Y. Kamdar, and N. Bhatt, "Automatic image captioning using convolution neural networks and LSTM," *J. Phys. Conf. Ser.*, vol. 1362, no. 1, 2019, doi: 10.1088/1742-6596/1362/1/012096.

[2] M. D. Zakir Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, 2019, doi: 10.1145/3295748.

[3] K. C. Nithya and V. V. Kumar, "A Review on Automatic Image Captioning Techniques," *Proc. 2020 IEEE Int. Conf. Commun. Signal Process. ICCSP 2020*, pp. 432–437, 2020, doi: 10.1109/ICCSP48568.2020.9182105.

[4] C. Wang, H. Yang, and C. Meinel, "Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 14, no. 2s, 2018, doi: 10.1145/3115432.

[5] C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, pp. 1–4, 2018, doi: 10.1109/ICCUBEA.2018.8697360.

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

[7] A. A. Nugraha, A. Arifianto, and Suyanto, "Generating image description on Indonesian language using convolutional neural network and gated recurrent unit," *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, pp. 1–6, 2019, doi: 10.1109/ICoICT.2019.8835370.

[8] E. Mulyanto, E. I. Setiawan, E. M. Yuniarno, and M. H. Purnomo, "Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset," *2019 IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl. CIVEMSA 2019 - Proc.*, 2019, doi: 10.1109/CIVEMSA45640.2019.9071632.

[9] R. Mulyawan, A. Sunyoto, and A. H. Muhammad, "Automatic Indonesian Image Captioning using CNN and Transformer-Based Model Approach," in *2022 5th International Conference on Information and Communications Technology (ICOIACT)*, 2022, pp. 355–360, doi: 10.1109/ICOIACT55506.2022.9971855.

[10] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 2641–2649, 2015, doi: 10.1109/ICCV.2015.303.

[11] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional Image Captioning," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 5561–5570, 2018, doi: 10.1109/CVPR.2018.00583.

[12] S. Liu, L. Bai, Y. Hu, and H. Wang, "Image Captioning Based on Deep Neural Networks," *MATEC Web Conf.*, vol. 232, pp. 1–7, 2018, doi: 10.1051/matecconf/201823201052.

[13] H. Shi, P. Li, B. Wang, and Z. Wang, "Image captioning based on deep reinforcement learning," *ACM Int. Conf. Proceeding Ser.*, vol. 01052, pp. 1–7, 2018, doi: 10.1145/3240876.3240900.

[14] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," *MM 2017 - Proc. 2017 ACM Multimed. Conf.*, pp. 1549–1557, 2017, doi: 10.1145/3123266.3123366.

[15] X. Li, W. Lan, J. Dong, and H. Liu, "Adding Chinese captions to images," *ICMR 2016 - Proc. 2016 ACM Int. Conf. Multimed. Retr.*, pp. 271–275, 2016, doi: 10.1145/2911996.2912049.

[16] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR captions: Constructing a large-scale Japanese image caption dataset," *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, pp. 417–421, 2017, doi: 10.18653/v1/P17-2066.

[17] H. A. Al-muzaini, T. N. Al-yahya, and H. Benhidour, "Automatic Arabic image captioning using RNN-LSTM-based language model

and CNN," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 6, pp. 67–73, 2018, doi: 10.14569/IJACSA.2018.090610.

[18] M. R. S. Mahadi, A. Arifianto, and K. N. Ramadhani, "Adaptive Attention Generation for Indonesian Image Captioning," *2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020*, 2020, doi: 10.1109/ICoICT49345.2020.9166244.

[19] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 3, pp. 2048–2057, 2015.

[20] L. Chen *et al.*, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 6298–6306, 2017, doi: 10.1109/CVPR.2017.667.

[21] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Show, observe and tell: Attribute-driven attention model for image captioning," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2018-July, pp. 606–612, 2018, doi: 10.24963/ijcai.2018/84.

[22] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 4651–4659, 2016, doi: 10.1109/CVPR.2016.503.

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

[24] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 1–11, 2019.

[25] V. Atliha and D. Šešok, "Text augmentation using BERT for image captioning," *Appl. Sci.*, vol. 10, no. 17, 2020, doi: 10.3390/app10175978.

[26] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu, "Captioning transformer with stacked attention modules," *Appl. Sci.*, vol. 8, no. 5, 2018, doi: 10.3390/app8050739.

[27] W. Zhang, W. Nie, X. Li, and Y. Yu, "I m age Caption Generation With Adaptive Transfor m er," pp. 521–526, 2019.

[28] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image Captioning Through Image Transformer," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12625 LNCS, pp. 153–169, 2021, doi: 10.1007/978-3-030-69538-5_10.

[29] C. Cormier, "Bleu," *Landscapes*, vol. 7, no. 1, pp. 16–17, 2005, doi: 10.3917/chev.030.0107.

[30] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," *Intrinsic Extrinsic Eval. Meas. Mach. Transl. and/or Summ. Proc. Work. ACL 2005*, no. June, pp. 65–72, 2005.

[31] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Proc. Work. text Summ. branches out (WAS 2004)*, no. 1, pp. 25–26, 2004, [Online]. Available: papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85.

[32] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 07-12-June, pp. 4566–4575, 2015, doi: 10.1109/CVPR.2015.7299087.

[33] R. Staniute and D. Šešok, "A systematic literature review on image captioning," *Appl. Sci.*, vol. 9, no. 10, 2019, doi: 10.3390/app9102024.