Image Captioning Menggunakan *Transformer* Dan Area *Attention*

DHIKI ROMADINUR[1*], JASMAN PARDEDE[1]

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

Received 04 09 2023 **|** *Revised* 11 09 2023 **|** *Accepted* 11 09 2023

ABSTRAK

Image Captioning adalah menghasilkan deskripsi teks yang akurat dan relevan dari sebuah gambar. Penelitian dilakukan dengan menggunakan dataset dari MS COCO 2014. Metode yang diterapkan pada penelitian ini adalah Transformer dengan Area Attention MobilenetV3Small untuk membangun model. Dalam penelitian ini menggunakan ekstraksi fitur MobilenetV3Small. Pengujian yang dilakukan dengan evaluasi skor BLEU. Pengukuran BLEU menggunakan 4-gram terdiri skor BLEU-1, BLEU-2, BLEU-3, BLEU-4. Dengan proses epoch sebanyak 100 kali, Dengan hasil skor BLEU yang dihasilkan dengan rata-rata skor yang didapatkan adalah {0.557348, 0.354169, 0.183363, 0.098632}.

***Kata kunci**: Image Caption ing, Transformer, Attention, MobilenetV3Small, BLEU*

ABSTRACT

Image Captioning is producing an accurate and relevant text description of an image. The research was conducted using the dataset from MS COCO 2014. The method used in this study was Transformer with MobilenetV3Small Attention Area to build the model. In this study using MobilenetV3Small feature extraction. The test was carried out by evaluating the BLEU score. The BLEU measurement uses 4-grams consisting of BLEU-1, BLEU-2, BLEU-3, BLEU-4 scores. With an epoch process of 100 times, the resulting BLEU score results with an average score obtained is {0.557348, 0.354169, 0.183363, 0.098632}.

Keywords: *Image Caption ing, Transformer, Attention, MobilenetV3Small, BLEU.*

1. PENDAHULUAN

Image Captioning adalah menghasilkan deskripsi teks yang akurat dan relevan dari sebuah gambar. Pembuatan teks adalah masalah kecerdasan buatan yang menantang dimana dekripsi tekstual harus dibuat untuk foto tertentu **(Radhakrishnan, 2017).** Masalah teks gambar otomatis oleh sistem kecerdasan buatan telah menerima banyak perhatian dalam beberapa tahun terakhir, karena keberhasilan model pembelajaran mendalam untuk bahasa dan pemprosesan gambar. Sebagian besar pendekatan teks gambar dalam literatur didasarkan pada terjemahan pendekatan, dengan encoder visual dan decoder linguistik. Salah satu tantangan dalam penerjemahan otomatis adalah bahwa hal itu tidak dapat dilakukan kata demi kata, tetapi kata-kata lain mempengaruhi makna, dan oleh karena itu, terjemahan dari sebuah kata ini bahkan lebih benar ketika menerjemahkan lintas modalitas, dari gambar ke teks, dimana sistem harus memutuskan apa harus dijelaskan dalam gambar **(Sen He et al, 2020).**

Pembuatan teks adalah masalah kecerdasan buatan yang menantang dimana deskripsi tekstual harus dibuat untuk foto tertentu. Dibutuhkan kedua metode dari visi komputer untuk memahami isi gambar dan model bahasa dari bidang pemrosesan bahasa alami untuk mengubah pemahaman gambar menjadi kata-kata dalam urutan yang benar **(Jason Brownlee, 2019).**

Arsitektur Transformer merupakan sebuah arsitektur yang dapat mengubah satu urutan menjadi urutan lain dengan menggunakan dua bagian yaitu yaitu enkoder dan dekoder **(Vaswani et al., 2017)**. Area *Attention* adalah model berbasis perhatian baru untuk teks gambar otomatis. Pendekatan memodelkan ketergantungan antara wilayah gambar, kata keterangan, dan status model bahasa RNN, menggunakan tiga interaksi berpasangan. Pada penelitian sekarang menggunakan arsitektur baru yang disebut transformer. Dimana arsitektur ini sepenuhnya dibangun di atas-*Self Attention* sendiri tanpa mengandalkan Recurrent network (GRU, LSTM). Transformer arsitektur ini untuk mengubah satu urutan ke urutan lain dengan bantuan enkoder dan dekoder.

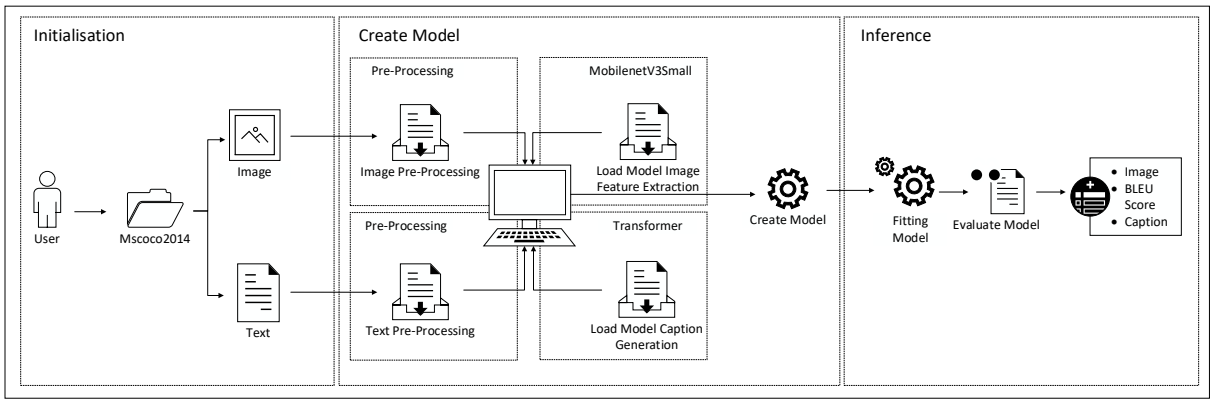
Image Captioning adalah mendeskripsikan gambar secara otomatis. Dilakukan dengan dua metode yaitu menggunakan *natural language processing* (NLP) dan computer vision. *Natural language processing* untuk menghasilkan deskripsi dari gambar berupa bahasa alami sedangkan computer vision untuk mengenali objek pada gambar **(Vinyal et al., 2015).**

Penelitian ini menggunakan Area *Attention* MobilenetV3Small dan metode Transformer. Menggunakan dataset MS COCO 2014 dengan pengujian hasil dari program yang telah dibuat menggunakan evaluasi BLEU (Bilingual Evaluation Understudy). BLEU merupakan sebuah metode yang dirancang untuk mengevaluasi sistem terjemahan mesin otomatis **(Diana et al.,2005).** Sebuah algoritma yang hanya mencerminkan bagaimana kinerja sistem pada rangkaian kalimat sumber dan terjemahan yang dipilih untuk pengujian. Metrik BLEU mengukur penilaian terjemahan dalam rentang skala 0 hingga 1. Semakin mendekati 1 menunjukkan bahwa semakin baik sistemnya.

METODE PENELITIAN

2.1 Perancangan Umum

Perancangan pada sistem *Image Captioning* terdiri dari initialization, pre-processing, image feature extraction, *caption* generation, create model, dan inference. Blok diagram penelitian secara umum ditunjukkan pada Gambar 1.



Gambar 1. Blok Diagram Penelitian

Berikut penjelasan dari blok diagram penelitian, pada tahap initialization user mempersiapkan dataset MS COCO 2014 yang berisi data gambar, dan data text yaitu. Setelah data berhasil diperoleh selanjutnya menampilkan data gambar berdasarkan 5 keterangan yang menggambarkan gambar tersebut. Kemudian pada bagian pre-processing dilakukan text pre-processing terhadap data *caption* dengan membuat standarisasi, tokenizing, proses word to index – index to word. Lalu dilakukan image pre-processing pada data gambar yang meliputi mengubah ukuran gambar yang sesuai dengan *input* model MobilenetV3Small. Selanjutnya image feature extraction menggunakan MobilenetV3Small dan *caption* generation menggunakan Transformer. Berikutnya melakukan model fitting dengan menggunakan epoch sebanyak 100 kali untuk melatih model. Model yang sudah dibuat lalu dievaluasi dengan menggunakan BLEU scoring. Untuk hasil *caption* generation dan evaluasi menampilkan kalimat *caption* asli, *caption* prediksi, skor BLEU, dan gambar dari *caption* .

2.2 Dataset

Dataset yang digunakan adalah MS COCO 2014 yang diambil dari cocodataset.org, dataset ini memiliki 82.783 gambar dengan 413.915 kalimat dalam training, 40.504 gambar dengan 202.520 kalimat dalam *Validation*, dan 40.775 gambar dengan 379.249 kalimat dalam testing. Pada setiap data gambar memiliki 5 kalimat *caption* yang mendeskripsikan gambar. Pada penelitian ini menggunakan 8000 kosakata dari keterangan karena sekitar 8.000 kata unik yang muncul pada keterangan dataset.

2.3 Pre-Processing

Image Pre-processing untuk peningkatan data citra yang menekan distorsi yang tidak diinginkan atau meningkatkan fitur citra yang penting. pada gambar image Pre-processing ini dengan mengubah gambar menjadi 224 x 224 x 3 dan menormalisasi gambar *input* agar sesuai dengan ketentuan *input* MobilenetV3Small.

Text pre-processing merupakan proses mengubah teks mentah menjadi bentuk yang lebih terstruktur, mudah diolah, dan sesuai dengan kebutuhan analisis teks atau pemodelan bahasa alami untuk meningkatkan kualitas data teks dan menghilangkan gangguan yang tidak relevan agar pemprosesan dapat dilakukan dengan lebih efektif. Text pre-processing yang dilakukan pada *caption* terdiri dari beberapa proses yaitu:

a. Standarisasi

Case Folding proses untuk semua karakter diubah menjadi huruf kecil.

Remove Punctuation untuk menghapus tanda baca seperti '!"#$%&()*+.,-/:;=?@[\]^_`{|}~'. dan untuk melepaskan karakter apa pun dalam tanda baca.

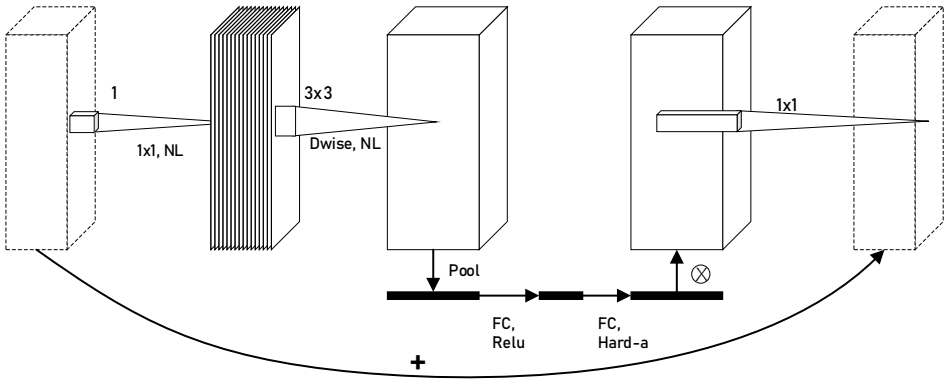
Mark *Caption* untuk penambahan token "<start>" dan "<end>" sebagai kata-kata terpisah untuk menandai awal dan akhir teks. Proses ini dilakukan agar sistem dapat mengenali bagian awal dan akhir kalimat *caption* dan memutuskan untuk mulai menghasilkan kalimat dan berhenti untuk menghasilkan kalimat yang diprediksi.

b. Tokenizing

Tokenizing adalah memisahkan setiap kata pada *caption* menjadi token, mengganti kata yang tidak termasuk ke dalam word list menjadi '<unk>' unknown.

2.4 MobilenetV3Small

MobileNet adalah sebuah model neural network arsitektur yang dirancang khusus untuk dijalankan pada perangkat mobile atau embedded system. Model ini sangat efisien dan memiliki ukuran yang relatif kecil. MobilenetV3Small ditargetkan pada penggunaan sumber daya tinggi dan rendah. Pada MobilenetV3Small memiliki latensi deteksi yang lebih cepat dan memiliki akurasi yang lebih kecil dari MobilenetV3Large **(Howard et al., 2019)**. Berikut blok dari MobilenetV3Small pada Gambar 2.



Gambar 2. MobilenetV3Small Block

Model MobilenetV3Small ini digunakan untuk melakukan image feature extraction. Ukuran *input* gambar yang dapat dimasuk ke dalam arsitektur berukuran 224 x 224 x 3 pixel. Untuk melakukan proses ekstraksi fitur, lapisan terakhir atau lapisan klarifikasi pada model dihapus karena penelitian ini tidak melakukan klarifikasi. Konfigurasi arsitektur MobilenetV3Small ditunjukkan pada Tabel 1.

Tabel 1. Konfigurasi Arsitektur MobilenetV3Small

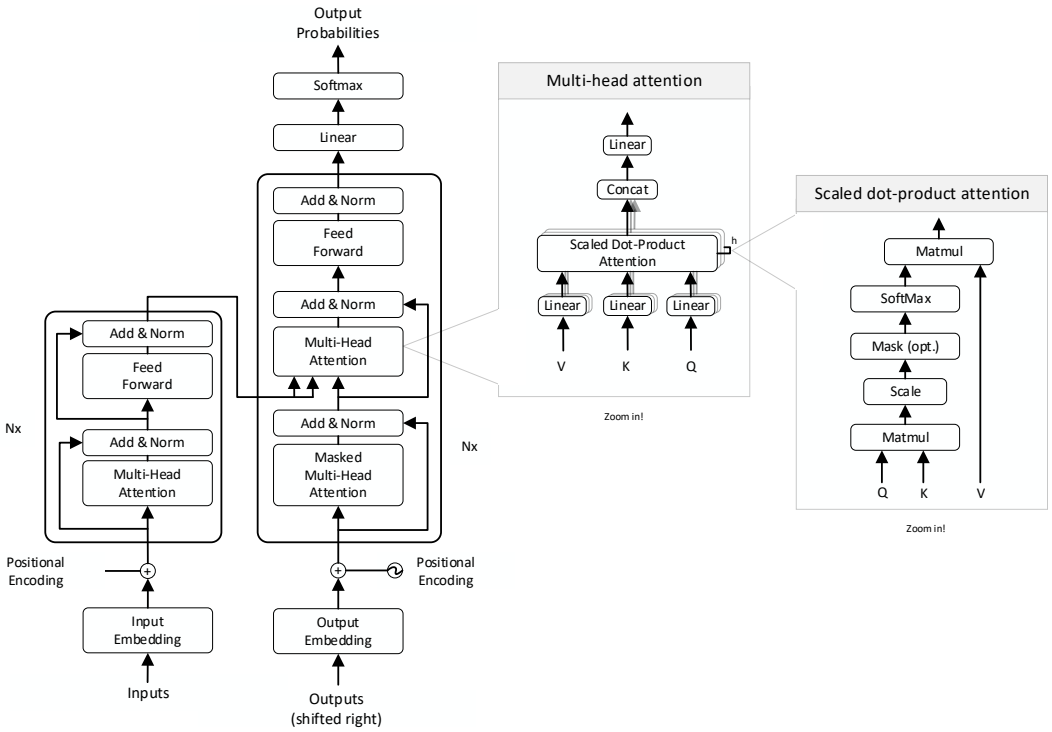
Input	**Operator**	**exp size**	***#out***	**SE**	**NL**	***s***
$224^2 \; x \; 3$	Conv2d, 3x3	-	16	-	HS	2
$112^2 \; x \; 16$	bneck, 3x3	16	16	✓	RE	2
$56^2 \; x \; 16$	bneck, 3x3	72	24	-	RE	2
$28^2 \; x \; 24$	bneck, 3x3	88	24	-	RE	1
$28^2 \; x \; 24$	bneck, 5x5	96	40	✓	HS	2
$14^2 \; x \; 40$	bneck, 5x5	240	40	✓	HS	1
$14^2 \; x \; 40$	bneck, 5x5	240	40	✓	HS	1
$14^2 \; x \; 40$	bneck, 5x5	120	48	✓	HS	1
$14^2 \; x \; 48$	bneck, 5x5	144	48	✓	HS	1

14^2 x 48	bneck, 5x5	288	96	✓	HS	2
7^2 x 96	bneck, 5x5	576	96	✓	HS	1
7^2 x 96	bneck, 5x5	576	96	✓	HS	1
7^2 x 96	conv2d, 1x1	-	576	✓	HS	1
7^2 x 576	pool, 7x7	-	-	-	-	1
1^2 x 576	conv2d 1x1, NBN	-	1024	-	HS	1
1^2 x 1024	conv2d 1x1, NBN	-	k	-	-	1

Exp size (expansion size), #out (jumlah filter), SE (Squeeze and Excite), HS (hard-swish) dan RE Rectified Linear Unit (ReLu). NBN (no batch normalization), dan s (stride).

2.4 Transformer

Salah satu karya yang sangat berpengaruh adalah makalah berjudul "*Attention* is All You Need" **(Vaswani et al., 2017)**. Dalam makalah ini, diperkenalkan suatu struktur baru yang dikenal dengan Transformer. Arsitektur ini sepenuhnya dibangun di atas *Self-Attention* tanpa mengandalkan Recurrent Network (GRU, LSTM). Transformer adalah arsitektur untuk mengubah satu urutan ke urutan lain dengan dua komponen utama yaitu encoder dan decoder. Berikut model Arsitektur Transformer pada Gambar 3.



Gambar 3. Arsitektur Transformer

Transformer memiliki tampilan arsitektur lengkap seperti urutan sumber dan urutan target pertama-tama memasuki bagian embedding layer untuk dapat menghasilkan data dengan dimensi yang sama yaitu dmodel = 512, untuk mempertahankan informasi posisi dari *input* data maka diterapkan sebuah sinusoid-wave-based positional encoding dan dijumlahkan dengan *output* embedding, dan softmax layer dan linear layer ditambahkan ke decoder *output* akhir.

a. Query, Key, dan Query

Transformer terdapat Komponen utama adalah unit dari multi-head *Self-Attention*. Transformer melihat representasi encoder dari *input* sebagai pasangan key-value, (K, V), n

dimensi (panjang urutan *input*). Dalam konteks NMT, keduanya adalah *encoder hidden states*. Dalam decoder, *output* sebelumnya dikompresi ke dalam sebuah query (Q dari m dimensi) dan *output* teks yang dihasilkan oleh pemetaan query, keys dan values **(Weng, 2018).**

$$\textit{Attention}\ (Q, K, V) = \text{softmax}\ (\frac{QK^T}{\sqrt{d_k}})\ V \quad (1)$$

Keterangan :
Q = Matriks dari query
K = Matriks dari key
V = Matriks dari value
M = Optional mask
dk = dimensi dari ke

b. Multi-Head *Self-Attention*

Multi-head *Attention* berjalan melalui scaled dot-product *Attention* berkali-kali dalam parallel. *Attention* yang independen pada *output* diintegrasikan dan disesuaikan linier ke dalam dimensi yang diinginkan **(Weng, 2018).** Multi-head *Self-Attention* terdiri dari lapisan linier, scaled dot-product, concat, dan linier akhir. Berikut rumus perhitungan multi-head *Self-Attention*.

$$MultiHead(Q,K,V) = Concat(head_1, \dots, head_h)W^o \quad (2)$$
$$Whereheadi = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

Di mana $W_i^Q, W_i^K, W_i^V, dan\ W_i^O$ adalah matrik parameter untuk dipelajari.

c. Embedding dan Softmax

Embedding digunakan untuk mengubah representasi kata dari token *output* dan token *output* menjadi bentuk yang dapat diproses oleh transformer yaitu vektor dimensi d_{model}. Softmax digunakan untuk menghasilkan probabilitas prediksi pada tahap *output* **(Vaswani et al., 2017).**

d. Position Encoding

Positional encoding menggunakan untuk menyuntikkan informasi posisi ke model atau untuk menangkap urutan secara beraturan. Encoding posisi menggunakan sinusoid tetap dari frekuensi berbeda yang ditambahkan langsung ke embedding *input*. Fungsi cos jika posisi indeks ganjil pada vektor *input*, sedangkan fungsi sin jika posisi indeks genap pada vektor *input*. Menambahkan vektor-vektor tersebut ke embedding *input* yang sesuai yang berhasil memberikan informasi jaringan pada posisi setiap vektor **(Vasnawi et al., 2017**).

$$PE(pos, 2i) = sin(\frac{pos}{10000^{2i/d_{model}}}) \quad (4)$$

$$PE(pos,2i+1) = cos(\frac{pos}{10000^{2i/d_{model}}}) \quad (5)$$

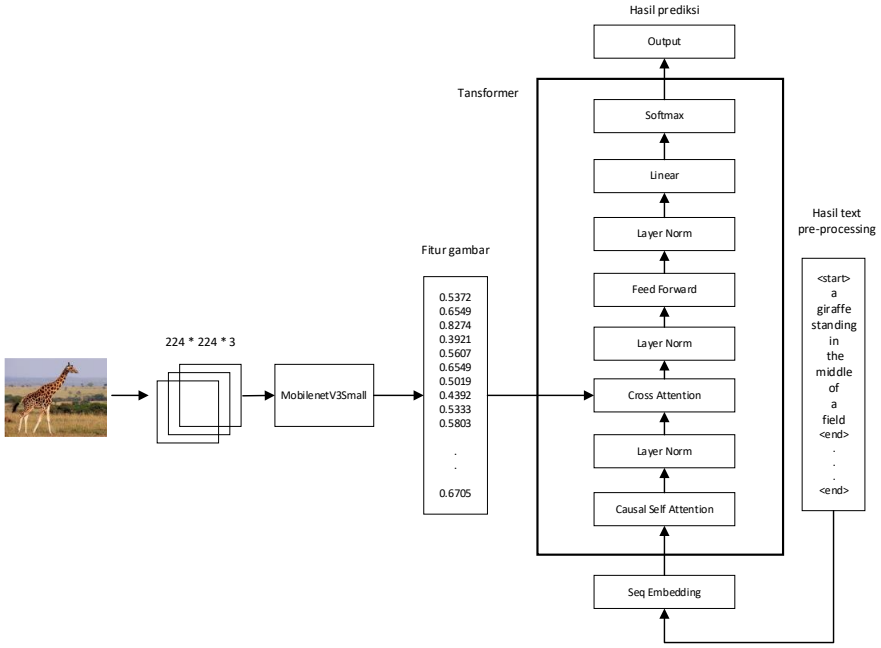
e. Feed Forward Neural Network

Feed Forward Neural Network ditambahkan ke setiap sub-layer dari *Attention* yang ada pada encoder dan decoder. Lapisan ini ditambah pada posisi yang terpisah tetapi memiliki lapisan identik antara satu sama lainnya. *Feed forward network* memiliki dua transformasi linier dan menggunakan aktivasi Relu didalamnya **(Vasnawi et al., 2017).**

$$FFN(x) = max(0, xW_1 + b_1)\ W_2 + b_2 \quad (6)$$

2.5 Pembangunan Model

Bagian ini membahas mengenai pembangunan model yang digunakan yaitu Area *Attention* menggunakan MobilenetV3Small dan Transformer. Gambar 4 merupakan Arsitektur yang digunakan pada penelitian ini.



Gambar 4. Pembangunan Model

Untuk encoder menggunakan Area *Attention* dari MobilenetV3Small dan dekoder menggunakan transformer dibangun dari *Attention* layer. *Self Attention* untuk memproses urutan kata yang dihasilkan, dan menggunakan *cross Attention* untuk menghubungkan antara representasi kata-kata dan fitur gambar untuk menghasilkan *Attention* scores yang mencerminkan sejauh mana saling berhubungan setiap elemen teks dengan gambar dan sebaliknya, menggunakan area perhatian untuk menentukan bagian gambar mana yang relevan dengan kata-kata yang sedang dihasilkan yang memungkinkan *cross Attention* untuk lebih spesifik dalam mendeskripsi teks dari gambar dengan menghubungkan setiap lokasi keluaran dari teks dapat memperhatikan gambar masukan, seperti hubungan antara teks dan gambar. feedforward untuk memproses representasi dari fitur visual dan teks dan memiliki proses keluaran secara mandiri. mobilenetV3small dioptimalkan untuk ekstraksi gambar yang mampu menangkap hasil visual dan spesial yang penting dari gambar, yang diperlukan untuk menghasilkan deskripsi yang tepat

Decoder terdiri dari beberapa bagian seperti memiliki 6 lapisan identik, setiap lapisan memiliki dua sub-lapisan dari *multi-head Self-Attention* dan satu-lapisan *feed forward network*, setiap sub-lapisan terdapat koneksi residual dan lapisan normalization yang semua sub-layer memiliki dimensi yang sama yaitu dmodel = 512, dan setiap sub-lapisan masked *multi-head Self-Attention* yang pertama dimodifikasi untuk mencegah berpindah ke posisi berikutnya. Terjadi proses penggabungan *output* dan *output* sebelumnya untuk menutupi *output* yang akan datang sehingga setiap *output* hanya akan memperlihatkan *output-output* sebelumnya **(Vaswani et al., 2017).**

2.6 BLEU (Bilingual Evaluation Understudy)

BLEU (*Bilingual Evaluation* Understudy) adalah algoritma untuk mengevaluasi kualitas teks yang telah diterjemahkan dari mesin terjemahan. Skor BLEU hanya mencatat sejauh mana

sistem berkinerja pada kumpulan kalimat sumber dan terjemahan yang dipilih untuk pengujian. Nilai BLEU untuk menilai terjemahan berada pada skala 0 sampai 1. Semakin mendekati 1 skor kalimat uji, semakin tumpang tindih dengan terjemahan referensi manusia maka semakin baik sistemnya. Untuk menghitung hasil BLEU, python menyedikan library Natural language toolkit atau NLTK yang dapat digunakan untuk mengevaluasi teks yang di gene*rate* terhadap suatu referensi teks. Dalam pustaka NLTK, terdapat fungsi skor BLEU kalimat yang dimanfaatkan untuk menilai kalimat-kalimat kandidat terhadap satu atau beberapa kalimat referensi.

$$BP_{BLEU} = \{ 1 \quad if\ c > r\ e\ (1 - \frac{r}{c})\ if\ c \leq \mathrm{r} \tag{7}$$

$$P_n = \frac{\sum_{\mathrm{C} \in \text{ corpus n-gram} \in \mathrm{C}} \quad \sum \quad count_{clip(n-gram)}}{\sum_{\mathrm{C} \in \text{ corpus n-gram} \in \mathrm{C}} \quad \sum \quad count_{(n-gram)}} \tag{8}$$

$$BLEU = BP_{BLEU}.\ e\ \sum_{n-1}^{N} W_n \log P_n \tag{9}$$

Keterangan:
BP = brevity penalty
c = jumlah kata dalam hasil terjemahan otomatis
r = jumlah kata dalam kata rujukan
Pn = modified precision score
Wn = 1/N (nilai N untuk standar BLEU adalah 4)
pn = jumlah n-gram dalam hasil terjemahan yang cocok dengan n-gram dalam rujukan, dibagi dengan jumlah total n-gram yang dihasilkan terjemahan.

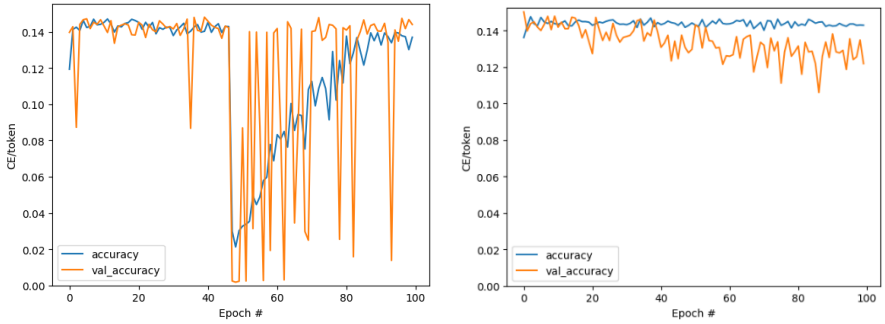
3. HASIL DAN PEMBAHASAN

Berdasarkan pelatihan ditujukan untuk membuat sebuah model terbaik Transformer dan menggunakan Area *Attention* MobilenetV3Small. Pelatihan terdapat 4 *learning rate* berbeda yang digunakan pada penelitian ini diantaranya, *learning rate* 0,1 dan minimum *learning rate* 0,0001 pada saat training untuk kemudian dibandingkan berdasarkan hasil pelatihan yang dihasilkan. Pelatihan didukung dengan penggunaan 100 epoch, dikarenakan tidak terjadi peningkatan akurasi dan validasi akurasi pada saat training, dengan menggunakan optimizer adam.
Dari hasil pelatihan dapat dibandingkan untuk *learning rate* paling tinggi didapatkan pada *learning rate* 0,0001 dengan nilai 0.4504. Perbandingan *learning rate* pada Tabel 2

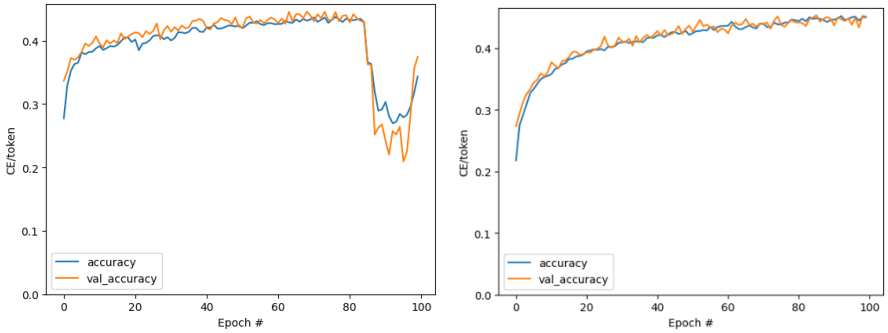
Tabel 2. Hasil Training

Learning Rate	***Accuracy***	***Validation Accuracy***
0,1	0.1463	0.1476
0,01	0.1436	0.1241
0,001	0.4338	0.4305
0,0001	0.4504	0.4515

Berikut grafik dari proses training dengan 100 kali *epoch model* menunjukkan *accuracy* serta *Validation accuracy* dan *accuracy* dari setiap *learning rate* ditunjukkan pada Gambar 5 dan Gambar 6.



Gambar 5. Grafik *Accuracy*, dan *Validation Accuracy learning rate* 0,1 dan 0,01



Gambar 6. Grafik *Accuracy*, dan *Validation Accuracy learning rate* 0,001 dan 0,0001

Pengujian kinerja menggunakan parameter skor BLEU (*Bilingual Evaluation Understudy*) digunakan untuk menilai sebuah kalimat dibandingkan dengan satu kalimat atau lebih kalimat referensi. Hasil *caption* akan dibandingkan dengan *caption* asli dari gambar yang telah ada sebelumnya. Perhitungan skor BLEU dilakukan dengan skor BLEU-1, BLEU-2, BLEU-3, BLEU-4, skor BLEU dari 0-1 dengan skor 1 hanya akan didapatkan apabila kalimat yang dihasilkan mesin benar-benar sama dengan kalimat referensi yang sudah ada. Untuk menghitung Skor BLEU menggunakan Natural Language Toolkit Library (NLTK), pendekatan ini mempermudah perbandingan antara kalimat yang dihasilkan oleh mesin dengan referensi. Di sini, konsep n-gram digunakan satu set 'n' kata berurutan dalam sebuah kalimat. Terdiri dari 1-gram (unigram) terdiri dari satu kata, 2-gram (bigram) kombinasi dua kata yang muncul secara berurutan, 3-gram (trigram) memperlihatkan kombinasi dari tiga kata yang muncul secara berurutan, dan 4-gram memperlihatkan kombinasi dari empat kata yang muncul secara berurutan. Misalnya dalam kalimat " I ate an apple today", untuk mendapatkan hasil dari setiap n-gram yaitu :

- 1-gram (unigram) : "I", "ate", "an", "apple", "today"
- 2-gram (bigram) : "I ate", "ate an", "an apple", "apple today"
- 3-gram (trigram) : "I ate an", "at an apple", " an apple today"
- 4-gram : "I ate an apple", "ate an apple today"

Fungsi utama BLEU adalah membandingkan n-gram dalam kalimat kandidat dengan n-gram dalam referensi terjemahan. Di sini, n-gram merujuk pada urutan kata yang muncul dalam kalimat, dan nilai n menunjukkan seberapa panjang urutan tersebut. Semakin besar nilai n, semakin baik evaluasinya terhadap terjemahan kandidat, karena kata yang lebih panjang.

Dilakukan pengujian berdasarkan parameter skor BLEU. Berdasarkan hasil dari pengujian maka dihasilkan rata-rata untuk masing-masing dari skor BLEU ditunjukkan ditunjukkan pada Tabel 3.

Tabel 3. Hasil Kinerja

No	BLEU	Hasil Skor
1	BLEU-1	0,557348
2	BLEU-2	0,354169
3	BLEU-3	0,183363
4	BLEU-4	0,098632

Berdasarkan (*Lavie*, 2010) nilai BLEU diatas 0.3 dapat mengindikasikan terjemahan yang dapat dimengerti, sementara nilai BLEU diatas 0.5 dapat menunjukkan terjemahan yang baik dan fasih. Ada sebuah konsep yang menjadi

4. KESIMPULAN

Pada penelitian ini telah diimplementasikan Area *Attention* dari MobilenetV3Small dan Transformer dalam melakukan *Image Captioning* sebagai image *feature extraction* dan *caption generation*. Model yang diusulkan mampu mengekstraksi fitur gambar dan mengubah menjadi kalimat untuk mendeskripsikan gambar.

Dengan menggunakan dataset MS COCO 2014, yang terdiri dari 82.783 data latih, 40.504 data validasi, dan 40.775 data uji. Hasil pengujian dengan perbandingan yang dilakukan menggunakan beberapa *learning rate* menunjukan bahwa pada *learning rate* 0,0001 menggunakan optimizer adam dengan epoch 100 mendapatkan hasil terbaik.

Model uji dengan parameter skor BLEU scoring yang dimana nilai paling tinggi maka model mampu menghasilkan kalimat deskriptif yang serupa dengan kalimat kandidat. BLEU menggunakan beberapa parameter skor yang disebut BLEU-1, BLEU-2, BLEU-3, dan BLEU-4. Berdasarkan model tersebut nilai skor BLEU-1, BLEU-2, BLEU-3, BLEU-4 dengan rata-rata skor masing-masing adalah 0.557348, 0.354169, 0.183363, 0.098632. Dengan skor BLEU tersebut dapat disimpulkan bahwa model yang dibangun sudah cukup baik untuk menghasilkan kalimat dari suatu gambar.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Program Studi Informatika Institut Teknologi Nasional (ITENAS) Bandung. Penulis juga berterima kasih kepada pihak yang telah membantu dan mendukung sehingga penelitian ini dapat berjalan dengan baik.

DAFTAR RUJUKAN

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

Pérez, D., Alfonseca, E., & Rodríguez, P. (n.d.). Application of the BLEU method for evaluating free-text answers in an e-*learning* environment. http://labs.google.com/glossary

Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019). Image *caption* ing: Transforming objects into words. Advances in Neural Information Processing Systems, 32, 1–11.

He, S., Liao, W., Tavakoli, H. R., Yang, M., Rosenhahn, B., & Pugeault, N. (2021). Image *Captioning* Through Image Transformer. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12625 LNCS, 153–169. https://doi.org/10.1007/978-3-030-69538-5_10

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. v., & Adam, H. (2019). Searching for MobileNetV3. http://arxiv.org/abs/1905.02244

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

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014). Microsoft COCO: Common Objects in Context. http://arxiv.org/abs/1405.0312

Li, Y., Kaiser, L., Bengio, S., & Si, S. (2019). [Query] 2 Area *Attention*. 36th International Conference on Machine *Learning*, ICML 2019, 2019-June, 6833–6846.

Lavie, A. (2010). Evaluating the *output* of machine translation systems. AMTA 2010 - 9th Conference of the Association for Machine Translation in the Americas.

Liu, W., Chen, S., Guo, L., Zhu, X., & Liu, J. (2021). CPTR: Full Transformer Network for Image *Caption* ing. 1–5. Retrieved from http://arxiv.org/abs/2101.10804

Pan, Y., Yao, T., Li, Y., & Mei, T. (2020). X-Linear *Attention* Networks for Image *Caption* ing. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 10968–10977. https://doi.org/10.1109/CVPR42600.2020.01098\

Pedersoli, M., Lucas, T., Schmid, C., & Verbeek, J. (2017). Areas of *Attention* for Image *Caption* ing. Proceedings of the IEEE International Conference on Computer Vision, 2017-October(ii), 1251–1259. https://doi.org/10.1109/ICCV.2017.140

Radhakrishnan, P. (2017, Jan 19). Image *Caption* ing. Retrived from www.towardsdatascience.com

Tavakoliy, H. R., Shetty, R., Borji, A., & Laaksonen, J. (2017). Paying *Attention* to Descriptions Gene*rate*d by *Image Captioning* Models. Proceedings of the IEEE International

Conference on Computer Vision, 2017-October, 2506–2515. https://doi.org/10.1109/ICCV.2017.272

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

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2014). Show and Tell: A Neural Image *Caption* Generator. http://arxiv.org/abs/1411.4555

Wang, W., Chen, Z., & Hu, H. (2019). Hierarchical *Attention* network for image *caption* ing. 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, 8957–8964. https://doi.org/10.1609/aaai.v33i01.33018957

Weng, L. (2018). *Attention*? *Attention*. Github. https://lilianweng.github.io/lillog/2018/06/24/*Attention*-*Attention*.html#whats-wrong-with-seq2seq-model.