ISSN [e]: XXXX-XXX
DOI: xxx

Xception Dan Gated Recurrent Unit Pada Image Captioning

Josua Sirait[1], Jasman Pardede[2]
[1, 2] Program Studi Informatika, Institut Teknologi Nasional Bandung, Indonesia
Email : josuasirait25@mhs.itenas.ac.id
Received DD/01/2022 **|** *Revised* DD/01/2022 **|** *Accepted* DD/01/2022

ABSTRAK

Image Captioning adalah proses menghasilkan deskripsi tekstual untuk gambar yang diberikan. Dengan melibatkan area dari Computer Vision untuk memahami konten gambar dan model bahasa dari bidang Natural Language Processing (NLP) untuk mengubah pemahaman gambar menjadi kata-kata dalam urutan yang benar. Berdasarkan dari penelitian-penelitian yang ada, penelitian ini mencoba mengimplementasikan arsitektur encoder-decoder melihat dari penelitian sebelumnya pada Image Captioning. Metode yang digunakan yaitu image based model; Xception (Extreme Inception) dan caption based model; GRU (Gated Recurrent Unit). Pengujian dilakukan dengan parameter skor BLEU pada setiap model yang terbentuk dari epoch. Pengukuran skor BLEU menggunakan 4-gram yang terdiri dari skor BLEU-1, BLEU-2, BLEU-3, dan BLEU-4. Dengan proses epoch yang dilakukan sebanyak 10 kali, maka didapatkan skor BLEU-1 yang dimana nilai paling mendekati 1.0 merupakan kalimat yang sama dengan kalimat kandidat yaitu skor BLEU-1 tertinggi ada pada epoch ke-15 dengan skor BLEU-1, BLEU-2, BLEU-3, dan BLEU-4 yaitu (0.642851, 0.449463, 0.347998, 0.212118).

Kata kunci: *Xception, Gated Recurrent Unit (GRU), BLEU score, Computer Vision (ComVis), Natural Language Processing (NLP), dan Beam Search*

ABSTRACT

Image Captioning is the process of generating a textual description for a given image. By involving areas of Computer Vision to understand image content and language models from the field of Natural Language Processing (NLP) to transform understanding images into words in the correct order. Based on the existing studies, this research tries to implement the encoder-decoder architecture based on previous research on Image Captioning. The method used is image based model; Xception (Extreme Inception) and caption based models; GRU (Gated Recurrent Units). The test is carried out with the BLEU score parameter on each model that is formed from epochs. BLEU score measurement using 4-grams consisting of BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores. With the epoch process carried out 10 times, a BLEU-1 score is obtained where the value closest to 1.0 is the same sentence as the candidate sentence, namely the highest BLEU-1 score is in the 15th epoch with a BLEU-1, BLEU-2, and BLEU-2 score. BLEU-3, and BLEU-4 are (0.642851, 0.449463, 0.347998, 0.212118).

***Keywords**: Image Captioning, Xception, Gated Recurrent Unit (GRU), BLEU score, Computer Vision (ComVis), Natural Language Processing (NLP), and Beam Search*

1. PENDAHULUAN

1.1. LATAR BELAKANG

Image Captioning merupakan area penelitian yang aktif hingga sekarang, banyaknya pengaplikasian digunakan pada bermacam bidang seperti judul pada gambar berita, deskripsi pada gambar medis, temu balik citra berbasis teks, membantu penyandang tunanetra dalam melakukan tugas sehari-hari dan aplikasi lainnya (Hrga & Ivašic-Kos, 2019).

Teknik-teknik yang digunakan pada Image Captioning ada beberapa seperti template-based, retrieval-based, dan novel caption generation / deep learning based. Pada teknik template-based terdapat beberapa template tetap dengan slot kosong untuk menghasilkan kalimat deskriptif. Kelemahan dari teknik ini tidak bisa menghasilkan variabel panjang kalimat, dan penggunaan template predefined (Singh, 2020). Teknik retrieval-based merupakan teknik yang menggunakan temu balik pembangkit kalimat pada set kalimat deskriptif yang ada pada dataset (Zakir Hossain et al., 2019).

Penelitian ini menggunakan dataset Flickr 8k lalu menggunakan hasil evaluasi Bilingual Evaluation Understudy (BLEU), sebuah algoritma yang menilai kualitas teks yang dihasilkan pada output oleh model. Output dari BLEU selalu antara 0 dan 1. Nilai yang mendekati 1 menunjukkan bahwa teks tersebut lebih analog dengan deskripsi yang dibuat oleh manusia (ground-truth).

1.2. RUMUSAN MASALAH

Berdasarkan identifikasi yang telah ditetapkan, maka muncul masalah yang akan ditemui yaitu:

1. Bagaimana cara mengimplementasikan model Xception dan GRU pada Image Captioning ?
2. Bagaimana hasil skor evaluasi metrik BLEU pada model deep learning Xception dan GRU ?

1.3. TUJUAN

Tujuan dari penelitian ini dengan membangkitkan kalimat deskriptif atau tekstual pada citra dengan implementasi model deep learning Xception dan GRU pada Image Captioning, untuk mengukur hasilnya, penelitian dinilai dengan evaluasi skor BLEU. Hasil yang diperoleh sebagai pembanding dari penelitian yang dilakukan pada model – model sebelumnya dan menjadi referensi untuk penelitian Image Captioning lainnya.

1.4. RUANG LINGKUP

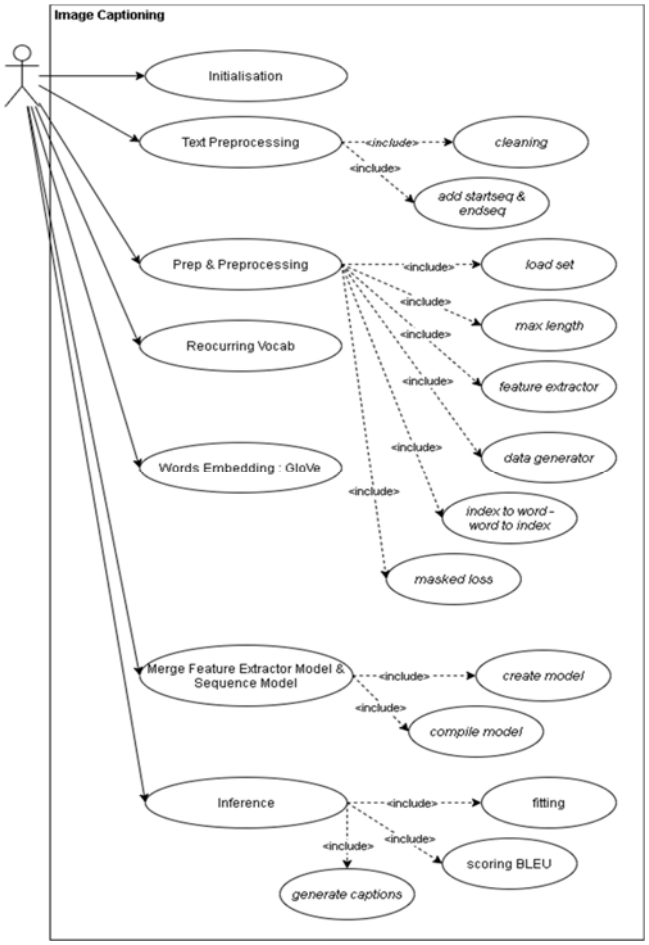
Dalam penelitian yang dilakukan, dibatasi ruang lingkup yang akan dibahas yaitu sebagai berikut :

1. Implementasi arsitektur encoder-decoder dengan model Xception dan GRU.
2. Flickr 8k dataset sebagai data latih, uji, dan validasi.
3. BLEU (Bilingual Evaluation Understudy) sebagai evaluasi metrik dengan n-grams.
4. Google Collaboratory sebagai executable document. Dengan GPU : NVIDIA Tesla K80.
5. Menggunakan library Keras.
6. Menggunakan bahasa pemrograman Python 3.6. 9.
7. Sistem berbasis website.

2. METODOLOGI PENELITIAN

2.1. PERANCANGAN UMUM

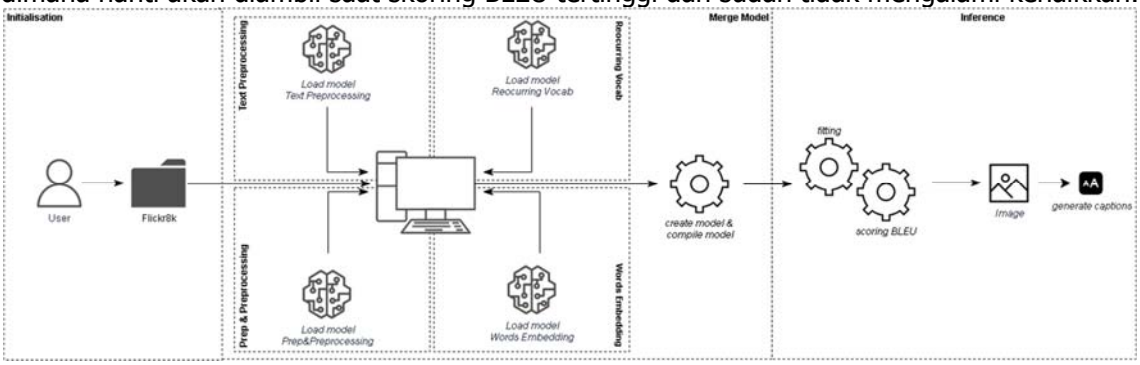
Gambaran umum di aplikasi ini diinterpretasikan melalui *use case*, terdapat *use case* diagram yang menunjukan interaksi yang bisa dilakukan oleh *user* pada sistem berdasarkan fungsionalitasnya. Dalam implementasi *Image Captioning* dengan menggunakan model CNN dan RNN yaitu Xception dan GRU, diperlukan sebuah model yang dapat menggabungkan kedua hal ini.



Gambar 1. Use Case Diagram

Konsep yang digunakan yaitu *merge model*, sebuah model yang mengkombinasikan *input* gambar dan teks deskripsi secara *encoded*. Kemudian kombinasi dari kedua *encoded* tersebut akan tersimpan sebagai model *decoder*. Dalam penelitian ini, untuk memperkaya kata atau kalimat deskriptif yang dihasilkan nanti, model menggunakan *class* pendekatan pada representasi kata yaitu GloVe. Kemudian dengan penggunaan *decoder* dalam membangkitkan kalimat deskriptif yang lebih mendekati pengucapan manusia, digunakan *Greedy search* dan *Beam search decoder*. Selanjutnya model terlebih dilatih secara *Generator* yang dapat dilatih

dalam *batch*. Model akan secara perlahan mengalami kenaikkan dalam skoring BLEU yang dimana nanti akan diambil saat skoring BLEU tertinggi dan sudah tidak mengalami kenaikkan.

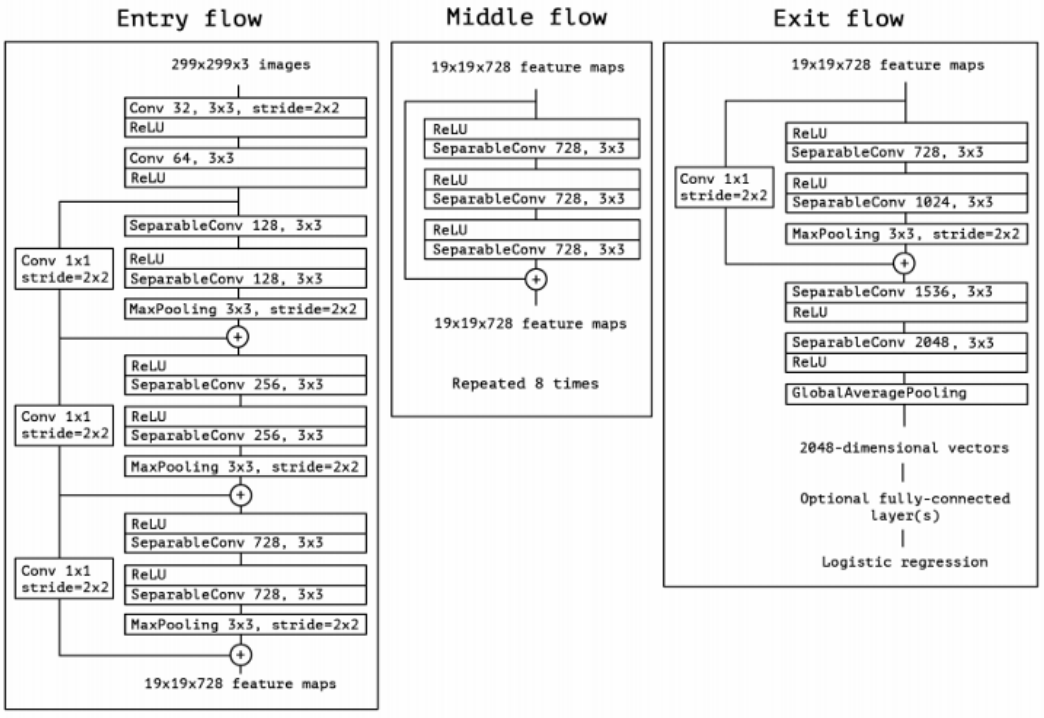


Gambar 2. Proses Bisnis Sistem

1. Proses bisnis bagian initialization yaitu, mempersiapkan dataset Flickr 8k berupa Flickr8k_Dataset yang berisi seluruh citra (.jpg) dan Flickr8k_Text yang berisi (Flickr8.token.txt,Flickr8k.testImages.txt, Flickr8k.trainImages.txt,dan Flickr8k.devImages.txt
2. Proses bisnis bagian text preprocessing yaitu membuat proses fungsi cleaning, membuat proses fungsi add startseq dan endseq
3. Proses bisnis bagian prep & preprocessing yaitu membuat proses fungsi load set, membuat proses fungsi max length, membuat proses fungsi feature extractor, membuat proses fungsi data generator, membuat proses fungsi index to word – word to index, membuat proses fungsi masked loss
4. Proses bisnis bagian reoccurring vocab yaitu membuat proses fungsi reoccurring, mencari ukuran vocab size
5. Proses bisnis bagian words embedding : GloVe yaitu , memuat GloVe, membuat embedding layer
6. Proses bisnis bagian merge model yaitu, membuat proses fungsi create model, melakukan compile model
7. Proses bisnis bagian inference yaitu melakukan fitting, mengukur model scoring BLEU melakukan proses generate caption

2.2. XCEPTION

Xception – Extreme Inception pertama kali dikenalkan dari pengembangan model Inception oleh F. Chollet (Chollet, 2017). Inception merupakan model arsitektur Convolutional Neural Network dengan mempelajari filtrasi dalam ruang 3 dimensi; 2 dimensi spasial (lebar dan tinggi) dan saluran dimensi. Sebuah convolutional kernel bertugas untuk korelasi pemetaan lintas saluran dan korelasi spasial. Bertujuan agar proses menjadi lebih mudah dan efisien dengan membagikannya secara factorial menjadi sekumpulan operasi yang secara independen mengawasi cross-chanel correlations dan spatial correlations.



Gambar 2. Xception

2.3. GRU

Dalam RNN (Recurrent Neural Network) munculnya masalah seperti short-term memory dan vanishing gradient point, dimana dalam sebuah sekuensial data yang panjang RNN menghadapi masalah untuk menyampaikan informasi yang didapatkan pada urutan awal hingga akhir. RNN didesain untuk mengerjakan tugas sekuensial data, dimana data tersebut merupakan teks, audio, video dan lainnya. (Phi, 2018) Sehingga untuk melakukan prediksi untuk teks dalam paragraf, kemungkinan adanya informasi penting yang terlewat dalam RNN. Untuk kelemahan vanishing gradient point pada RNN dalam backpropagation, dimana nilai gradien yang digunakan untuk menjadi pembaharuan bobost neural network menjadi menyusut selama proses backpropagation jika bobot semakin kecil maka tidak memberikan dampak saat melatih data.

Pada *update gate*, menghitung formula *update_gate z for time step t* menggunakan rumus :

$$z_t = \sigma(W^{(z)}X_t + U^{(z)}h_{t-1})$$

dimana saat X_t dihubungkan ke dalam unit jaringan, akan dikalikan dengan bobot nya ($W^{(z)}$). Sementara untuk (h_{t-1}) yang digunakan untuk menyimpan informasi pada waktu yang sebelumnya *t*-1 akan dikalikan bersama bobotnya ($U^{(z)}$). Kedua nilai akan dijumlahkan dan dengan fungsi aktivasi sigmoid untuk mengubah nilai antara 0 dan 1.

Pada *reset gate* menentukan jumlah informasi dari tahapan sebelumnya untuk dieliminasi. Menggunakan formula :

$$r_t = \sigma(W^{(r)}X_t + U^{(r)}h_{t-1}),$$

formula ini sama seperti yang digunakan pada *update gate*. Perbedaannya terletak pada bobot dan penggunaan *gate*, seperti tahapan yang sama, dengan memasukkan h_{t-1} dan X_t dikalikan dengan bobotnya, dan dijumlahkan kemudian diterapkan fungsi sigmoid.

Hasil dari kedua vektor tersebut (*reset gates* & *update gates*) akan mempengaruhi nilai *output* pada GRU, sebelumnya pada *reset gate* adanya formula untuk menyimpan konten informasi yang akan digunakan yaitu

$$h'_t = tanh(WX_t + r_t \odot Uh_{t-1})$$

1. Mengalikan *input* X_t dengan bobot *W* dan h_{t-1} dengan bobot *U*.
2. Menghitung fungsi *Hadamard* antara *reset gate* dan Uh_{t-1}, yang dimana akan menentukan apa yang harus dieliminasi pada tahapan sebelumnya.
3. Jumlahkan tahapan (1) dan (2).
4. Menerapkan fungsi aktivasi nonlinear *tanh*.

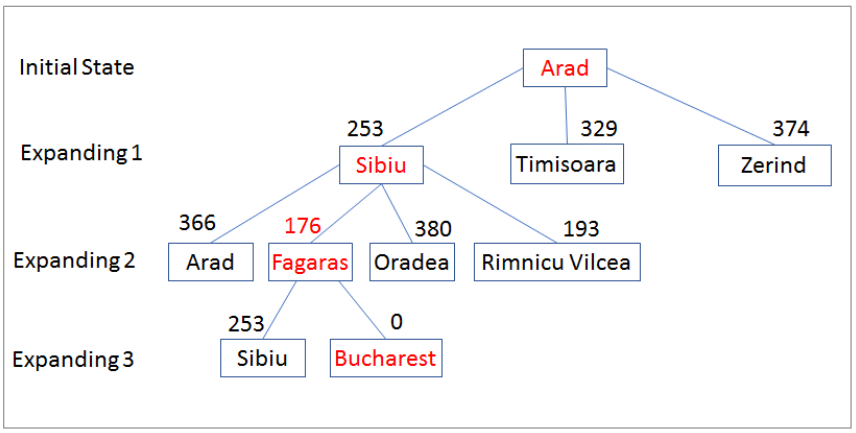
2.4. WORDS EMBEDDING (GLOVE)

Word Embedding merupakan sebuah kelas pendekatan untuk merepresentasikan teks atau kata yang memiliki makna yang serupa, dengan dense vector representation. Class ini merupakan improvisasi dari Bag of Words (BoW) yang lebih tradisional, dimana sparse vector besar digunakan untuk mewakili setiap kata atau untuk menilai setiap kata dalam vektor untuk mewakili seluruh kosakata. Dalam embedding, kata direpresentasikan oleh dense vector dimana vektor merepresentasikan proyeksi dari kata ke dalam continuous vector space. Posisi dari kata dalam vector space dipelajari oleh teks dan berdasarkan kata yang meliputi kata tersebut. Dimana posisi kata dalam pembelajaran vector space disebut sebagai embedding.

Global Vector (GloVe) merupakan salah satu teknik pada word embedding yang digunakan untuk merepresentasikan kata. Berupa algoritma unsupervised learning yang dikembangkan oleh peneliti untuk menghasilkan word embedding dengan agregasi kata global co-occurrence matrices dari korpus kata. GloVe menangkap kedua statistik global dan statistik local dari korpus, untuk memunculkan word vectors

2.5. GREEDY SEARCH

Greedy Search merupakan sebuah pendekatan sederhana dengan menggunakan pencarian keseluruhan kata yang paling mungkin atau probabilitas tertinggi pada setiap langkah dalam urutan *output*. Pendekatan ini memiliki kelebihan yang sangat cepat, tetapi kualitas urutan *output* akhir mungkin jauh dari optimal. Dengan memilih hanya satu kandidat pada satu langkah mungkin optimal di beberapa urutan, tetapi jika melanjutkan sisa kalimat penuh, hal ini mungkin menjadi lebih buruk jika dalam menggunakan *Greedy Search*.



Gambar 3. Greedy Search

2.6. BEAM SEARCH

Beam search memilih beberapa token untuk posisi dalam urutan tertentu berdasarkan probabilitas bersyarat. Algoritma dapat mengambil sejumlah N alternatif terbaik melalui *hyperparameter* yang dikenal sebagai *Beam width*. Dalam *Greedy* hanya mengambil kata terbaik untuk setiap posisi dalam urutan, di mana *Beam search* memperluas pencarian atau "lebar" untuk memasukkan kata-kata lain yang mungkin lebih cocok. Pencarian *Greedy* melihat setiap posisi dalam urutan output secara terpisah. Sebuah kata diputuskan berdasarkan probabilitas tertinggi dan terus melanjutkan sisa kalimat, tidak kembali ke kalimat sebelumnya. Dengan pencarian Beam, mengambil N urutan keluaran terbaik dan melihat kata-kata sebelumnya saat ini dan probabilitas dibandingkan dengan posisi saat ini yang dekode dalam urutan.

2.7. BILINGUAL EVALUATION UNDERSTUDY (BLEU)

BLEU adalah salah satu standar pengukuran terjemahan mesin evaluasi dan metrik pertama yang digunakan untuk mengevaluasi *Image Captioning*. BLEU menghitung rata-rata geometris skor presisi *n-gram* dikalikan dengan *Brevity Penalty* (nilai yang mendekati 1, menunjukkan bahwa teks tersebut lebih mendekati dengan deskripsi yang dibuat oleh manusia). Pada dasarnya BLEU merupakan algoritma yang memeriksa kualitas teks yang dihasilkan pada *output* oleh model, untuk memeriksa *output* yang dihasilkan model dan membandingkan teks yang dihasilkan dengan satu / lebih referensi.

$$P_n = \frac{C \in (\sum \quad candidate) \sum_{n-gram \in c} \quad Count_{clip}(n-gram)}{C \in (\sum \quad candidate) \sum_{n-gram' \in c'} \quad Count_{clip}(n-gram')}$$

3. HASIL DAN PEMBAHASAN

3.1. HASIL

```
Evaluating Completion: [------------------->] 100%

BLEU-1: 0.628728
BLEU-2: 0.439507
BLEU-3: 0.339994
BLEU-4: 0.205870
[0.6287278372416047,
 0.43950700183491376,
 0.3399942231621965,
 0.2058695981979836]
```

Gambar 4. Skor BLEU Pada Model

Mengukur performa model dengan BLEU scoring, dimana skor BLEU yang mendekati 1.0 yaitu kalimat pengucapan yang mendekati anotasi manusia. Dimana dalam hal ini yaitu skor BLEU-1 yaitu skor unigram BLEU dengan skor 0.628.


```
Greedy: a brown dog runs through the grass
Beam K= 3: a brown dog is running through the grass
Beam K= 5: a brown dog runs through the grass
Beam log K= 3: a brown dog runs through the grass
Beam log K= 5: a brown dog is running through the grass
```

Gambar 5. Generate Captions

Menguji model untuk membangkitkan kalimat deskriptif, kemudian setelah dilakukan fitting dengan algoritma pencarian pembangkit kalimat dengan Greedy Search kemudian Beam Search, kalimat yang ditampilkan merupakan kalimat hasil model Greedy, Beam Search k-beams=3, Beam Search k-beams=5, Beam Search log k-beams=3, dan Beam Search log k-beams=5.

3.2. PEMBAHASAN

Pengujian kinerja sistem dilakukan untuk mengukur kinerja pada model apakah *output* tersebut mampu membangkitkan kalimat yang 'baik' dimana kalimat tersebut apakah mampu mendekati kalimat dari referensi (*ground-truth*), disini model diuji dengan parameter skor BLEU pada setiap model yang tersimpan dari setiap 3 *epoch*. Pengukuran skor BLEU menggunakan *cumulative* 4-gram yang terdiri dari skor BLEU-1, BLEU-2, BLEU-3, dan BLEU-4. Disini n-gram merupakan konsep yang digunakan satu set 'n' kata berurutan dalam sebuah kalimat. Misalnya dalam kalimat "*The ball is red*", 4-gram yang didapatkan, yaitu :

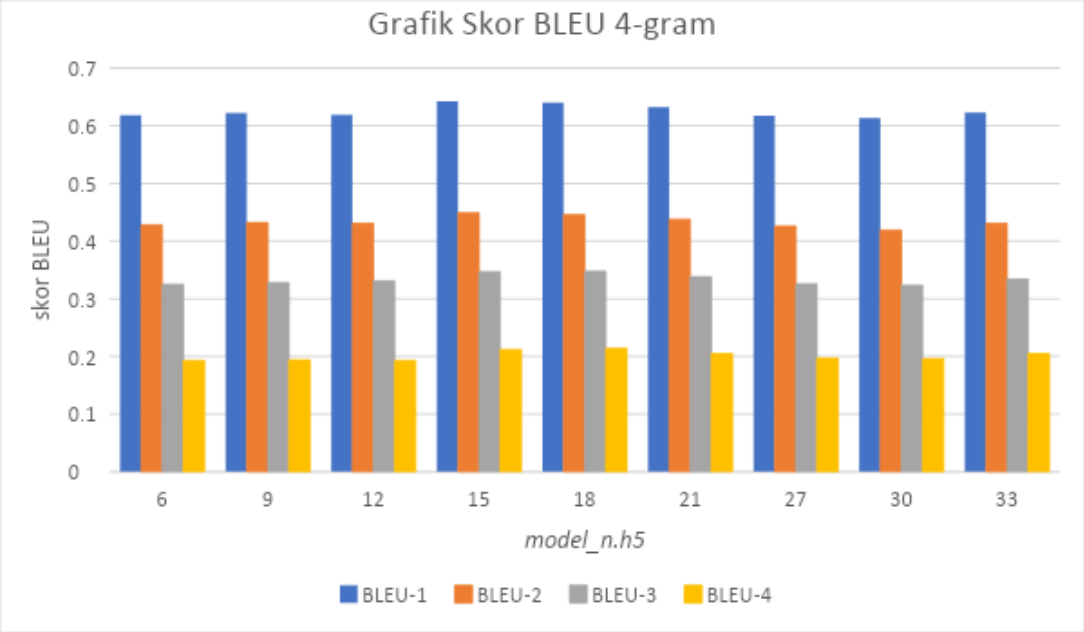
- 1-gram *(unigram) : "The", "ball", "is", "red"*
- 2-gram *(bigram) : "The ball", "ball is", "is red"*
- 3-gram *(trigram) : "The ball is", "ball is red"*
- 4-gram *: "The ball is red"*

Pada setiap individu n-gram yang didapatkan akan dihitung skor *precision* dengan formula yang ada pada (6) dalam 2.7 *Bilingual Evaluation Understudy* (BLEU) dan diimplementasikan pada program. Dengan *cumulative scores* dimana perhitungan dari setiap individual n-gram dengan bobot BLEU-4 sebesar 1/4 atau 25% untuk tiap 1-gram, 2-gram, dan 3-gram. Berikut merupakan skor BLEU-1 pada Tabel 1 yang didapatkan dari setiap model yang disimpan. Disini BLEU-1 merupakan skor yang memiliki bobot 4/4 atau bernilai 1.

Tabel 1. Skor BLEU-1 per epoch

***epoch* ke-**	**Skor BLEU-1**	**Nama Model**
6	0.618013	*model_6.h5*
9	0.622148	*model_9.h5*
12	0.619911	*model_12.h5*
15	0.642851	*model_15.h5*
18	0.640448	*model_18.h5*
21	0.632756	*model_21.h5*
24	0.627205	*model_24.h5*
27	0.617421	*model_27.h5*
30	0.613098	*model_30.h5*
33	0.623807	*model_33.h5*

Dengan model yang disimpan sebanyak 10 kali atau 33 *epoch*, maka didapatkan skor BLEU-1 dengan nilai paling mendekati 1.0, dimana model tersebut merupakan model pembangkit kalimat yang paling mendekati *ground-truth*. Model tersebut berada di epoch ke -15 atau "*model_15.h5*". Dalam proses *fitting* model mengalami kenaikkan *learning* sehingga mampu memberikan skor BLEU yang tinggi dan setiap model nilai *loss* mengalami penurunan secara berkala, disini menandakan model mengalami kenaikkan akurasi dalam melakukan pembelajaran. Untuk skor BLEU-2, BLEU-3, dan BLEU-4 pada model "*model_15.h5*" dapat dilihat pada grafik dibawah ini. Berikut merupakan grafik skor BLEU 4-gram sebanyak 10 *epoch* pada Gambar 6



Gambar 6. Grafik skor BLEU4-gram pada setiap model

Berdasarkan model yang telah disimpan dengan skor BLEU-1 sebesar 0.642851. Peneliti dapat menilai bahwa kalimat yang dihasilkan memiliki kelayakan untuk melakukan *Image Captioning*. Dalam beberapa tinjauan pustaka yang digunakan, skor BLEU-1 yang berada di antara nilai 0.6 dan 0.7 memiliki kualitas yang memadai untuk melakukan pembangkitan kalimat deskriptif.

4. KESIMPULAN

Pada penelitian ini telah diimplementasikan model Xception dan GRU dalam Image Captioning sebagai model encoder-decoder yang berfungsi untuk mengubah suatu fitur dari citra menjadi kata-kata. Dengan menggunakan dataset Flickr 8k, yang terdiri dari 1000 data uji, 6000 data latih, dan 1000 data validasi. Model melakukan ekstraksi fitur dalam citra data latih beserta kalimat referensinya, yang kemudian disimpan sebagai sebuah fitur dengan one hot encoding vector. Sehingga vektor tersebut dapat diterima oleh model bahasa yang dimana dalam arsitekturnya merupakan sebuah sequence model. Model ini akan memproses setiap nilai vektor yang kemudian diubah menjadi sebuah kata-kata. Sebelumnya untuk merepresentasikan sebuah vektor menjadi kata diperlukan embedding layer, disini penelitian menggunakan pre-trained words representation yaitu GloVe. Sehingga fase training ataupun learning tidak dibebankan untuk mempelajari representasi vektor lagi menjadi kata.

Kemudian dalam proses pencarian kata ataupun prediksi, teknik yang digunakan dalam penelitian ini dibantu menggunakan Greedy Search dan Beam Search yang mampu memberikan bobot pada setiap kata untuk mencari nilai probabilitas yang tertinggi ataupun mencari kandidat yang terbaik. Dalam teknik Beam Search penggunaan beam width mampu memberikan jumlah kalimat yang terbaik berdasarkan n-beam width. Penelitian ini menggunakan beamwidth 3 dan 5 yang kemudian akan diambil terbaiknya. Untuk setiap beam width akan menggunakan kondisi dimana log = True ataupun False, dimana nilai indeks kata akan ditambahkan atau tidak, jadi teknik ini akan mengambil jumlah skor dari indeks dan probabilitas, atau hanya probabilitas saja. Dengan ini untuk melihat kinerja pada model yang telah dibangun, dan menguji model tersebut. Model akan diukur dengan menggunakan parameter skor BLEU-4. Dalam skor BLEU-4 terdiri dari skor BLEU-1, BLEU-2, dan BLEU-4

memiliki nilai cumulative score dengan bobot yang berbeda. Sebelumnya model telah disimpan melalui proses pelatihan dan penyesuaian terhadap data uji. Model diuji dengan parameter skor BLEU yang dimana nilai paling mendekati 1.0 maka model mampu menghasilkan kalimat deskriptif yang serupa dengan kalimat kandidat. Hingga akhirnya model ini mencapai dengan nilai rata-rata skor BLEU-1, BLEU-2, BLEU-3, dan BLEU-4 (0.642, 0.449, 0.347, 0.212). Dengan ini dapat disimpulkan bahwa model memiliki kelayakan untuk melakukan image captioning ditinjau dari skor BLEU-1 yang memiliki nilai 0.642, dan cukup 'baik' dalam membangkitkan kalimat deskriptif pada sebuah citra.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Prodi Informatika Institut Teknologi Nasional Bandung. Penulis juga berterima kasih kepada semua pihak yang telah mendukung sehingga penelitian ini terlaksana dengan baik.

DAFTAR RUJUKAN

Brownlee, J. (2018, January 1). *Encoder-Decoder Recurrent Neural Network Models for Neural Machine Translation*. Retrieved from https://machinelearningmastery.com/: https://machinelearningmastery.com/encoder-decoder-recurrent-neural-network-models-neural-machine-translation/

Chowdhry, A. (2021, June 30). *Image Caption Generator*. Retrieved from blog.clairvoyantsoft.com: https://blog.clairvoyantsoft.com/image-caption-generator-535b8e9a66ac

IBM Cloud Education. (2020, May 1). *www.ibm.com/cloud/learn/deep-learning*. Retrieved from www.ibm.com: https://www.ibm.com/cloud/learn/deep-learning

Mori, Y., Takahashi, H., & Oka, R. (1999). Image-to-word transformation based on dividing and vector. 1-9.

Phi, M. (2018, September 25). *Illustrated Guide to LSTM's and GRU's: A step by step explanation*. Retrieved from https://towardsdatascience.com/: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

Roy, A. (2020, December 9). *A Guide to Image Captioning*. Retrieved from towardsdatascience.com: https://towardsdatascience.com/a-guide-to-image-captioning-e9fd5517f350

San Pa Pa Aung†, W. P. (2020). Automatic Myanmar Image Captioning using CNN and LSTM-Based Language. 139-43.

Singh, N. P. (2020). *Automatic Image Annotation / Image Captioning*. Retrieved from https://iq.opengenus.org/: https://iq.opengenus.org/automatic-image-annotation/

Utami, S. N. (2021, July 05). *Artificial Intelligence (AI): Pengertian, Perkembangan, Cara Kerja, dan Dampaknya*. Retrieved from kompas.com: https://www.kompas.com/skola/read/2021/07/05/121323869/artificial-intelligence-ai-pengertian-perkembangan-cara-kerja-dan?page=all

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11211 LNCS*, 833–851. https://doi.org/10.1007/978-3-030-01234-2_49

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 1724–1734. https://doi.org/10.3115/v1/d14-1179

Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, *2017-Janua*, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

D., B., Krishna, K. C., K., T., Vikas, N. V., & Sahithya, A. N. V. (2021). *Image Captioning Using Deep Learning*. 381–395. https://doi.org/10.4018/978-1-7998-6870-5.ch026

Hrga, I., & Ivašic-Kos, M. (2019). Deep Image Captioning: An overview. *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2019 - Proceedings*, 995–1000. https://doi.org/10.23919/MIPRO.2019.8756821

Nugraha, A. A., Arifianto, A., & Suyanto. (2019). Generating Image Description on Indonesian Language using Convolutional Neural Network and Gated Recurrent Unit. *2019 7th International Conference on Information and Communication Technology, ICoICT 2019*, 1–6. https://doi.org/10.1109/ICoICT.2019.8835370

Nursikuwagus, A., Munir, R., & Khodra, M. L. (2020). Image Captioning menurut Scientific Revolution Kuhn dan Popper. *Jurnal Manajemen Informatika (JAMIKA)*, *10*(2), 110–121. https://doi.org/10.34010/jamika.v10i2.2630

Pa Pa Aung, S., Pa Pa, W., & Nwe, T. L. (2020). Automatic Myanmar Image Captioning using CNN and LSTM-Based Language Model. *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, *May*, 139–143. https://www.aclweb.org/anthology/2020.sltu-1.19

Parikh, H., Sawant, H., Parmar, B., Shah, R., Chapaneri, S., & Jayaswal, D. (2020). Encoder-Decoder Architecture for Image Caption Generation. *2020 3rd International Conference on Communication Systems, Computing and IT Applications, CSCITA 2020 - Proceedings*, 174–179. https://doi.org/10.1109/CSCITA47329.2020.9137802

Shrimal, A., & Chakraborty, T. (2020). *Attention Beam: An Image Captioning Approach*. https://doi.org/10.1109/CVPR.2015.7298932.Lin

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

Yang, S., Yu, X., & Zhou, Y. (2020). LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example. *Proceedings - 2020 International Workshop on Electronic Communication and Artificial Intelligence, IWECAI 2020*, 98–101. https://doi.org/10.1109/IWECAI50956.2020.00027

Zakir Hossain, M. D., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Computing Surveys*, *51*(6). https://doi.org/10.1145/3295748

Arnav, A., Hangkyu, J., & Maloo, P. (2021). *Image Captioning Using Deep Learning*. 381–395. https://doi.org/10.4018/978-1-7998-6870-5.ch026

Liu, S., Bai, L., Hu, Y., & Wang, H. (2018). Image Captioning Based on Deep Neural Networks. *MATEC Web of Conferences*, *232*, 1–7. https://doi.org/10.1051/matecconf/201823201052

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

https://doi.org/10.1109/PARC49193.2020.236619