Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2573−2581
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6958

Penerapan CNN dan RNN untuk Pembuatan Deskripsi Konten Visual Menggunakan Deep Learning

Aldy Agil Hermanto*, Giat Karyono, Imam Tahyudin

Program Studi Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Amikom Purwokerto, Purwokerto, Indonesia
Email: [1,*]aldyagilh@gmail.com, [2]giatkaryono@amikompurwokerto.ac.id, [3]imam.tahyudin@amikompurwokerto.ac.id
Email Penulis Korespondensi: aldyagilh@gmail.com
Submitted: **09/02/2025**; Accepted: **15/02/2025**; Published**: 16/02/2025**

Abstrak−Perkembangan teknologi di bidang pengolahan citra dan suara telah membawa dampak signifikan dalam meningkatkan aksesibilitas informasi bagi berbagai kalangan, terutama bagi individu dengan gangguan penglihatan. Salah satu inovasi yang muncul adalah sistem image to speech, yang memungkinkan konversi gambar menjadi suara yang dapat dimengerti oleh penggunanya. Permasalahan utama terletak pada rendahnya akurasi pengenalan objek dalam gambar dengan variabilitas tinggi, seperti pencahayaan buruk atau latar belakang kompleks, serta tantangan dalam menghasilkan deskripsi teks yang sesuai untuk dikonversi menjadi audio. Metode yang digunakan melibatkan ekstraksi fitur gambar menggunakan CNN berbasis InceptionV3 dan pembentukan urutan teks deskriptif melalui RNN dengan mekanisme attention . Dataset terdiri dari 40.455 keterangan dan 8.091 gambar, diproses dengan teknik pra-pemrosesan teks dan gambar sebelum dilatih menggunakan teknik teacher forcing . Hasil evaluasi menunjukkan skor BLEU sangat rendah (5.154827976372712e-153), menandakan ketidakmampuan model mereplikasi caption asli dengan baik. Meskipun demikian, audio hasil konversi teks ke suara menggunakan Google Text-to-Speech cukup jelas. Solusi ke depan mencakup peningkatan dataset, penerapan regularisasi, serta penyesuaian arsitektur model untuk meningkatkan akurasi prediksi caption dan relevansi audio terhadap gambar. Dengan perbaikan ini, diharapkan sistem dapat memberikan aksesibilitas informasi visual yang lebih inklusif bagi individu dengan gangguan penglihatan.

Kata Kunci: CNN; Image To Speech; Penyandang Disabilitas Penglihatan; Skor BLEU; Pengolahan Citra

Abstract−The development of technology in the field of image and sound processing has had a significant impact on increasing the accessibility of information for various groups, especially for individuals with visual impairments. One of the innovations that emerged was the image to speech system, which allows the conversion of images into sounds that can be understood by its users. The main problem lies in the low accuracy of object recognition in images with high variability, such as poor lighting or complex backgrounds, as well as the challenge of producing suitable text descriptions to be converted into audio. The method used involves extracting image features using InceptionV3-based CNN and forming a sequence of descriptive texts through RNN with an attention mechanism. The dataset consists of 40,455 captions and 8,091 images, processed using text and image pre-processing techniques before being trained using the teacher forcing technique. The evaluation results show a very low BLEU score (5.154827976372712e-153), indicating the model's inability to replicate the original caption well. However, the audio from the text-to-speech conversion using Google Text-to-Speech is quite clear. Future solutions include increasing the dataset, applying regularization, and adjusting the model architecture to improve the accuracy of caption prediction and audio relevance to the image. With these improvements, it is hoped that the system can provide more inclusive visual information accessibility for individuals with visual impairments.

Keywords: CNN; Image To Speech; Visually Disabled; BLEU Score; Image Processing

1. PENDAHULUAN

Perkembangan teknologi di bidang pengolahan citra dan suara semakin pesat, memberikan peluang untuk menciptakan berbagai aplikasi yang mempermudah kehidupan manusia. Salah satu aplikasi yang semakin populer adalah pengkonversian gambar atau citra menjadi bentuk suara yang dapat dimengerti oleh pengguna. Teknologi ini dikenal dengan istilah image to speech (gambar ke suara), yang memberikan solusi bagi individu dengan gangguan penglihatan atau bagi mereka yang membutuhkan aksesibilitas informasi melalui media suara. Sebagai contoh, pengaplikasian image to speech dapat membantu pembaca tunanetra untuk memahami konten gambar dalam buku atau media lainnya secara lebih inklusif.

Namun, meskipun ada banyak kemajuan dalam bidang ini, tantangan besar masih ada dalam hal akurasi dan efisiensi dalam mengenali objek atau teks dari gambar dan mengonversinya menjadi bentuk suara yang jelas dan dapat dipahami. Salah satu permasalahan utama adalah pengenalan objek dalam gambar yang memerlukan analisis mendalam tentang konten gambar tersebut. Sebagian besar metode tradisional dalam pengolahan citra terbukti kurang efektif dalam menangani variabilitas gambar, seperti pencahayaan yang buruk, objek yang kabur, atau latar belakang yang kompleks. Hal ini mengarah pada kebutuhan untuk metode yang lebih robust dan efisien.

Untuk mengatasi permasalahan tersebut, metode Convolutional Neural Networks (CNN) dipilih sebagai solusi utama dalam pengembangan sistem image to speech. CNN memiliki keunggulan dalam mengidentifikasi fitur-fitur penting dalam gambar melalui proses ekstraksi fitur yang mendalam, yang memungkinkan sistem untuk mengenali objek dan teks dengan akurat [1]. Dalam konteks ini, CNN sangat cocok karena kemampuannya untuk belajar dari data visual yang besar dan mengatasi kompleksitas yang ada pada gambar [2][3]. Untuk melatih model CNN ini, digunakan dataset yang diperoleh dari platform Kaggle, yang menyediakan berbagai dataset gambar yang dapat digunakan untuk mengembangkan model pengenalan gambar dengan akurasi tinggi.

Dalam penelitian yang dilakukan oleh Inggis Kurnia Trisiawan, Yuliza, Fina Supegina, dan Said Attamimi, penerapan CNN untuk klasifikasi botol minuman menghasilkan akurasi 95,02% dengan loss 0,1064, dan akurasi

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2573−2581
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6958

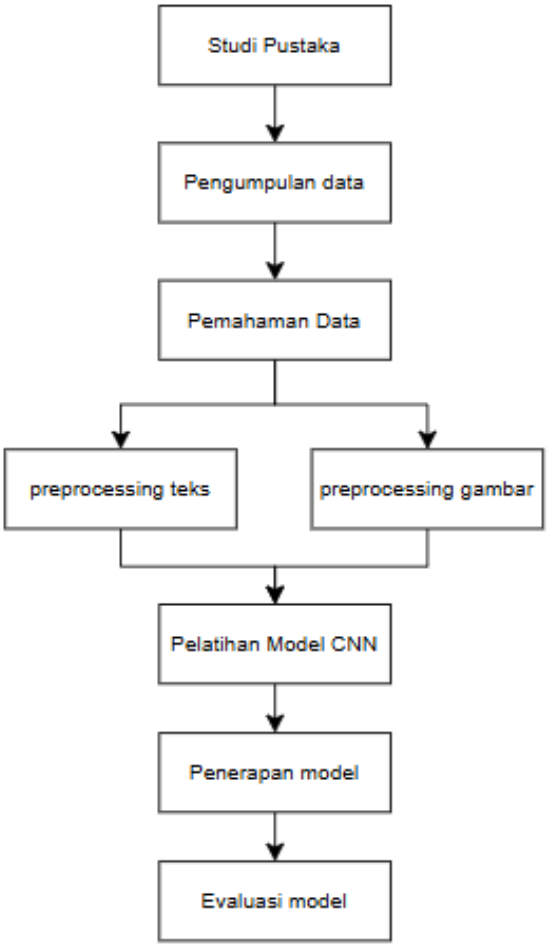
98,526% pada dataset baru[4]. Farah Afi Febriyanti menunjukkan bahwa CNN efektif untuk deteksi penyakit kulit dengan akurasi validasi 96%, meskipun masih perlu peningkatan lebih lanjut pada dataset dan arsitektur model [5]. Ego Oktafanda menemukan bahwa CNN dengan transfer learning ResNet50 menghasilkan akurasi 95% untuk klasifikasi citra kualitas bibit kelapa sawit, dengan dataset kecil memengaruhi performa model [6]. Yoke Annisa Putri Vandalis, Sopian Soim, dan Lindawati Lindawati mengembangkan CNN dengan penambahan layer konvolusi yang berhasil mencapai akurasi 88% pada pelatihan dan 83% pada pengujian untuk klasifikasi sampah [7]. Sementara itu, Ar'rafi Akram, Kun Fayakun, dan Harry Ramza berhasil mencapai akurasi 93,81% pada training dan 80% pada pengujian dalam klasifikasi hama serangga untuk pengendalian hama pertanian [8].

Recurrent Neural Network(RNN) adalah sebuah metodeDeep Learningyang dapat melakukan pembelajaran terhadap pola yang terdapat di dalam data sekuensial. Recurrent Neural Network(RNN) memiliki kemampuan untuk "mengingat" elemen data yang telah "dipelajari" sebelumya [9]. RNN bertugas untuk memproses fitur gambar yang dihasilkan oleh CNN dan mengubahnya menjadi urutan teks atau deskripsi suara [10]. RNN sangat cocok untuk tugas ini karena kemampuannya dalam memproses data sekuensial, seperti urutan kata dalam kalimat [11]. Solusi yang diusulkan adalah pengembangan sistem image to speech berbasis CNN:RNN yang dapat mengenali berbagai elemen dalam gambar dan mengubahnya menjadi suara yang dapat dimengerti. Tujuan dari penelitian ini adalah untuk meningkatkan kemampuan sistem image to speech dengan menggunakan metode CNN dan dataset dari Kaggle, dengan fokus pada peningkatan akurasi pengenalan objek dan teks serta kejelasan dalam konversi suara. Keberhasilan sistem akan diukur menggunakan metrik akurasi pengenalan gambar, rasio kesalahan (loss), serta kualitas dan ketepatan deskripsi suara yang dihasilkan, yang akan diuji dengan evaluasi subjektif oleh pengguna, khususnya individu dengan gangguan penglihatan. Dengan solusi ini, diharapkan dapat tercipta sistem yang lebih efisien dan akurat dalam mengonversi gambar menjadi suara, serta memberikan akses yang lebih baik bagi individu dengan gangguan penglihatan.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian pada Gambar 1 bertujuan untuk menghasilkan model berbasis Convolutional Neural Network (CNN) yang mampu melakukan tugas tertentu dengan memanfaatkan data berupa teks dan gambar.

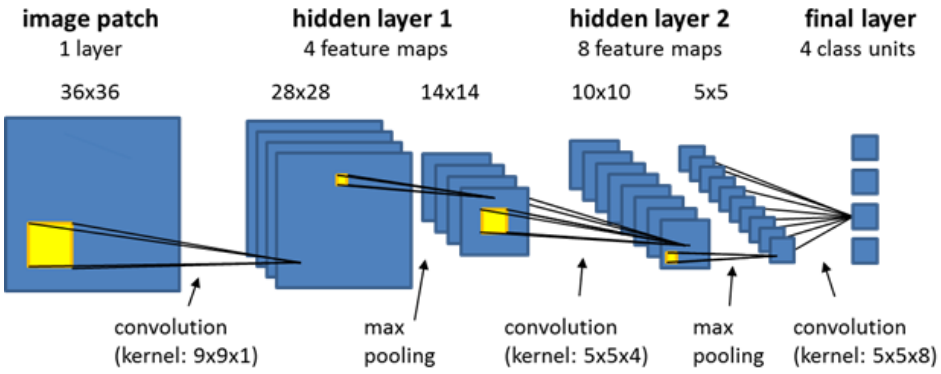


Gambar 1. Tahapan Penelitian

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2573−2581
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6958

Metode yang digunakan dalam penelitian ini adalah CNN yang merupakansalah satu algoritma dari deep learningyang banyak digunakan dengan data citrayang dapat mengenali objek-objek pada suatu citra [12]. Berikut adalah tahapan yang dilibatkan dalam penelitian ini:

a. Studi Pustaka
   Pada tahap ini peneliti mengumpulkan artikel terkait penelitian sebelumnya mengenai klasifikasi objek dari jurnal, proseding, buku, internet dan lain –lain yang mendukung penelitian ini [13].
b. Pengumpulan Data
   Agar CNN dapat melakukan klasifikasi objek maka diperlukan data pendahuluan yang diperoleh dari Kaggle Dataset [13]. Dataset yang digunakan adalah "Flickr 8k Dataset" yang terdiri dari 8.000 gambar dengan lima deskripsi berbeda per gambar . Gambar dipilih dari enam grup Flickr yang menggambarkan berbagai adegan dan situasi. Data dibagi dengan rasio 80% untuk pelatihan dan 20% untuk pengujian untuk memastikan model dapat belajar secara efektif dan diuji dengan data yang belum pernah dilihat sebelumnya.
c. Preprocessing Data
   Sebelum data dianalisis perlu dilakukan preprocesing data, yang mempunyai tujuan yaitu untuk melihat karakteristik data. Karakteristik data adalah gambaran umum bagaimana komputer membaca sebuah gambar menjadi array yang bermakna untuk proses selanjutnya seperti adanya efek spasial dalam data gambar [14]. Data preprocessing merupakan tahapan dimana data akan dilakukan pengisian data yang kosong, menghilangkan duplikasi data, memeriksa inkonsistensi data, pembersihan data serta memperbaiki kesalahan pada data [15].
d. Pelatihan Model
   CNN adalah pengembangan dari Multilayer Perceptron (MLP) yang didesain untuk mengolah data dua dimensi. CNN termasuk dalam jenis Deep Neural Network karena kedalaman jaringan yang tinggi dan banyak diaplikasikan pada data citra [16][17]. CNN memiliki kemampuan untuk mengidentifikasi pola hirarki pada data dan mengintegrasikan piksel yang lebih kompleks daripada piksel yang lebih kecil dan sederhana [18]. Namun, untuk mengonversi gambar menjadi suara, CNN hanya bertanggung jawab dalam tahap ekstraksi fitur gambar. Arsitektur CNN yang digunakan pada penelitian ini yaitu arsitektur Alexnet dapat dilihat pada Gambar 2.



Gambar 2. Penerapan CNN

Selanjutnya, Tugas dari RNN sendiri adalah menemukan hubungan antara inputsaat ini dan input yang diterapkan sebelumnya [19]. Recurrent Neural Network (RNN) diterapkan sebagai decoder untuk menghasilkan urutan kata deskripsi berdasarkan fitur gambar yang diperoleh dari CNN. Encoder menghasilkan fitur gambar dengan bentuk, sementara decoder, yang merupakan bagian dari RNN, memprediksi kata demi kata hingga mencapai kata penutup *caption*.

e. Penerapan Model
   Pada tahap ini akan ditampilkan dataset yang digunakan apakah kode yang digunakan sudah berhasil atau belum. Untuk data gambar yang disertai teks, gambar dapat diproses menggunakan CNN untuk ekstraksi fitur, sementara teks diproses menggunakan RNN. Training dilakukan untuk memberikan pengetahuan tentang arsitektur Convolutional Neural Network (CNN) yang dibangun [20]. Dalam tahap ini, data teks dan gambar yang telah diproses digunakan untuk mengajarkan model mengenali pola tertentu. Model dilatih menggunakan teknik machine learning dengan optimasi parameter.
f. Evaluasi model
   Tahap ini dilakukan menggunakan pencarian greedy untuk menghasilkan prediksi audio secara efisien, dan pencarian berkas (beam search) secara opsional untuk meningkatkan akurasi dengan mempertimbangkan konteks yang lebih luas. Hasil model diuji pada data sampel menggunakan skor BLEU untuk membandingkan transkripsi audio dengan teks referensi. Evaluasi juga mencakup penilaian kualitas audio dan kesesuaian output dengan masukan.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2573−2581
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6958

3. HASIL DAN PEMBAHASAN

3.1 Hasil

3.1.1 Pemahaman Data

Tahap ini dimulai dengan mengimpor dataset dan membaca data berupa gambar serta keterangan (caption) ke dalam dua variabel yang terpisah. Dataset terdiri atas 40.455 keterangan dan 8.091 gambar. Hasil pemahaman dataset gambar serta teks dapat dilihat pada Gambar 3.

```
print("Total captions present in the dataset: "+ str(len(annotations)))
print("Total images present in the dataset: " + str(len(all_img_path)))

Total captions present in the dataset: 40455
Total images present in the dataset: 8091
```

Gambar 3. Hasil Pemahaman Data

Visualisasi gambar-gambar serta teks keterangan dilakukan untuk memahami struktur dataset. Selanjutnya, sebuah kerangka data dibangun untuk merangkum informasi berupa gambar, jalur gambar, dan keterangan yang terorganisir dalam format dataframe.

Gambar 4. Visualisasi Dataset Gambar

Untuk memudahkan proses lebih lanjut, dibuat pula daftar yang mencakup semua keterangan dan jalur gambar.

```
df=doc;
df["Path"]=doc.ID.apply(lambda x: "/content/drive/MyDrive/Flickr/Dataset/Flicker8k_Dataset/"+x)

df.head()
```

	ID	Captions	Path
0	1000268201_693b08cb0e.jpg	A child in a pink dress is climbing up a set of stairs in an entry way .	/content/drive/MyDrive/Flickr/Dataset/Flicker8k_Dataset/1000268201_693b08cb0e.jpg
1	1000268201_693b08cb0e.jpg	A girl going into a wooden building .	/content/drive/MyDrive/Flickr/Dataset/Flicker8k_Dataset/1000268201_693b08cb0e.jpg
2	1000268201_693b08cb0e.jpg	A little girl climbing into a wooden playhouse .	/content/drive/MyDrive/Flickr/Dataset/Flicker8k_Dataset/1000268201_693b08cb0e.jpg
3	1000268201_693b08cb0e.jpg	A little girl climbing the stairs to her playhouse .	/content/drive/MyDrive/Flickr/Dataset/Flicker8k_Dataset/1000268201_693b08cb0e.jpg
4	1000268201_693b08cb0e.jpg	A little girl in a pink dress going into a wooden cabin .	/content/drive/MyDrive/Flickr/Dataset/Flicker8k_Dataset/1000268201_693b08cb0e.jpg

Gambar 5. Visualisasi Dataset Dalam Bentuk Teks

3.1.2 Pra-Pemrosesan Teks

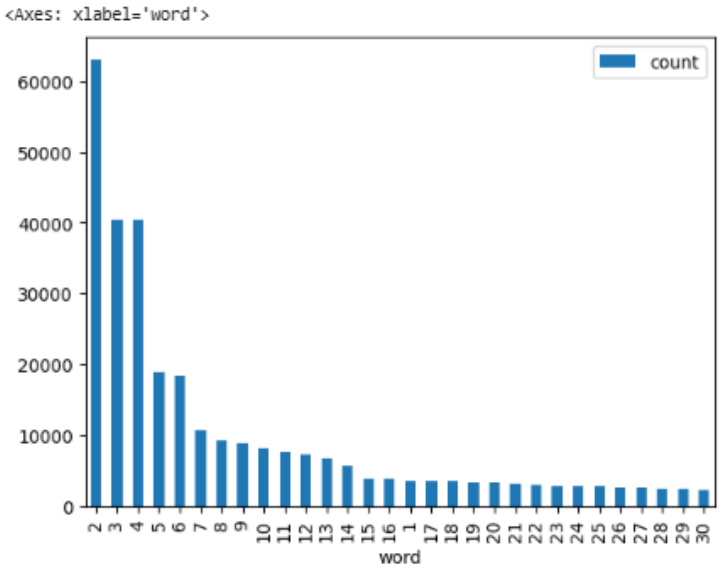
Tahap pra-pemrosesan teks dilakukan dengan membuat tokenisasi teks, yaitu memisahkan kata-kata dalam keterangan berdasarkan spasi dan menyaring elemen-elemen yang tidak relevan. Hasilnya adalah kosakata yang berisi hingga 5.000 kata unik guna menghemat memori. Kata-kata yang tidak termasuk dalam kosakata ini diganti dengan token "UNK". Kemudian, dilakukan pemetaan kata ke indeks dan sebaliknya. Semua urutan kata dipadatkan hingga memiliki panjang yang sama dengan urutan terpanjang. Selain itu, token "<start>" dan "<end>" ditambahkan ke setiap keterangan untuk menandai awal dan akhir kalimat. Hasil prapemrosesan teks dapat dilihat pada Gambar 6.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2573−2581
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6958

```
0                 <start> A child in a pink dress is climbing up a set of stairs in an entry way . <end>
1                                              <start> A girl going into a wooden building . <end>
2                                   <start> A little girl climbing into a wooden playhouse . <end>
3                                <start> A little girl climbing the stairs to her playhouse . <end>
4                              <start> A little girl in a pink dress going into a wooden cabin . <end>
5                                          <start> A black dog and a spotted dog are fighting <end>
6                  <start> A black dog and a tri-colored dog playing with each other on the road . <end>
7    <start> A black dog and a white dog with brown spots are staring at each other in the street . <end>
8                        <start> Two dogs of different breeds looking at each other on the road . <end>
9                                    <start> Two dogs on pavement moving toward each other . <end>
Name: Captions, dtype: object
```

Gambar 6. Hasil Pra-pemrosesan Teks

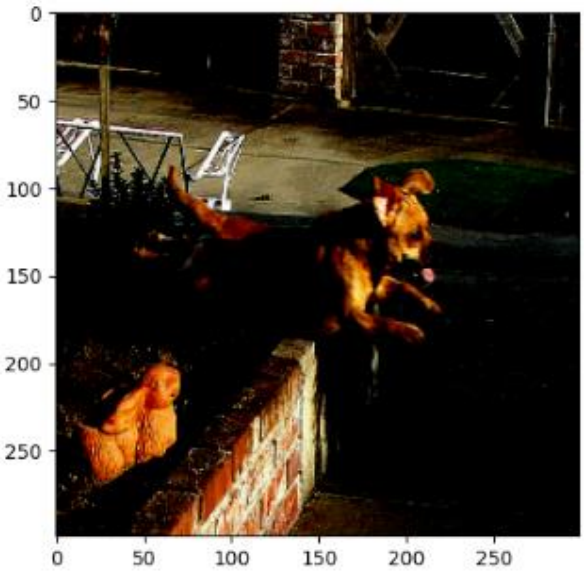
Tokenizer dari Keras digunakan untuk mengubah kata-kata menjadi angka. Visualisasi tokenizer dapat dilihat pada Gambar 7.



Gambar 7. Visualisasi Tokenizer Teks

3.1.3 Pra-Pemrosesan Gambar

Penerapan CNN (Convolutional Neural Network) dalam penelitian ini terletak pada tahap pra-pemrosesan gambar dan ekstraksi fitur. Gambar yang diimpor diubah ukurannya menjadi 299x299 piksel dan dinormalisasi, agar sesuai dengan format input untuk model InceptionV3. Fitur gambar diekstraksi menggunakan lapisan terakhir model InceptionV3 yang telah dilatih sebelumnya dengan bobot Imagenet, menghasilkan bentuk fitur sebesar 8x8x2048. Fitur ini kemudian disimpan menggunakan format kamus atau numpy untuk efisiensi. Hasil pra-pemrosesan gambar dapat dilihat pada Gambar 8.



Gambar 8. Hasil Pra-Pemrosesan Gambar

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2573−2581
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6958

3.1.4 Pembagian Dataset

Dataset dibagi menjadi data pelatihan dan pengujian dengan rasio 80:20. Setiap jalur gambar dan keterangan diproses menjadi pasangan fitur gambar dan urutan teks caption. Proses ini mencakup pengacakan dan pembagian ke dalam batch untuk memastikan performa model yang lebih baik. Hasil pembagian dataset dapat dilihat pada gambar 9.

```
Training data for images: 32364
Testing data for images: 8091
Training data for Captions: 32364
Testing data for Captions: 8091
```

Gambar 9. Train dan Test Dataset

Setelah diproses, bentuk data pelatihan dan pengujian adalah (batch_size, 8x8, 2048) untuk gambar dan (batch_size, max_len) untuk keterangan. Pembagian batch dataset dapat dilihat pada Gambar 10.

```
[ ] sample_img_batch, sample_cap_batch = next(iter(train_dataset))
    print(sample_img_batch.shape)  #(batch_size, 8*8, 2048)
    print(sample_cap_batch.shape) #(batch_size,max_len)

    (32, 64, 2048)
    (32, 39)
```

Gambar 10. Pembagian Batch Dataset

3.1.5 Pembuatan Model

Model dikembangkan dengan mengatur parameter encoder, attention model, dan decoder. Dalam tahap ini, Recurrent Neural Network (RNN) diterapkan pada bagian decoder untuk memprediksi urutan kata pada deskripsi teks. Encoder bertugas untuk menghasilkan fitur gambar dengan bentuk (32, 64, 256), sementara decoder, yang merupakan bagian dari RNN, bertugas menghasilkan urutan teks dari fitur gambar yang telah diproses. Decoder menerima informasi dari encoder dan memprediksi kata demi kata dalam urutan teks hingga mencapai kata penutup ("<end>"). RNN, dengan kemampuannya dalam memproses urutan data, digunakan di sini untuk memodelkan hubungan temporal antara kata-kata dalam deskripsi gambar.

Attention model memberikan bobot perhatian (attention weights) pada bagian-bagian tertentu dari gambar yang dianggap lebih penting saat memprediksi kata berikutnya dalam deskripsi teks. Attention ini bekerja dengan cara mengarahkan fokus model ke fitur gambar yang relevan, sehingga membantu decoder untuk menghasilkan deskripsi yang lebih akurat. Hasil encoder serta decoder dapat dilihat pada Gambar 11.

```
Feature shape from Encoder: (32, 64, 256)
Predcitions shape from Decoder: (32, 5001)
Attention weights shape from Decoder: (32, 64, 1)
```

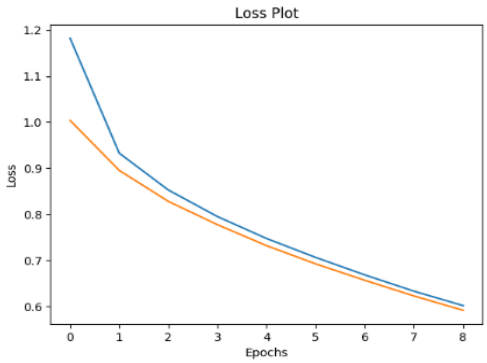
Gambar 11. Pembuatan Model

3.1.6 Pelatihan dan Pengoptimalan Model

Pelatihan dilakukan dengan menerapkan teknik teacher forcing, yang merupakan teknik yang umum digunakan dalam pelatihan model RNN, di mana kata target yang benar digunakan sebagai input berikutnya untuk decoder selama pelatihan. Dengan menggunakan teknik ini, model dapat belajar lebih cepat dan lebih efektif dalam menghasilkan urutan teks yang sesuai. Pada saat pengujian, prediksi sebelumnya digunakan sebagai input berikutnya untuk decoder, yang memperkenalkan perbedaan antara pelatihan dan pengujian tetapi menghindari overfitting.

Penggunaan RNN pada decoder memungkinkan model untuk memprediksi urutan kata yang lebih tepat, meskipun masih terdapat tantangan dalam menghasilkan deskripsi yang sangat akurat sesuai dengan gambar yang dianalisis. Meskipun demikian, proses pelatihan dengan menggunakan teknik teacher forcing telah membantu dalam memperbaiki performa model dalam memproduksi teks deskriptif. Untuk pengujian, prediksi sebelumnya digunakan sebagai input berikutnya. Teknik ini menyebabkan perbedaan antara kerugian pelatihan dan pengujian, tetapi tidak menunjukkan adanya overfitting. Proses ini mencakup pengaturan fungsi kerugian, pengoptimalan, dan jalur penyimpanan model. Model dapat dilatih lebih lama untuk hasil yang lebih baik, tetapi tujuan utamanya adalah memperkenalkan mekanisme perhatian pada arsitektur encoder-decoder untuk gambar.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2573−2581
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6958



Gambar 12. Visualisasi Pelatihan Model

Grafik pada Gambar 12 menunjukkan penurunan loss pada model selama 8 epoch pelatihan. Sumbu X mewakili jumlah epoch, sedangkan sumbu Y menunjukkan nilai loss. Dua garis pada grafik tersebut mewakili data loss dari model: garis biru untuk data pelatihan dan garis oranye untuk data validasi. Kedua loss secara konsisten menurun seiring bertambahnya epoch, menunjukkan bahwa model semakin baik dalam menyesuaikan parameter untuk meminimalkan kesalahan prediksi. Selain itu, jarak antara loss pelatihan dan validasi tetap kecil, mengindikasikan bahwa model tidak mengalami overfitting selama pelatihan ini. Dengan nilai loss yang mendekati 0, model menunjukkan performa yang baik dalam mempelajari pola data.

3.1.7 Evaluasi Model

Evaluasi dilakukan menggunakan pencarian greedy dan skor BLEU. Hasil prediksi caption untuk contoh tertentu adalah "a woman walks down a sidewalk", yang memiliki perbedaan signifikan dengan keterangan asli "a small boy holding two fireworks sparklers". Skor BLEU yang dihasilkan sangat rendah (5.154827976372712e-153), menunjukkan bahwa model belum mampu mereplikasi caption asli dengan baik. Skor BLEU yang sangat rendah tersebut kemungkinan disebabkan oleh dataset yang masih terbatas, kurangnya regularisasi untuk mengurangi overfitting, atau arsitektur model yang perlu dioptimalkan, seperti penambahan kapasitas decoder atau penyesuaian hyperparameter untuk meningkatkan kemampuan model dalam memahami hubungan antara gambar dan teks. Gambar 13 hasil prediksi teks yang telah berhasil dievaluasi oleh model.

Gambar 13. Hasil Prediksi Caption Pada Gambar

Hasil audio yang dihasilkan dari sistem ini, menggunakan Google Text-to-Speech (gTTS), berhasil mengonversi teks prediksi menjadi suara dengan jelas. Gambar 15 adalah hasil caption yang didapat dari pembacaan gambar.

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2573−2581
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6958

```
[ ] captions=beam_evaluate(test_image)
    print(captions)

    a woman walks down the sidewalk

[ ] from gtts import gTTS
    from IPython.display import Audio
    tts = gTTS("Predicted Caption is: "+ pred_caption)
    tts.save('s.wav')
    sound_file = 's.wav'
    Audio(sound_file, autoplay=True)
```

Gambar 14. Hasil Pembuatan Audio Prediksi Gambar

Audio yang dihasilkan mengucapkan kalimat "a woman walks down a sidewalk" dengan cukup baik. Meskipun demikian, meskipun audio tersebut mudah dipahami, masih terdapat perbedaan yang signifikan antara prediksi model dan keterangan asli dari gambar yang diproses. Hasil ini menunjukkan bahwa meskipun proses konversi teks ke suara berjalan dengan baik, ada tantangan dalam hal akurasi model dalam menghasilkan teks yang tepat sesuai dengan gambar yang dianalisis. Namun, langkah selanjutnya dapat difokuskan untuk meningkatkan kemampuan model dalam memprediksi caption yang lebih akurat, dengan hasil audio yang lebih sesuai dengan konteks gambar tersebut.

3.2 Pembahasan

Dalam penelitian ini, penerapan CNN untuk ekstraksi fitur gambar dan pemodelan teks menunjukkan hasil yang menggembirakan, meskipun masih terdapat beberapa tantangan. Berdasarkan penelitian terdahulu, banyak studi yang menggunakan CNN untuk klasifikasi gambar telah mencapai akurasi tinggi, seperti yang dilakukan oleh Inggis Kurnia Trisiawan yang mencapai 95,02% dalam klasifikasi botol minuman [4], serta Farah Afi Febriyanti yang memperoleh akurasi 96% dalam deteksi penyakit kulit [5]. Penelitian ini memperlihatkan bahwa meskipun CNN dapat mengenali pola gambar dengan baik, ada peluang untuk meningkatkan akurasi lebih lanjut, terutama pada dataset dan arsitektur model.

Namun, tantangan utama dalam penelitian ini terletak pada prediksi teks atau caption yang dihasilkan oleh model. Meskipun CNN efektif dalam memproses gambar, model ini belum mampu menghasilkan deskripsi yang akurat dan relevan dengan gambar yang dianalisis. Sebagai contoh, hasil evaluasi menggunakan skor BLEU menunjukkan perbedaan signifikan antara prediksi caption dan keterangan gambar asli, seperti yang terlihat pada contoh "a woman walks down a sidewalk" yang sangat berbeda dengan keterangan asli "a small boy holding two fireworks sparklers". Ini menunjukkan bahwa penghubungan antara gambar dan teks memerlukan pengembangan lebih lanjut, khususnya dalam hal penerapan mekanisme perhatian (attention model).

Berdasarkan perbandingan dengan penelitian terdahulu, meskipun model CNN yang digunakan dalam penelitian ini berhasil mengolah gambar dengan baik, hasil yang diperoleh menunjukkan bahwa akurasi pada prediksi teks masih perlu diperbaiki. Penelitian seperti yang dilakukan oleh Yoke Annisa Putri Vandalis dalam klasifikasi sampah mencapai akurasi 88%, namun model tersebut masih perlu penguatan pada tugas yang lebih kompleks seperti pengenalan teks [7]. Oleh karena itu, langkah selanjutnya dalam penelitian ini adalah fokus pada peningkatan akurasi dalam prediksi teks, dengan mempertimbangkan penggunaan model yang lebih canggih dan dataset yang lebih besar serta lebih beragam. Hal ini diharapkan dapat memperkecil gap antara prediksi model dengan keterangan gambar yang lebih akurat dan relevan.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa penggabungan model CNN:RNN untuk ekstraksi fitur gambar dan model encoder-decoder untuk generasi teks memiliki potensi dalam menghasilkan caption gambar, meskipun hasilnya belum optimal. Proses ini melibatkan pemrosesan gambar menggunakan InceptionV3 untuk mengekstraksi fitur visual, yang kemudian dikombinasikan dengan model perhatian (attention) untuk menghasilkan deskripsi. Meskipun sistem berhasil mengonversi teks hasil prediksi menjadi audio dengan Google Text-to-Speech, evaluasi menunjukkan skor BLEU yang sangat rendah, menandakan adanya keterbatasan model dalam mengidentifikasi elemen-elemen gambar secara akurat. Ke depannya, penelitian dapat ditingkatkan dengan mengoptimalkan proses pelatihan, memperbaiki akurasi caption, serta memanfaatkan model yang lebih kompleks agar menghasilkan deskripsi gambar dan audio yang lebih relevan dan sesuai dengan konteks gambar. Hal ini akan meningkatkan kemampuan model dalam menghasilkan deskripsi gambar yang lebih akurat dan relevan, yang tentunya akan menghasilkan audio yang lebih sesuai dengan

Building of Informatics, Technology and Science (BITS)
Volume 6, No 4, March 2025 Page: 2573−2581
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v6i4.6958

gambar yang dianalisis. Seiring dengan kemajuan teknologi, pemanfaatan model-model yang lebih kompleks dan teknik pelatihan yang lebih lanjut dapat memberikan hasil yang lebih memuaskan.

REFERENCES

[1] A. I. Pradana and W. Wijiyanto, "Identifikasi Jenis Kelamin Otomatis Berdasarkan Mata Manusia Menggunakan Convolutional Neural Network (CNN) dan Haar Cascade Classifier," *G-Tech: Jurnal Teknologi Terapan*, vol. 8, no. 1, 2024, doi: 10.33379/gtech.v8i1.3814.

[2] M. F. Prayuda, "Classification of Sad Emotions and Depression Through Images Using Convolutional Neural Network (CNN)," *Jurnal Informatika Universitas Pamulang*, vol. 6, no. 1, 2021, doi: 10.32493/informatika.v6i1.8433.

[3] D. Iswantoro and D. Handayani UN, "Klasifikasi Penyakit Tanaman Jagung Menggunakan Metode Convolutional Neural Network (CNN)," *Jurnal Ilmiah Universitas Batanghari Jambi*, vol. 22, no. 2, 2022, doi: 10.33087/jiubj.v22i2.2065.

[4] I. K. Trisiawan and Y. Yuliza, "Penerapan Multi-Label Image Classification Menggunakan Metode Convolutional Neural Network (CNN) Untuk Sortir Botol Minuman," *Jurnal Teknologi Elektro*, vol. 13, no. 1, 2022, doi: 10.22441/jte.2022.v13i1.009.

[5] F. A. Febriyanti, "Image Processing Dengan Metode Convolutional Neural Network (CNN) Untuk Deteksi Penyakit Kulit Pada Manusia," *Kohesi: Jurnal Sains dan Teknologi*, vol. 3, no. 10, pp. 21–30, 2024, doi: https://doi.org/10.3785/kohesi.v3i10.4088.

[6] E. Oktafanda, "Klasifikasi Citra Kualitas Bibit dalam Meningkatkan Produksi Kelapa Sawit Menggunakan Metode Convolutional Neural Network (CNN)," *Jurnal Informatika Ekonomi Bisnis*, 2022, doi: 10.37034/infeb.v4i3.143.

[7] Y. A. P. Vandalis, S. Soim, and Lindawati, "Pengembangan Algoritma Convolutional Neural Networks (CNN) untuk Klasifikasi Objek dalam Gambar Sampah," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 2, pp. 797–806, 2023, doi: https://doi.org/10.47065/bits.v6i2.5585.

[8] A. Akram, K. Fayakun, and H. Ramza, "Klasifikasi Hama Serangga pada Pertanian Menggunakan Metode Convolutional Neural Network," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 2, 2023, doi: 10.47065/bits.v5i2.4063.

[9] N. Lubis, Mhd. Z. Siambaton, and R. Aulia, "Implementasi Algoritma Deep Learning pada Aplikasi Speech to Text Online dengan Metode Recurrent Neural Network (RNN)," *Sudo Jurnal Teknik Informatika*, vol. 3, no. 3, pp. 113–126, 2024, doi: https://doi.org/10.56211/sudo.v3i3.583.

[10] Y. C. Adi, W. Priharti, and I. Hidayat, "Implementasi Pengenal Tulisan Tangan Menggunakan Optical Character Recognition Dengan Metode Cnn Dan Rnn Pada Dokumen Resi Dan Kuitansi," *e-Proceeding of Engineering*, vol. 11, no. 1, pp. 32–38, 2024.

[11] D. T. Adherda, M. Hikmatyar, and Ruuhwan, "Gender Classification Based On Voice Using Recurrent Neural Network (Rnn)," *Antivirus : Jurnal Ilmiah Teknik Informatika*, vol. 17, no. 1, 2023, doi: 10.35457/antivirus.v17i1.3049.

[12] Y. A. Suwitono and F. J. Kaunang, "Implementasi Algoritma Convolutional Neural Network (CNN) Untuk Klasifikasi Daun Dengan Metode Data Mining SEMMA Menggunakan Keras," *Jurnal Komtika (Komputasi dan Informatika)*, vol. 6, no. 2, 2022, doi: 10.31603/komtika.v6i2.8054.

[13] Herdianto and D. Nasution, "Implementasi Metode Cnn Untuk Klasifikasi Objek," *METHOMIKA: Jurnal Manajemen Informatika & Komputersisasi Akuntansi*, vol. 7, no. 1, pp. 54–60, 2023, doi: https://doi.org/10.46880/jmika.Vol7No1.pp54-60.

[14] J. V. P. Putra, F. Ayu, and B. Julianto, "Implementasi Pendeteksi Penyakit pada Daun Alpukat Menggunakan Metode CNN," *Stains (Seminar Nasional Teknologi & Sains)*, vol. 2, no. 1, 2023, doi: https://doi.org/10.29407/stains.v2i1.2888.

[15] R. H. Alfikri, M. S. Utomo, H. Februariyanti, and E. Nurwahyudi, "Pembangunan Aplikasi Penerjemah Bahasa Isyarat Dengan Metode Cnn Berbasis Android," *Jurnal Teknoinfo*, vol. 16, no. 2, pp. 183–197, 2022, doi: https://doi.org/10.33365/jti.v16i2.1752.

[16] F. N. Cahya, N. Hardi, D. Riana, and S. Hadiyanti, "Klasifikasi Penyakit Mata Menggunakan Convolutional Neural Network (CNN)," *SISTEMASI*, vol. 10, no. 3, 2021, doi: 10.32520/stmsi.v10i3.1248.

[17] Y. B. E. Purba, N. F. Saragih, A. P. Silalahi, S. Sitepu, and A. Gea, "Perancangan Alat Pendeteksi Kematangan Buah Nanas Dengan Menggunakan Mikrokontroler Dengan Metode Convolutional Neural Network (CNN)," *Methotika: Jurnal Ilmiah Teknik Informatika*, vol. 2, no. 1, 2022.

[18] A. Zalvadila, "Klasifikasi Penyakit Tanaman Bawang Merah Menggunakan Metode SVM dan CNN," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 8, no. 3, 2023, doi: 10.30591/jpit.v8i3.5341.

[19] J. R. Aisya and A. Prasetiadi, "Klasifikasi Penyakit Daun Kentang dengan Metode CNN dan RNN," *Jurnal Tekno Insentif*, vol. 17, no. 1, 2023, doi: 10.36787/jti.v17i1.888.

[20] Diki Hananta Firdaus, Bahtiar Imran, Lalu Darmawan Bakti, and Emi Suryadi, "Klasifikasi Penyakit Katarak Berdasarkan Citra Menggunakan Metode Convolutional Neural Network (Cnn) Berbasis Web," *Jurnal Kecerdasan Buatan dan Teknologi Informasi*, vol. 1, no. 3, 2022, doi: 10.69916/jkbti.v1i3.6.