JUISI Jurnal Informatika dan Sistem Informasi

*Volume 10 Nomor 1*
*e-ISSN: 2477-5452; p-ISSN: 2460-1306*
DOI: 10.37715/juisi.v10i1.4685

# Sistem Pendeskripsian Gambar Pemandangan Sekitar Bagi Penyandang Tunanetra Dengan Metode Reflective Decoding Network

# Image Captioning System of View for the Visually Impaired with Reflective Decoding Network

Vincent Leonard Santoso, Caecilia Citra Lestari*
*Program Studi Informatika, Universitas Ciputra Surabaya, Surabaya 60219, Indonesia*
*(*Email korespondensi: caecilia.citra@ciputra.ac.id)*

**Abstrak:** Terdapat 217 juta orang yang tergolong *Mild to Severe Visual Impairment* (MSVI) yang membuat penglihatan mereka sangat terganggu. Orang tunanetra juga perlu melakukan aktivitas sehari-hari yang memerlukan informasi tentang keadaan atau pemandangan sekitar. Dengan keterbatasan penglihatan yang dialami, tidak mudah bagi orang tunanetra untuk mendapatkan informasi mengenai pemandangan sekitarnya tanpa bantuan orang lain. Salah satu metode untuk menyelesaikan masalah ini adalah dengan menggunakan *image captioning* yaitu sistem yang dapat mendeskripsikan sebuah foto menggunakan *Natural Language Processing*. *Reflective Decoding Network* adalah model untuk *image captioning* yang dapat membuat *caption* pada foto dengan tingkat metric METEOR yang bagus. Adanya sistem ini dapat membantu orang tunanetra untuk mendengarkan deskripsi pemandangan tanpa perlu mencari bantuan orang lain untuk mendeskripsikan pemandangan itu sendiri. *Reflective Decoding Network* ini berhasil diimplementasikan ke dalam aplikasi berbasis iOS dimana pengguna dapat mengambil gambar dan aplikasi mengeluarkan deskripsi dalam bentuk suara menggunakan *library AVSpeechSynthesizer*. Sistem ini bekerja dengan mengirim gambar yang diambil atau diupload dari aplikasi ke sebuah server yang memiliki model *image captioning* sebagai pembuat deskripsi gambar. Deskripsi tersebut lalu akan dikirimkan kembali ke aplikasi dan diubah ke dalam bentuk suara. Model *Reflective Decoding Network* pada sistem ini memiliki nilai METEOR 20,1%. Objek deteksi pada sistem memiliki akurasi paling bagus di pencahayaan pagi dengan akurasi 53,9% dan jarak deteksi terjauh 20 meter.

**Kata Kunci:** Image Captioning, Reflective Decoding Network, Tunanetra

**Abstract:** There are 217 million people classified as Mild to Severe Visual Impairment (MSVI) which makes their vision very impaired. Blind people also need to carry out daily activities that require information about the surrounding conditions or scenery. With limited vision, it is not easy for blind people to get information about the surrounding scenery without the help of other people. One method to solve this problem is to use Image Captioning, a system that can describe a photo using Natural Language Processing. Reflective Decoding Network is a model for image captioning that can create captions for photos with good METEOR metric levels. The existence of this system can help blind people to listen to descriptions of scenes without needing to seek help from other people to describe the scenery themselves. This Reflective Decoding Network was successfully implemented into an iOS-based application where users can take pictures and the application produces a description in the form of sound using the AVSpeechSynthesizer library. This system works by sending images taken or uploaded from the application to a server that has an image captioning model as an image description maker. The description will then be sent back to the application and converted into sound. The Reflective Decoding Network model in this system has a METEOR value of 20.1%. Object detection in the system has the best accuracy in morning light with an accuracy of 53.9% and the longest detection distance is 20 meters.

JUISI Jurnal Informatika dan Sistem Informasi

*Volume 10 Nomor 1*
*e-ISSN: 2477-5452; p-ISSN: 2460-1306*
DOI: 10.37715/juisi.v10i1.4685

**Keywords**: Image Captioning, Reflective Decoding Network, Visually Impaired

*Naskah diterima 2 April 2024; direvisi 24 Mei 2024; dipublikasi 31 Mei 2024.*


## 1. Pendahuluan

Orang tunanetra adalah orang yang memiliki gangguan pada penglihatan mereka. Terdapat 2,2 miliar orang di dunia yang tunanetra dan terdapat 217 juta orang yang tergolong *Mild to Severe Visual Impairment* (MSVI) yang berarti adanya gangguan penglihatan tingkat rendah hingga tingkat tinggi tetapi tidak sepenuhnya buta (Ackland et al., 2017; Anonymous, 2022b). Dengan adanya kekurangan tersebut yaitu gangguan pada penglihatan, maka mereka membutuhkan bantuan dalam keseharian mereka.

Dengan majunya teknologi zaman sekarang, orang tunanetra sudah bisa menggunakan *smartphone* dengan adanya fitur-fitur bantuan seperti *voice over*, penjelasan *text* melalui *audio*, dan *audio descriptions*. Meskipun teknologi ini sangat membantu, fitur-fitur tersebut memiliki batasan yaitu ketidakmampuan mereka dalam mendeskripsikan sebuah gambar.

Terdapat sebuah aplikasi bernama TapTapSee yang dapat mendefinisikan sebuah objek dalam foto dan menjelaskannya dalam bentuk *audio*. Aplikasi ini sangat membantu orang tunanetra dalam mendefinisikan objek di dalam foto tetapi aplikasi ini hanya dapat menjelaskan objek tanpa menjelaskan lingkungan yang ada di sekitar objek. Dengan itu peneliti ingin membuat sebuah aplikasi yang dapat mendeskripsikan pemandangan lingkungan sekitar pada sebuah foto.

Metode yang akan digunakan untuk membuat pendeskripsian gambar adalah *Reflective Decoding Network* (RDN). RDN adalah model *image captioning* yang memiliki nilai kosakata yang tinggi sehingga penjelasan yang dibuat memiliki bahasa koheren atau jelas. RDN juga dapat memaksimalkan informasi yang akan masuk ke dalam hasil deskripsi yang dibuat. RDN juga memiliki tingkat efektivitas dan performa yang lebih tinggi dari metode-metode sebelumnya.

Dari survei yang dilakukan WebAIM pada tahun 2021 yang berjumlah 1568 peserta yang memiliki disabilitas dengan 1246 responden penderita kebutaan, 71,9% menggunakan iOS sebagai *mobile phone* mereka (WebAIM, 2021). Hal ini dikarenakan iOS memiliki banyak fitur aksesibilitas seperti VoiceOver. Karena itu sistem pendeskripsian video pemandangan akan dibuat di *platform* iOS menggunakan bahasa Swift karena banyak orang tunanetra menggunakan *platform* ini.

Penelitian ini bermaksud untuk membuat sebuah sistem pendeskripsian pemandangan sekitar bagi penyandang tunanetra dengan menerapkan model RDN. Model dilatih ulang dengan foto-foto bertemakan perkotaan yang diambil dari data set Visual Genome. Hasil deskripsi gambar yang berupa teks kemudian diubah menjadi audio. Sistem ini lalu dibuat di aplikasi *mobile* dengan *platform* iOS menggunakan bahasa Swift.

## 2. Metodologi Penelitian

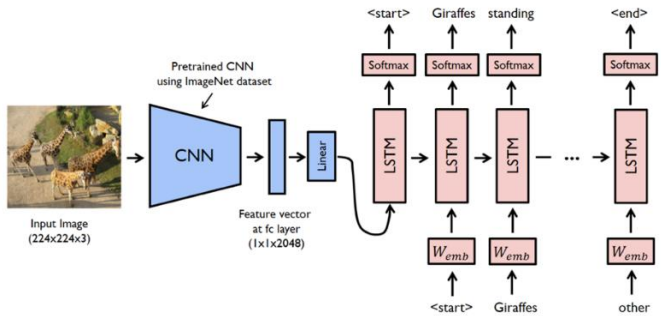
### 2.1 Kajian Pustaka

#### 2.1.1. Image Captioning

*Image captioning* adalah proses translasi dari gambar ke bentuk *text* menggunakan *Natural Language Processing* (NLP). *Image captioning* memiliki *encoder* dan *decoder* dimana *encoder* bekerja untuk melakukan *feature extraction* pada gambar dan *decoder* bekerja sebagai *language generation* yang menghasilkan deskripsi dari gambar tersebut.

Di penelitian terdahulu, *Convolutional Neural Network* (CNN) digunakan sebagai *encoder*. *Recurrent Neural Network* (RNN) digunakan sebagai *decoder* dan menggunakan Long Short-Term Memory (LSTM) yang berguna untuk menghasilkan *caption* atau kata-kata deskripsi dari sebuah gambar (Kim et al., 2019). Gambar 1 di bawah menjelaskan tentang arsitektur *image captioning*.

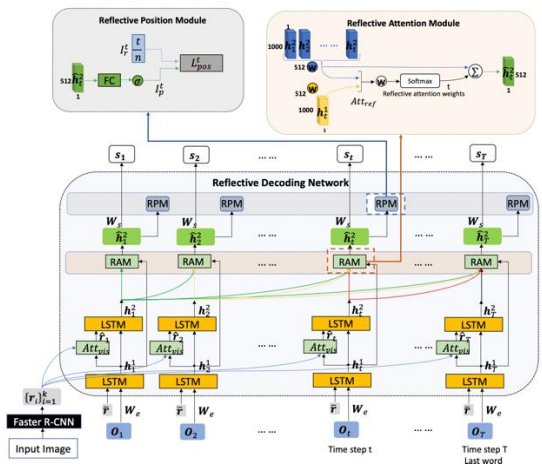
Jurnal Informatika dan Sistem Informasi
Volume 10 Nomor 1
e-ISSN: 2477-5452; p-ISSN: 2460-1306
DOI: 10.37715/juisi.v10i1.4685




**Gambar 1.** Arsitektur Image Captioning (Kim et al., 2019)

### 2.1.2. Reflective Decoding Network

Reflective Decoding Network adalah salah satu metode/algoritma untuk *image captioning*. Pada penelitiannya, Ke dan kawan-kawan menggunakan data set COCO yang berjumlah 82,783 gambar untuk melatih RDN dan 40,504 gambar dari COCO serta108.000 gambar dari Visual Genome data set untuk *validation work* (Ke et al., n.d.). RDN diawali dengan melakukan *feature extraction* menggunakan model ResNet-101 sebagai *encoder*. Model ResNet101 ini adalah model *pre-trained* yang melakukan *image classification* pada ImageNet. *Encoder* ini dilakukan untuk mendeteksi objek-objek yang ada di dalam gambar. Pada RDN, *decoder* yang digunakan menggunakan LSTM yaitu salah satu tipe RNN yang sering digunakan untuk kebutuhan NLP. Gambar 1 di bawah ini menunjukkan *framework* dari model RDN.




**Gambar 2.** Framework model Reflective Decoding Network (Ke et al., n.d.)

Pada Gambar 2, *decoder* dari RDN terdapat dua bagian yaitu Reflective Attention Module (RAM) dan Reflective Position Module (RPM). RAM melihat kompabilitas antara *hidden state* saat ini dan sebelumnya. Dengan itu RAM mengambil informasi yang lebih lengkap untuk menetapkan kata. RPM memberikan posisi relatif untuk tiap kata pada *caption*. Hal ini digunakan untuk memprediksi struktur pada kalimat.

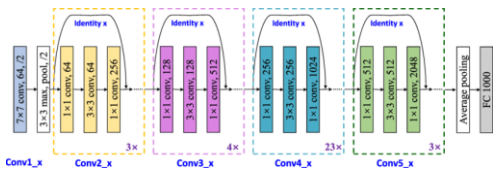
Pelatihan pada metode ini dimulai dengan melakukan *preprocess* pada teks di deskripsi gambar dengan membuat semua huruf menjadi *lower-case* dan mengurangi kata yang jarang digunakan. Di bagian *encoder*, dilakukan penetapan *Intersection over Union* (IoU) pada proposal *Suppresion* dan *Object Prediction* di angka 0,7 dan 0,3. IoU digunakan untuk mengevaluasi *object detection* pada *encoder* dengan membandingkan prediksi daerah objek yang

Jurnal Informatika dan Sistem Informasi
Volume 10 Nomor 1
e-ISSN: 2477-5452; p-ISSN: 2460-1306
DOI: 10.37715/juisi.v10i1.4685

dideteksi dengan daerah asli objek pada gambar. IoU ini digunakan untuk mengevaluasi hasil dari *encoder*. Pada LSTM, *word embedding size* dan *hidden size* ditetapkan pada angka 1,000 dan dimensi pada *attention layers* ditetapkan di angka 512.

### 2.1.3. ResNet-101

RDN menggunakan Residual Network atau ResNet untuk mendeteksi objek pada gambar. Model ResNet memiliki banyak macam berdasarkan jumlah layer dan pada penelitian ini model ResNet yang digunakan adalah model ResNet yang memiliki 101 *layer* yang memiliki arsitektur seperti Gambar 3.



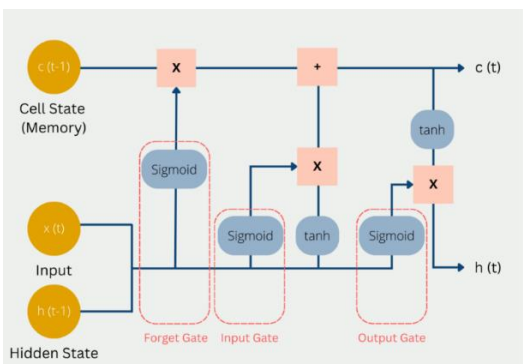

**Gambar 3.** Model ResNet-101 (Chen et al., 2021)

ResNet adalah model Convolutional Neural Network (CNN) yang diciptakan untuk menyelesaikan masalah *vanishing gradient*. Masalah ini terjadi saat *network* sudah terlalu dalam hingga *gradient* pada *loss function* dapat mengecil ke angka 0 yang membuat *weight* tidak lagi terbarui dan tidak lagi ada pembelajaran. ResNet melakukan *skip connection* yang dapat melewati *layer* dan mengambil *activation* sebelumnya. Hal ini membuat *residual block* yaitu sebuah set dari *layer* dimana hasil *output* dari *layer* dapat dikirim ke *layer* yang lebih dalam.

### 2.1.4. Long Short-Term Memory

LSTM atau Long Short-Term Memory adalah salah satu tipe Recurrent Neural Network atau RNN yang dibuat untuk menangani kelemahan RNN dalam melakukan *training* jangka panjang atau *long term memory* (Anonymous, 2022a). Pada RDN, LSTM digunakan untuk menghasilkan *caption* dari gambar.

Seperti ditunjukkan pada Gambar 4, LSTM mengeluarkan output berdasarkan 3 hal yaitu *cell state* yang merupakan memori jangka panjang, *hidden state* yaitu *output* pada langkah sebelumnya, dan *input data* pada langkah saat ini. *Cell state* adalah tempat dimana informasi jangka panjang disimpan yang membuat *cell state* sebagai *long term memory*. *Hidden state* pada LSTM memiliki fungsi menyimpan informasi kalkulasi pada step sebelumnya yang membuat *hidden state* sebagai *short term memory*.




**Gambar 4.** Arsitektur Long Short-Term Memory (Anonymous, 2022a)

LSTM memiliki tiga *gate* yang mengontrol masuk dan keluarnya informasi dalam *cell state*. *Gates* ini bisa dianggap sebagai filter dalam data yang memberikan limit terhadap informasi yang bisa masuk sehingga hanya informasi relevan yang digunakan. Tiga *gates* ini adalah *forget gate*, *input gate*, dan *output gate*. *Forget gate* berguna untuk memutuskan informasi *cell state* yang berguna berdasarkan *hidden state* dan *new input* data. Artinya *forget*

*Volume 10 Nomor 1*
*e-ISSN: 2477-5452; p-ISSN: 2460-1306*
DOI: 10.37715/juisi.v10i1.4685

*gate* memutuskan informasi bagian mana dari *cell state* yang tidak berguna dan akan dilupakan untuk mengurangi *weight*. *Input gate* berguna untuk memutuskan informasi baru apa yang akan diberikan ke *cell state* berdasarkan *hidden state* sebelumnya dan input data yang baru. Input baru yang didapat merupakan input yang sama dari *forget gate*. Kebalikan dari *forget gate*, *input gate* memutuskan informasi baru mana yang akan diingat dan disimpan ke *cell state*. *Output gate* berguna untuk memutuskan *hidden state* yang baru. *Output gate* juga memiliki input berupa *hidden state* sebelumnya dan data baru.

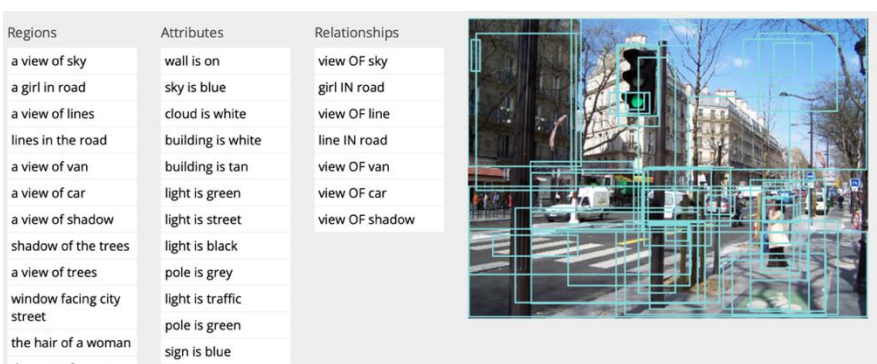
### 2.1.5. Analisis Kebutuhan Dataset

Menurut sebuah hasil survei Kominfo yang dilakukan pada tahun 2017, 83,04% penduduk perkotaan di Indonesia menggunakan smartphone dibandingkan 50,39% penduduk desa di Indonesia yang menggunakan smartphone (Finaka, 2018). Karena itu pemandangan yang akan digunakan adalah pemandangan perkotaan. Karena *view* yang ingin digunakan adalah *view* kota maka data set yang digunakan hanyalah data set yang mengandung unsur-unsur perkotaan. Menurut Kevin Lynch seorang urban planner asal Amerika, elemen-elemen dari perkotaan meliputi jalan, trotoar, transportasi umum maupun pribadi, laut pada tepi kota, taman, perumahan, alun-alun, *landmark*, dan monumen (Archi_com, 2022). Tetapi dari unsur-unsur perkotaan tersebut, yang digunakan sebagai kata kunci pada data set adalah jalan, trotoar, transportasi, dan gedung. Kata kunci ini dipilih karena unsur-unsur tersebut mudah dan sering ditemukan di dalam kota.

## 2.3. Desain

### 2.3.1. Desain Dataset

Di penelitian ini, data set yang digunakan adalah Visual Genome. Dataset yang memiliki 108.251 gambar yang sudah memiliki deskripsi secara relasional. Dataset ini diambil dari website visualgenome.org pada tanggal 20 Februari 2023. Gambar yang digunakan pada penelitian ini adalah gambar dengan resolusi minimal 600x600. Kategori pemandangan yang digunakan adalah perkotaan. Elemen-elemen yang ada di dalam kota meliputi jalan, trotoar, transportasi umum maupun pribadi, tepi laut, taman, perumahan, alun-alun, *landmark*, dan monumen (Archi_com, 2022). Pada dataset ini, terdapat tiga properti yang dapat mendapatkan objek yaitu region, *attribute*, dan *relationship*. Gambar 5 adalah contoh salah satu gambar yang ada di dalam dataset. *Attribute* adalah objek pada gambar. *Relationship* adalah hubungan antar beberapa objek. *Regions* adalah kalimat singkat yang menjelaskan gambar. Data *attribute* digunakan untuk melakukan filter sehingga data yang digunakan memiliki objek-objek tertentu yang diinginkan. Dengan itu dataset yang digunakan adalah gambar yang mengandung objek-objek pada kategori perkotaan yaitu “Street”, “Building”, “Sidewalk”, dan “Cars”



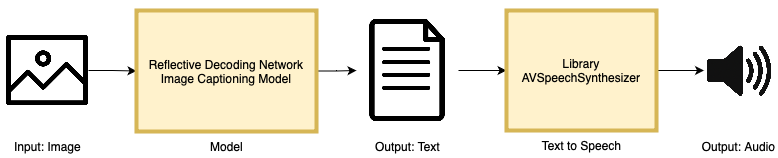

**Gambar 5.** Contoh Dataset dengan Objek “City”

### 2.3.2. Desain Sistem Deskripsi Gambar Pemandangan

Model yang dibuat, peneliti menggunakan *pipeline* pembuatan model RDN. Model menggunakan input gambar

Volume 10 Nomor 1
e-ISSN: 2477-5452; p-ISSN: 2460-1306
DOI: 10.37715/juisi.v10i1.4685

dan menghasilkan output berupa deskripsi dalam bentuk teks. Pada tahap *pre-processing*, gambar yang digunakan adalah gambar dengan resolusi minimal 600x600 sehingga gambar dengan resolusi dibawah itu tidak akan digunakan. Dataset gambar yang digunakan untuk training juga harus memiliki objek "Street", "Building", "Sidewalk", dan "Cars" pada gambar tersebut. Model RDN pertama akan melakukan *feature extraction* menggunakan model ResNet-101 yang mendapatkan informasi mengenai objek-objek pada gambar. Lalu LSTM digunakan sebagai *decoder* untuk membentuk kalimat berdasarkan objek-objek yang didapatkan dari *encoder*. *Decoder* ini akan mampu membuat kalimat yang koheren antar kata yang dibuat. Setelah *training* model dilakukan, hasil tersebut akan dievaluasi menggunakan metric METEOR. Setelah itu deskripsi yang dibuat akan dikirimkan ke aplikasi dalam bentuk *text*. Dengan menggunakan *AVSpeechSynthesizer*, *text* deskripsi diubah ke dalam bentuk suara yang akan digunakan di aplikasi. Penjelasan di atas tergambar pada Gambar 6.




**Gambar 6.** Model Sistem Deskripsi Foto Pemandangan

### 2.3.3. Desain Pengujian

Pengujian dilakukan untuk mengidentifikasi masalah yang ada pada prototipe dan melakukan perbaikan sehingga dapat melakukan perubahan pada *development cycle*. Testing juga dilakukan untuk memastikan produk yang dibuat sesuai dengan ekspektasi user. Testing yang akan dilakukan pada penelitian ini adalah pengujian deteksi model dan pengujian sistem melalui user testing.

Pengujian deteksi model dilakukan berdasarkan pencahayaan dan jarak. Pengujian berdasarkan pencahayaan dimulai dengan mengambil 10 gambar berbeda dengan waktu pengambilan pagi, sore dan malam. Setelah itu model deteksi akan diuji menggunakan gambar yang didapat untuk mengetahui pencahayaan apa yang paling akurat. Pengujian berdasarkan jarak dilakukan dengan mengambil beberapa gambar dengan objek yang sama tetapi pada jarak yang berbeda-beda. Model deteksi akan diuji untuk mengetahui jarak deteksi terjauh.

## 3. Hasil dan Pembahasan

Model *image captioning* dievaluasi menggunakan penilaian METEOR (Banerjee, S., & Lavie, A., 2005). Hasil yang didapat dari penilaian ini adalah 20.1% yang berarti hasil prediksi deskripsi dengan deskripsi asli memiliki kemiripan sebanyak 20.1%. Model kemudian dengan proses lain dalam sistem deskripsi gambar pemandangan dan dibangun menjadi aplikasi *mobile* berbasis iOS. Selanjutnya, dilakukan pengujian performa model untuk mendeteksi objek terhadap pencahayaan dan jarak. Hal ini dilakukan dengan cara melakukan pengujian pada model yang telah diterapkan di aplikasi.

Pengujian pencahayaan dilakukan dengan cara mengambil 10 gambar di tiga waktu yang berbeda yaitu pagi pukul 8, sore pukul 5, dan malam pukul 8. Total gambar yang digunakan untuk pengujian ini adalah 30 gambar. Setelah didapat gambar-gambar tersebut, dilakukan analisa terhadap objek apa saja yang ada di gambar dengan hasil objek yang terdeteksi oleh model.

Tabel 1 adalah tiga dari sepuluh pengujian yang dilakukan. Hasil pengujian ini menunjukkan bahwa model mendeteksi objek terbaik adalah pencahayaan pagi dengan akurasi 53,9%. Pada pencahayaan sore, performa model untuk mendeteksi objek adalah 42,3% sedangkan pada pencahayaan malam menghasilkan akurasi terendah yaitu 32,1%.

JUISI Jurnal Informatika dan Sistem Informasi
Volume 10 Nomor 1
e-ISSN: 2477-5452; p-ISSN: 2460-1306
DOI: 10.37715/juisi.v10i1.4685

**Tabel 1.** Pengujian Deteksi Objek Berdasarkan Pencahayaan

| Gambar | Deteksi Seharusnya | Deteksi Pagi | Deteksi Sore | Deteksi Malam |
|---|---|---|---|---|
|  | Building, tree, street, fence, sky | Large white building, red door, street, cars, tall tree (60%) | Building, street, cars, license plate (40%) | White building, black roof, silver car, street sign (20%) |
|  | Car, street, building, lamp, house, sky, trees | White car, street, trees, bushes, house (57%) | Large white building, street, cars, tree, door (57%) | White building, car, street sign, roof (29%) |
|  | Street, pole, trees, fence, metal box, building | Rectangular frame, building, clock, stone (17%) | Fire hydrant, rural street, house, brown roof, trees (33%) | Car, trees, shrubs, fence (17%) |

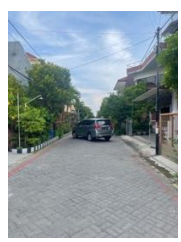
Pengujian kedua, yaitu pengujian jarak, dilakukan dengan cara mengambil gambar sebuah objek dari berbagai jarak. Objek yang digunakan adalah sebuah mobil sedangkan jarak yang digunakan adalah dua (2m), lima (5m), sepuluh (10m), dua puluh (20m), dan empat puluh (40m) meter. Total gambar yang digunakan pada pengujian kedua ini adalah lima gambar.

Tabel 2 adalah hasil pengujian performa model mendeteksi objek berdasarkan jarak. Dari tabel tersebut dapat diketahui bahwa model dapat mendeteksi mobil hingga jaran 20 meter. Mulai jarak 40 meter mobil sudah mulai tidak terdeteksi.

Dari hasil pengujian, pencahayaan paling akurat adalah pencahayaan pagi hari dimana objek yang terdeteksi paling akurat. Hal ini bisa dikarenakan pencahayaan pagi memiliki nilai lux 32.000 hingga 100.000 sedangkan pencahayaan sore hanya memiliki nilai lux 400 dan pencahayaan malam memiliki nilai lux di bawah 0,01. Lux adalah satuan nilai yang mengukur intensitas cahaya (Anonymous, n.d.). Karena rendahnya nilai lux pada pencahayaan sore dan malam, objek lebih susah untuk dideteksi sehingga akurasi menurun. Dari hasil pengujian jarak, mobil mulai tidak terdeteksi di jarak 40 meter. Hal ini dikarenakan objek mobil sudah terlihat kecil dan menjadi susah untuk dideteksi.

JUISI Jurnal Informatika dan Sistem Informasi
Volume 10 Nomor 1
e-ISSN: 2477-5452; p-ISSN: 2460-1306
DOI: 10.37715/juisi.v10i1.4685

**Tabel 2.** Pengujian Deteksi Objek Berdasarkan Jarak

| Gambar | Jarak | Hasil |
|---|---|---|
|  | 2m | Terdeteksi |
|  | 5m | Terdeteksi |
|  | 10m | Terdeteksi |
|  | 20m | Terdeteksi |
|  | 40m | Tidak terdeteksi |

## 4. Kesimpulan dan Saran

Sistem deskripsi gambar pemandangan dengan RDN berhasil dikembangkan. Performa dari model *image captioning* RDN yang dievaluasi dengan METEOR adalah 0,201 dan performa akurasi deteksi objek dari pengujian adalah 42,8%. Model bekerja mendeteksi objek paling baik saat pencahayaan pagi dengan rerata akurasi 53,9% dan paling buruk saat pencahayaan malam dengan rerata akurasi 32,1%. Model hanya dapat mendeteksi objek hingga jarak 20 meter dan di jarak 40 meter objek sudah tidak bisa terdeteksi.

*Volume 10 Nomor 1*
*e-ISSN: 2477-5452; p-ISSN: 2460-1306*
DOI: 10.37715/juisi.v10i1.4685

Dengan masih adanya batasan dari sistem yang berhasil dibangun, maka diberikan saran pengembangan dan perbaikan sebagai berikut: pertama, performa deskripsi objek, jarak deteksi objek, dan akurasi deteksi di pencahayaan tertentu dapat ditingkatkan; kedua bahasa yang digunakan pada aplikasi bisa diberikan variasi seperti bahasa Indonesia sehingga pengguna dapat lebih memahami deskripsi

**Daftar Pustaka**

Ackland, P., Resnikoff, S., & Bourne, R. (2017). World blindness and visual impairment: despite many successes, the problem is growing. Community Eye Health, 30(100), 71–73.

Anonymous. (2022a, June). Long Short-Term Memory Networks (LSTM)- simply explained! https://databasecamp.de/en/ml/lstms

Anonymous. (2022b, October 13). Blindness and vision impairment. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

Archi_com. (2022, August 19). The Elements of a City. https://archi-monarch.com/the-elements-of-a-city/#:~:text=In%20urban%20design%2C%20the%20elements,and%20character%20of%20a%20city.

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments.

Finaka, A. (2018). 66,3% masyarakat Indonesia Memiliki Smartphone #8. Indonesia Baik. https://indonesiabaik.id/infografis/663-masyarakat-indonesia-memiliki-smartphone-8

Hassan, M. (2019, January 23). ResNet (34, 50, 101): Residual CNNs for Image Classification Tasks.

Ke, L., Pei, W., Li, R., Shen, X., & Tai, Y.-W. (n.d.). Reflective Decoding Network for Image Captioning.

Kilani, M. (2021, May 31). Swift Text To Speech Synthesizer. https://medium.com/geekculture/swift-text-to-speech-synthesizer-ddf4e16f3fc6

Kim, D.-J., Choi, J., Oh, T.-H., So Kweon, I., & Korea, S. (2019). Dense Relational Captioning: Triple-Stream Networks for Relationship-Based Captioning.

WebAIM. (2021, June 30). Screen Reader User Survey #9 Results. https://webaim.org/projects/screenreadersurvey9/