Visual Question Answering Bahasa Indonesia Berbasis *Deep Learning* untuk Pembelajaran Visual Anak TK

Indonesian Visual Question Answering Based on Deep Learning for Kindergarten's Visual Learning

Asiyah Hanifah[1], Rizka Wakhidatus Sholikah[2], RV Hari Ginardi[3]
[1,2,3]Departemen Teknologi Informasi, Institut Teknologi Sepuluh Nopember
E-mail: [1]asiyahhanifah9@gmail.com, [2]wakhidatus@its.ac.id, [3]hari.ginardi@its.ac.id,

Abstrak

Indonesia semakin gencar melakukan persiapan transformasi digital dalam berbagai sektor, termasuk dalam bidang pendidikan. Salah satu upaya yang dilakukan pemerintah adalah dengan mengimplementasikan platform *e-learning* dalam kegiatan belajar mengajar hingga ke jenjang taman kanak-kanak. Metode pembelajaran visual pada taman kanak-kanak dapat diimplementasikan ke dalam *e-learning* yang lebih interaktif dan menarik dengan sistem *Visual Question Answering* (VQA). Sistem VQA dapat memberikan pertanyaan terkait dengan gambar yang ditampilkan dan mengecek kesesuaian jawaban dari siswa secara otomatis. Pada penelitian ini dibangun sistem VQA yang dapat menerima pertanyaan berbahasa Indonesia dan mengoreksi jawaban dalam bahasa Indonesia. Sistem dibangun dengan menggunakan model *Bootstrapping Language-Image Pre-training* (BLIP) untuk VQA dan model *No Language Left Behind* (NLLB) untuk penerjemahan. Uji coba dilakukan pada enam jenis jawaban yaitu ya/tidak, kata benda, kata kerja, kata sifat, kata keterangan, dan numeral. Hasil pengujian menunjukkan bahwa sistem dapat menjawab dengan nilai ketepatan 100 untuk jawaban ya/tidak, kata benda, kata kerja, dan numeral. Sementara untuk kata sifat dan kata keterangan masing-masing memiliki nilai ketepatan 62,5 dan 87,5.

Kata kunci: *Visual question answering*, *e-learning*, BLIP, NLLB, bahasa Indonesia.

Abstract

Indonesia is preparing for digital transformation in various sectors, including education. The government is implementing an e-learning platform for teaching and learning activities up to the kindergarten level. Visual learning methods in kindergarten can be implemented into more interactive and interesting e-learning with the Visual Question Answering (VQA) system. The VQA system can give questions related to the images displayed and automatically check the students' answers. In this research, a VQA system was built to receive questions and answers in Indonesian. The system applied the Bootstrapping Language-Image Pre-training (BLIP) model for VQA and the No Language Left Behind (NLLB) model for translation. The experiment was conducted on six types of answers: yes/no, nouns, verbs, adjectives, adverbs, and numerals. The results show that the system can answer with an accuracy of 100 for yes/no answers, nouns, verbs, and numerals, while adjectives and adverbs have an accuracy of 62.5 and 87.5, respectively.

Keywords: Visual question answering, *e-learning, BLIP, NLLB, Indonesian Language*.

1. PENDAHULUAN

Memasuki era revolusi 4.0, Indonesia telah mempersiapkan *roadmap* transformasi digital, termasuk di sektor pendidikan [1]. Salah satu upaya pemerintah adalah implementasi *e-learning*, yang telah mengalami pertumbuhan signifikan di Indonesia [2]. *E-learning* merupakan model

pembelajaran melalui situs di internet dengan menggunakan multimedia secara *synchronous* maupun *asynchronous* antara pengajar dan siswa [3]. Penerapan *e-learning* perlu dilakukan di semua tingkat pendidikan, termasuk tingkat paling mendasar, agar anak-anak dapat beradaptasi dengan perkembangan teknologi. Contohnya, pembelajaran visual dapat diimplementasikan menggunakan *e-learning*, dengan memanfaatkan sistem *Visual Question Answering* (VQA). Pada pembelajaran visual secara tradisional, biasanya dipergunakan alat peraga berupa kartu bergambar. Selanjutnya guru akan menunjukkan gambar tersebut kepada siswa dan menanyakan beberapa pertanyaan terkait gambar yang ditunjukkan seperti "gambar apakah itu?", "apakah ini gambar kucing?", "ada berapa kucing yang terlihat?", dll. Siswa akan memberi *feedback* dengan menjawab pertanyaan yang diajukan oleh guru secara langsung. Dalam sistem VQA, guru dapat memberikan inputan berupa gambar dan pertanyaan kemudian siswa dapat memberikan jawaban melalui sistem. Selanjutnya sistem akan secara otomatis melakukan pengecekan terhadap jawaban yang dikirimkan oleh siswa terhadap hasil jawaban model VQA. Sistem VQA ini membuat pembelajaran visual lebih interaktif dan efektif, sehingga anak-anak dapat fokus pada kegiatan belajar. Selain itu guru dapat lebih mudah untuk memantau siswa yang dapat manjawab dengan benar dan yang belum.

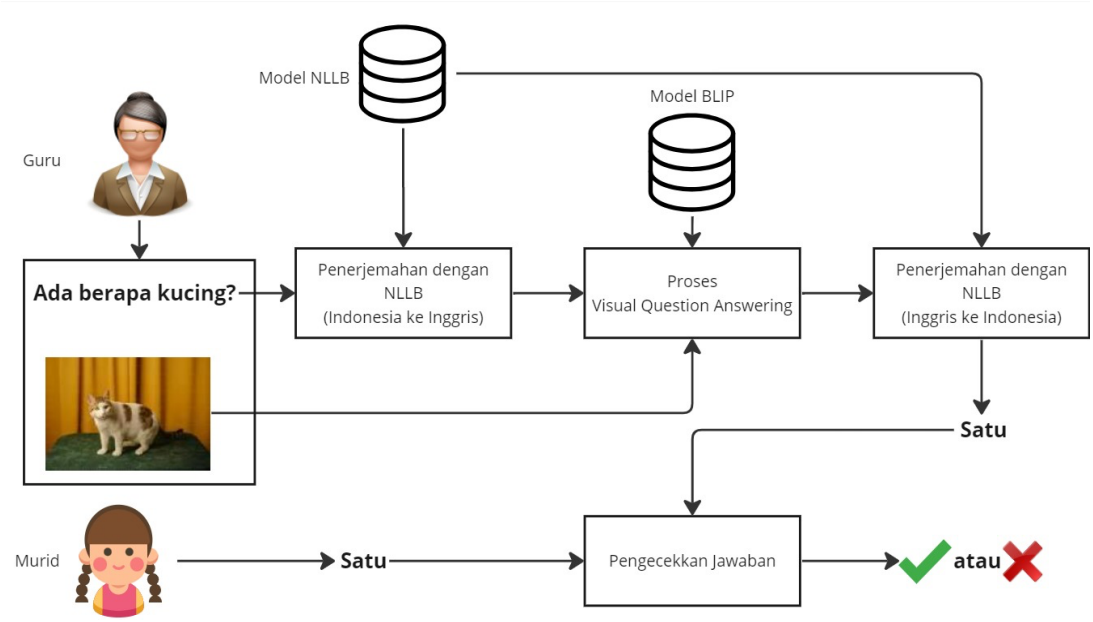
Penelitian terkait pembuatan sistem VQA untuk pembelajaran visual tekstual patologi telah dilakukan oleh He et al. [4]. Sistem ini menjawab pertanyaan terkait dengan temuan klinis yang terkandung dalam gambar yang diambil untuk pengambilan keputusan medis. Namun dataset yang digunakan terbatas dalam lingkup ilmu patologi [4]. Penelitian dengan data spesifik lainnya juga telah dilakukan untuk megembangkan VQA pada domain *remote sensing* data [5]. Dataset pada domain remote sensing dibangun sendiri oleh peneliti karena keterbatasan dataset terkait remote sensing untuk task VQA. Sementara itu peneliti menggunakan metode CNN untuk melakukan pengolahan pada gambar, sementara RNN dipergunakan untuk pengolahan pada teks [5].

Dalam sektor pariwisata, dilakukan Pembangunan VQA untuk mendeteksi gambar objek pariwisata monas berbahasa Indonesia [6]. Diketahui pada penelitian ini bahwa tidak ada data VQA yang tersedia dalam bahasa Indonesia, sehingga data VQA monas dikumpulkan dan disusun sendiri oleh peneliti [6].

Saat ini pengembangan sistem VQA pada bahasa Indonesia memiliki kendala dalam ketersediaan dataset berbahasa Indonesia. Salah satu cara untuk mengatasi permasalahan tersebut adalah dengan menggunakan teknologi penerjemah untuk merubah masukan pengguna dalam bahasa Indonesia ke bahasa Inggris. Beberapa penelitian terkait domain umum juga telah dilakukan, diantaranya pembangunan VQA dengan dataset gambar objek-objek umum yaitu DAQUAR-FULL, DAQUAR-37, dan COCO-QA [7]. Hasil uji coba dengan dataset COCO-VQA memiliki nilai akurasi terbaik di antaranya dengan menunjukkan perbaikan performa dengan akurasi 57,33%. Dataset ini merupakan kumpulan data keluaran microsoft yang berisi foto dari 91 jenis objek yang mudah dikenali oleh anak berusia 4 tahun dalam bahasa inggris [7]. Pengembangan VQA dalam domain umum dilakukan pada bahasa Arab oleh Kamel et al. [8]. Dataset yang dibangun berjumlah 138.000 yang berisi gambar sehari-hari. Selanjutnya sistem VQA dikembangkan dengan arsitektur VGG-16 untuk pengolahan gambar dan LSTM untuk pengolahan teks. Hasil uji coba menunjukkan akurasi tertinggi dari VQA yang dibangun mencapai 84,9% [8].

Beberapa penelitian telah dilakukan untuk mengembangkan VQA dalam mendukung proses pembelajaran pada anak pra-sekolah. He et al. [9] mengembangkan sistem VQA yang ditanamkan dalam robot untuk menemani pembelajaran anak pra-sekolah. Dataset yang dipergunakan meliputi empat kategori yaitu benda di dalam ruangan, kendaraan, hewan, dan manusia. Penelitian lain mengemas VQA ke dalam sebuah game edukatif untuk anak pra-sekolah [10]. Dalam permainan yang dibuat terdapat 20 pertanyaan dan melibatkan anak-anak dalam uji cobanya. Gupta et al. [11] mengusulkan sebuah aplikasi VQA untuk anak sekolah dasar yang diberi nama EDUVI. EDUVI dilatih dengan COCO dan VQA V2 dataset. Model yang dipergunakan dalam penyusunan sistem terdiri dari CNN untuk pengolahan gambar dan LSTM untuk pemrosesan teks [11]. Dari beberapa penelitian yang telah ada menunjukkan bahwa

pembelajaran visual dengan berbasis pada *e-learning* mampu meningkatkan *engagement* dari siswa pada level Pendidikan pra-sekolah hingga sekolah dasar.



Gambar 1 Gambaran Umum Sistem VQA

Oleh karena itu pada penelitian ini dilakukan pembuatan sistem visual question answering untuk pembelajaran anak TK berbahasa indonesia dengan menggunakan dataset COCO-VQAv2 yang sudah ada dan mudah dikenali oleh anak berusia 4 tahun. Pada implementasinya digunakan *Bootstrapping Language-Image Pre-training* (BLIP) sebagai *pretrained model visual-language understanding* yang dipergunakan untuk membangun model VQA. Selain itu dipergunakan pula *pretrained model No Language Left Behind* (NLLB) sebagai penerjemah dari bahasa Indonesia ke bahasa Inggris maupun sebaliknya, mengingat model BLIP belum bisa mengenali teks dalam bahasa Indonesia.

2. METODE PENELITIAN

Pada penelitian ini terdapat tiga proses utama yang dilakukan yaitu pemrosesan model NLLB, implementasi VQA dengan model BLIP, dan pengecekan jawaban. Gambaran umum dari cara kerja sistem dapat dilihat pada Gambar 1. Sementara detail penjelasan dari setiap proses akan dijelaskan pada subbab berikutnya.

2.1 Pemrosesan model NLLB

Model NLLB (*No Language Left Behind*) adalah model universal untuk penerjemahan teks yang mencakup hingga 200 bahasa. Penggunaan model NLLB didasarkan pada penelitian sebelumnya yang melakukan pembandingan antara NLLB dan google translate [12]. Hasil dari perbandingan menggunakan FLORES-200 dataset menunjukkan rata-rata kinerja NLLB lebih unggul dibandingkan google translate [12]. Model NLLB melakukan identifikasi bahasa sumber dan bahasa target yang dikehendaki, kemudian mengenali token bahasa pada dataset FLORE-200 [12]. Setelah tahapan *pre-processing* model melakukan *bitext mining* untuk menemukan pasangan kalimat terjemahan dalam korpus *monolingual*.

Model NLLB memiliki tiga variasi tergantung dimensi dan jumlah dataset yang dipergunakan untuk training. Ketiga variasi tersebut adalah *nllb-200-distilled-3.3B*, *nllb-200-*

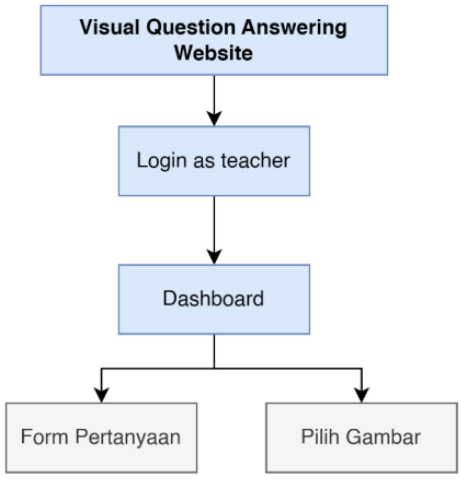
distilled-1.3B, dan *nllb-200-distilled-600M* [13]. Pada penelitian ini dipergunakan model *nllb-200-distilled-600M* yang memiliki ukuran paling kecil sehingga meringankan komputasi. NLLB dipergunakan sebagai model penerjemahan dikarenakan pada tahapan pemrosesan VQA hanya dapat menerima masukan berupa kalimat berbahasa Inggris, sehingga sistem perlu untuk menerjemahkan masukan pengguna yang berupa bahasa Indonesia ke dalam bahasa Inggris. Begitu juga untuk luaran, model VQA yang dipakai hanya mampu memberikan luaran dalam bahasa Inggris sehingga perlu diterjemahkan terlebih dahulu ke dalam bahasa Indonesia sebelum dilakukan pengecekan jawaban. Untuk memuat model NLLB, dilakukan impor *library* transformers yang digunakan untuk melatih model *pretrained state-of-the-art*. Teks dipecah menjadi subkata, contohnya "mengatakan" menjadi "meng-kata-kan". Subkata ini kemudian diproses oleh model NLLB.

2.2 Pemrosesan model BLIP

BLIP (*Bootstrapping Language-Image Pre-training*) merupakan model untuk pemrosesan teks dan gambar menggunakan arsitektur *encoder-decoder*. BILP memanfaatkan *pretrained language model* BERT sebagai *encoder* dan *decoder multimodal* dengan pembagian bobot yang cermat. BLIP juga mengusulkan *dataset bootstrapping* untuk meningkatkan kualitas teks dalam korpus *pretraining* dengan menghilangkan *noise* teks dan menghasilkan variasi teks baru [14].

BLIP memperkenalkan modul Q-former dalam arsitekturnya yang dilatih untuk mengatasi kesenjangan antara gambar yang dilatih dan *large language models* (LLM). Q-former terdiri dari dua submodul transformator dengan lapisan *self-attention* yang sama. Submodul pertama adalah transformator gambar yang berinteraksi dengan *encoder* gambar, sedangkan submodul kedua adalah transformator teks yang berfungsi sebagai *encoder* dan *decoder* teks. Pada proses pembelajaran, Q-former terhubung dengan *encoder* gambar dan dilatih menggunakan pasangan gambar-teks. Tiga tahap *pretraining* dilakukan pada Q-former: (1) *Image-Text Contrastive Learning* untuk mengoptimalkan representasi gambar dan teks sehingga informasi timbal baliknya ditingkatkan, (2) *Image-grounded Text Generation* untuk melatih Q-Former dalam menghasilkan teks dengan menggunakan gambar *input* sebagai kondisi, dan (3) *ImageText Matching* untuk memprediksi kesesuaian pasangan gambar-teks. Dalam tahap ini, kueri yang dapat dipelajari disematkan menjadi masukkan ke transformer gambar, di mana kueri tersebut saling berinteraksi melalui lapisan *self-attention* dan berinteraksi dengan gambar melalui lapisan *cross-attention* [14].

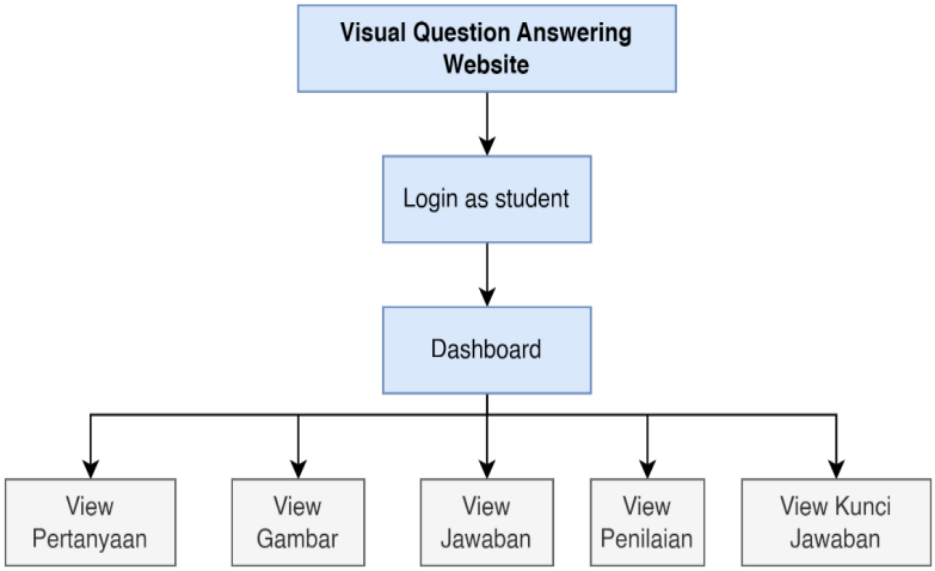
Dalam melakukan implementasi model *pre-trained* BLIP digunakan library LAVIS (*A Library for Language Vision Intelligence*) [15]. LAVIS adalah *library open-source* yang



Gambar 2 Arsitektur Informasi Pengguna Guru

mendukung *training* dan evaluasi model *state-of-the-art* pada permasalahan gambar-teks dan video-teks. LAVIS menyediakan lebih dari 30 model termasuk model BLIP. Model BLIP pada LAVIS merupakan model yang telah dilatih dengan dataset VQAv2. Framework pytorch dipergunakan untuk melakukan *load* model BLIP dari *library* LAVIS. Pada implementasi sistem model BLIP memperoleh masukkan dari pengguna berupa gambar dan teks pertanyaan dalam bahasa Inggris. Selanjutnya luaran dari BLIP berupa jawaban dalam bahasa Inggris sesuai dengan masukan pertanyaan dari pengguna.

2.3 Implementasi VQA



Gambar 3 Arsitektur Informasi Pengguna Murid

Tabel 1. Daftar Skenario Uji Coba

No. Pengujian	Gambar	Keterangan
1	Hewan	Dua gambar hewan yang sama tetapi memiliki jumlah berbeda
2		Dua gambar hewan yang memiliki spesies berbeda
3	Benda	Dua gambar benda yang sama tetapi memiliki jumlah yang berbeda
4		Dua gambar benda dari jenis yang berbeda

Tabel 2. Tujuan dari Skenario Uji Coba

No. Pengujian	Tujuan
1	Menguji apakah sistem VQA dapat menedeteksi jumlah hewan yang berbeda
2	Menguji apakah sistem VQA dapat mendeteksi hewan sejenis tetapi dari spesies yang berbeda
3	Menguji apakah sistem VQA dapat mendeteksi jumlah benda yang berbeda
4	Menguji apakah sistem VQA dapat mendeteksi benda dari jenis yang berbeda

Implementasi pada aplikasi website, diperlukan arsitektur informasi yang efektif dan mudah dipahami oleh pengguna. Aplikasi memiliki dua jenis akses yaitu untuk pengguna guru dan murid. Pengguna guru dapat mengakses halaman *dashboard* yang berisi form pertanyaan dan gambar untuk diajukan kepada murid. Guru dapat memasukkan pertanyaan dan gambar pada sistem, selanjutnya sistem memproses VQA untuk mendapatkan kunci jawaban. Ilustrasi dari arsitektur informasi pada pengguna guru dan pengguna murid dapat dilihat pada Gambar 2 dan Gambar 3.

Pengguna murid dapat mengakses halaman *dashboard* yang menampilkan pertanyaan dan gambar yang dikirim oleh guru, serta form jawaban untuk diisi. Setelah mengirim jawaban, pengguna murid akan diarahkan ke halaman *correction* untuk memeriksa hasil jawaban. Halaman *correction* ini menampilkan pertanyaan-gambar dari guru, jawaban murid, serta penilaian dari jawaban murid apakah benar atau salah. Jika jawaban yang dimasukkan oleh murid salah, maka sistem akan menampilkan jawaban yang benar. Jawaban yang benar dari VQA secara otomatis dihasilkan oleh sistem. Pengecekan jawaban dari murid juga dilakukan secara otomatis oleh sistem. Selanjutnya pada tahap *finishing*, data pertanyaan, gambar, dan jawaban akan dihapus dari *database* agar data tidak redundan saat pengguna ingin memasukkan data baru.

3. HASIL DAN PEMBAHASAN

3.1 Skenario Uji Coba

Pada penelitian ini dilakukan beberapa skenario uji coba untuk menguji sistem VQA yang dibangun. Detail skenario uji coba dapat dilihat pada Tabel 1, sementara untuk keterangan tujuan dari masing-masing skenario uji coba dapat dilihat pada Tabel 2. Dari setiap skenario uji coba dilakukan pengujian terhadap enam jenis pertanyaan yaitu pertanyaan ya/ tidak, kata benda, kata kerja, kata sifat, kata keterangan, dan numeral.

3.2. Hasil Pengujian

Pengujian pertama dilakukan pada dua gambar hewan “Gajah” yang memiliki jumlah berbeda. Untuk masing-masing gambar diajukan enam jenis pertanyaan dan dievaluasi apakah jawaban yang dihasilkan benar atau tidak. Pertanyaan dan Gambar merupakan masukan dari pengguna, sementara jawaban adalah hasil luaran otomatis dari sistem VQA. Tabel 3 menunjukkan hasil pengujian untuk skenario uji coba 1.

Pengujian kedua dilakukan dengan menggunakan dua buah gambar hewan sejenis tetapi dari spesies yang berbeda. Pada pengujian ini digunakan gambar “Burung Beo” dan “Burung Elang”. Pengujian dilakukan dengan mengajukan enam jenis pertanyaan untuk masing-masing gambar. Hasil pengujian dapat dilihat pada Tabel 4.

Pengujian ketiga dilakukan pada gambar benda dengan jenis yang sama tetapi memiliki jumlah yang berbeda. Untuk uji coba ketiga ini diberikan juga enam pertanyaan untuk masing-masing gambar seperti yang dapat dilihat pada Tabel 5. Pengujian terakhir dilakukan pada dua gambar dari jenis benda yang berbeda. Dalam uji coba ke enam digunakan gambar "meja" dan gambar "pizza" Setiap gambar diujikan dengan enam pertanyaan yang berbeda. Hasil pengujian dapat dilihat pada Tabel 6.

Tabel 3. Hasil Pengujian Skenario Uji Coba 1

Gambar 1		**Gambar 2**	
			
Pertanyaan	**Jawaban**	**Pertanyaan**	**Jawaban**
Apakah hewan itu besar?	Ya, itu	Apakah hewan itu besar?	Ya, itu
Hewan apakah itu?	Gajah	Hewan apakah itu?	Gajah
Apa yang dilakukan hewan itu?	Berjalan	Apa yang dilakukan hewan itu?	Berjalan
Bagaimana tekstur bulu hewan itu?	Berkerut	Bagaimana tekstur bulu hewan itu?	Berkerut
Dimanakah letak foto itu diambil?	Di jalan	Dimanakah letak foto itu diambil	Afrika
Berapakah jumlah hewan yang ada di dalam foto?	1	Berapakah jumlah hewan yang ada di dalam foto?	4

Tabel 5. Hasil Pengujian Skenario Uji Coba 2

Gambar 1		**Gambar 2**	
			
Pertanyaan	**Jawaban**	**Pertanyaan**	**Jawaban**
Apakah hewan itu monyet?	Tidak	Apakah hewan itu monyet?	Tidak
Burung jenis apakah itu?	Beo	Burung jenis apakah itu?	Elang
Apa yang dilakukan hewan itu?	Terbang	Apa yang dilakukan hewan itu?	Terbang
Bagaimana tekstur bulu hewan itu?	Berlemak	Bagaimana tekstur bulu hewan itu?	Berlemak
Dimanakah letak foto itu diambil?	Di hutan	Dimanakah letak foto itu diambil	Di gunung
Berapa jumlah hewan yang ada di dalam foto?	1	Berapa jumlah hewan yang ada di dalam foto?	1

Tabel 4. Hasil Pengujian Skenario Uji Coba 3

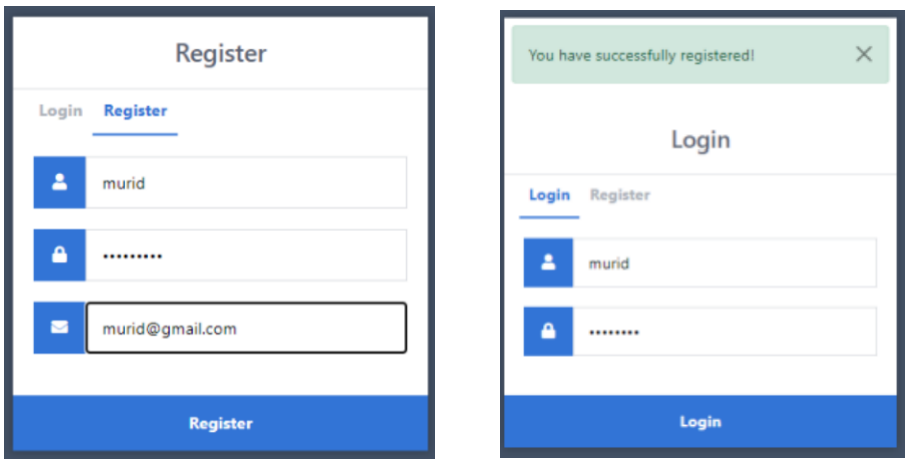
Gambar 1		**Gambar 2**	
			
Pertanyaan	**Jawaban**	**Pertanyaan**	**Jawaban**
Apakah benda itu motor?	Tidak	Apakah benda itu motor?	Tidak
Benda apakah itu?	Mobil	Benda apakah itu?	Mobil olahraga
Selain mobil, benda apa saja yang ada di dalam foto?	pohon	Selain mobil, benda apa saja yang ada di dalam foto?	Bangunan
Apa warna benda yang ada di dalam foto?	Perak	Apa warna mobil yang berada di tengah dari gambar?	Kuning
Dimanakah letak foto itu diambil?	Jalan	Dimanakah letak foto itu diambil?	Tempat parkir
Berapakah jumlah mobil yang ada di dalam foto	1	Berapakah jumlah mobil yang ada di dalam foto?	3

Tabel 6. Hasil Pengujian Skenario Uji Coba 4

Gambar 1		Gambar 2	
			
Pertanyaan	**Jawaban**	**Pertanyaan**	**Jawaban**
Apakah benda itu kursi?	Tidak	Apakah benda itu mie?	Tidak
Benda apakah itu?	Meja	Benda apakah itu?	Pizza
Apa warna meja itu?	Berwarna coklat	Bagaimana rasa pizza yang ada di dalam foto jika dimakan?	Baik
Selain meja ada benda apa lagi di dalam gambar?	Tidak ada	Selain pizza ada benda apa lagi di dalam gambar?	Garpu
Dimanakah meja itu diletakkan?	Di depan dinding	Dimanakah pizza itu diletakkan?	Di atas meja
Berapakah jumlah meja yang ada di dalam foto?	1	Berapakah jumlah pizza yang ada di dalam foto?	1 potong

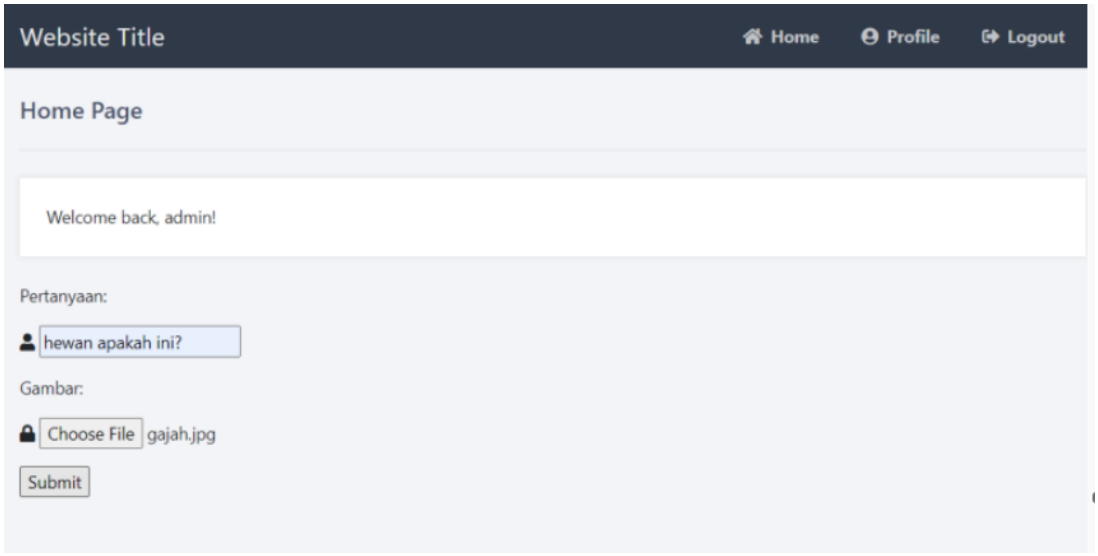
3.3. Hasil Implementasi

Sistem VQA diimplementasikan ke dalam website agar dapat diakses dengan mudah oleh pengguna guru maupun murid baik dalam pembelajaran daring maupun luring. Terdapat beberapa halaman yang dimiliki oleh sistem diantaranya halaman registrasi, halaman login, halaman dashboard guru dan halaman murid. Halaman registrasi berisi *username*, email, dan *password* seperti yang dapat dilihat pada Gambar 4 (a). Pengguna harus mendaftarkan akunnya terlebih dahulu sebelum dapat masuk ke dalam aplikasi. Setelah melakukan registrasi pengguna dapat melakukan *login* dengan *username* dan *password* yang sudah terdaftar. Halaman *login* dapat dilihat pada Gambar 4 (b). Saat pengguna berhasil login maka akan diarahkan ke halaman sesuai dengan hak aksesnya apakah sebagai guru atau murid. Bagi pengguna guru akan menuju ke

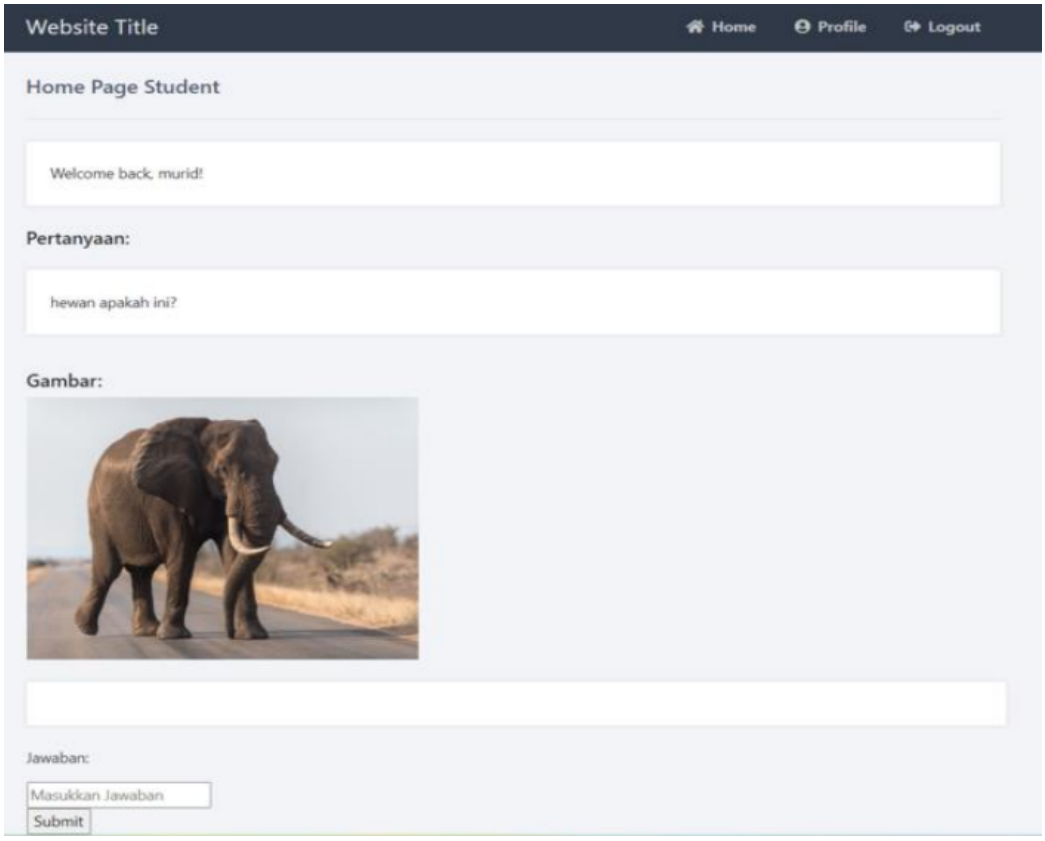


(a) (b)

Gambar 4 (a) Antar Muka Halaman Registrasi; (b) Antar Muka Halaman *Login*



Gambar 5 Antar Muka Halaman Guru



Gambar 6 Antar Muka Halaman Murid

halaman guru, Dimana guru dapat membuat soal dengan memasukkan pertanyaan dan gambar Antar muka dari halaman guru dapat dilihat pada Gambar 5. Sementara itu bagi pengguna murid akan menuju ke halaman soal seperti yang dapat dilihat pada Gambar 6. Pada halaman soal siswa dapat melihat pertanyaan, gambar terkait pertanyaan, *text box* untuk mengetikkan jawaban dan tombol untuk mengirimkan jawaban (*submit*).

3.4. Analisis Hasil Pengujian

Tabel 8. Rentang Penilaian

Nilai angka (76-100)	Baik
Nilai angka (51-75)	Cukup
Nilai angka (0-50)	Kurang

Tabel 7. Ringkasan Hasil Uji Coba

No. Pengujian	Jumlah Prediksi Jawaban Benar					
	Ya/ Tidak	Kata Benda	Kata Kerja	Kata Sifat	Numeral	Kata Keterangan
1.	2	2	2	2	2	2
2.	2	2	2	0	2	2
3.	2	2	2	2	2	1
4.	2	2	2	1	2	2
Total	8	8	8	5	8	7
Nilai Ketepatan	100	100	100	62,5	100	87,5

Pada pengujian sistem VQA berdasarkan setiap kategori *test case* yang dibuat, didapatkan hasil prediksi jawaban terkait jenis pertanyaan yang diajukan. Penilaian hasil prediksi jawaban ini dibuat berdasarkan justifikasi ketepatan jawaban. Hasil pengujian pertama terkait prediksi jawaban jumlah hewan, sistem menghasilkan prediksi jawaban yang sesuai dengan jumlah gambar yang diujikan. Namun untuk pertanyaan dengan jenis jawaban ya/tidak, jawaban yang dihasilkan kurang sempurna dengan menghasilkan prediksi jawaban, "ya, itu." Meskipun prediksi jawabannya benar dan berarti sistem sudah berhasil memprediksi jawaban yang benar sesuai gambar, namun, struktur kata dari jawaban yang dihasilkan kurang sempurna. Hal ini disebabkan karena pertanyaan "Apakah hewan itu besar?" dalam bahasa inggris menjadi, "*is the animal big?*" sehingga jawaban dalam bahasa inggris yang dihasilkan oleh sistem adalah "*yes, it is*." Karena jawaban ini dialihbahasakan menjadi bahasa Indonesia oleh model NLLB, jawaban yang dihasilkan menjadi "ya, itu." Kata "*it is*" inilah yang membuat hasil terjemahan menambahkan kata "itu". Dari 12 *test case* pengujian ini, prediksi jawaban benar adalah 12 dari 12 pertanyaan yang diujikan.

Pada uji coba ke-dua terkait prediksi jawaban spesies gambar hewan, sistem juga dapat menghasilkan prediksi jawaban yang tepat. Pada gambar pertama, sistem menjawab jenis "burung beo" dan pada gambar kedua sistem menjawab jenis "burung elang". Artinya, sistem ini dapat mendeteksi spesies hewan. Namun untuk pertanyaan dengan jenis jawaban kata sifat dari kedua gambar yang berbeda, sistem menjawab salah, seharusnya prediksi jawaban dari pertanyaan "bagaimana tekstur bulu hewan tersebut?" adalah "halus", dan prediksi jawaban yang dihasilkan adalah "berlemak". Hal ini disebabkan karena pertanyaan terkait kata sifat membutuhkan jawaban yang mendetail sedangkan dataset yang dipakai lebih umum. Dari 12 *test case* pengujian ini, prediksi jawaban benar adalah 10 dari 12 pertanyaan yang diujikan.

Dalam scenario uji coba ke-tiga terkait prediksi jawaban jumlah benda, sistem dapat menghasilkan prediksi jawaban yang benar. Sistem ini juga dapat mendeteksi benda-benda yang ada di dalam gambar. Dari 12 *test case* pengujian ini, prediksi jawaban benar adalah 12 dari 12 pertanyaan yang diujikan. Selanjutnya pada pengujian terakhir terkait prediksi jawaban dengan jenis benda yang berbeda. Sistem menjawab sebagian pertanyaan dengan benar, kecuali untuk jenis jawaban kata sifat. Prediksi jawaban untuk pertanyaan "Bagaimana rasa pizza yang ada di foto setelah dimakan?" pada gambar makanan pizza salah, seharusnya sistem menghasilkan prediksi jawaban "enak", atau "pedas", atau "asin", atau kata sifat indra perasa yang lain. Namun, prediksi jawaban yang dihasilkan adalah "baik". Jika dibandingkan dengan gambar benda selain makanan, prediksi jawaban dari sistem jauh lebih tepat. Hal ini disebabkan karena pertanyaan untuk gambar makanan terutama jenis jawaban kata sifat terlalu kompleks dan subjektif jika

dibandingkan dengan gambar benda. Dari 12 *test case* pengujian ini, prediksi jawaban benar adalah 11 dari 12 pertanyaan yang diujikan.

Pada pengujian dibuat rentang penilaian untuk tiap-tiap parameter agar mempermudah dalam melakukan analisis. Rentang penilaian dapat dilihat pada Tabel 7. Tabel 8 menunjukkan ringkasan dari pengujian untuk setiap *test case* dari semua scenario uji coba beserta nilai ketepatannya untuk setiap jenis pertanyaan. Dari hasil analisis nilai ketepatan prediksi jawaban, didapatkan bahwa untuk pengujian prediksi jawaban dengan jenis jawaban kata sifat memiliki nilai ketepatan yang cukup dikarenakan pertanyaan dengan jenis jawaban kata sifat terlalu kompleks, mendetil, dan subjektif sehingga sistem VQA belum menghasilkan prediksi jawaban yang tepat. Pada jenis jawaban kata lain memiliki nilai ketepatan yang baik dikarenakan pertanyaan dengan jenis jawaban ya/tidak, kata benda, kata kerja, kata sifat, kata numeral, dan kata keterangan lebih umum dan dapat dilihat secara visual.

4. KESIMPULAN DAN SARAN

Sistem VQA bahasa Indonesia untuk anak taman kanak-kanak berhasil dikembangkan dengan memanfaatkan model BLIP dan NLLB. Model BLIP berfungsi sebagai inti dari sistem VQA yang dapat menerima masukan berupa gambar dan pertanyaan serta luaran berupa jawaban dari pertanyaan yang diajukan. Sementara itu model NLLB dipergunakan sebagai model penerjemah dari bahasa Indonesia ke dalam bahasa Inggris dan sebaliknya, karena BLIP hanya dapat memproses bahasa Inggris sebagai masukan dan luaran. Hasil pengujian dengan enam jenis pertanyaan menunjukkan bahwa untuk jenis pertanyaan dengan jawaban Ya/ Tidak, kata benda, kata kerja, numeral memperoleh nilai ketepatan jawaban 100, sementara untuk kata keterangan 87,7, dan kata sifat sebesar 62,5.

Saran pengembangan selanjutnya adalah dengan melakukan pengujian sistem dari sisi *Human Computer Interaction* (HCI) dan menyesuaikan tampilan aplikasi dengan pengguna anak-anak. Dari sisi model dapat dilakukan pelatihan model BLIP dengan dataset berbahasa Indonesia agar dapat menghilangkan proses penerjemahan pada sistem.

DAFTAR PUSTAKA

[1] Yusuf, "Masuki Era Revolusi Industri 4.0, Indonesia Perlu Manfaatkan Teknologi Digital," Kominfo, Jakarta, 2020.

[2] I. Solehudin, "Pendidikan Online di Indonesia Masuk 10 Besar Dunia," Jawa Pos, 2017.

[3] M. Rusli, D. Hermawan and N. N. Supurwiningsih, Memahami E-learning: Konsep, Teknologi, dan Arah Perkembangan, Yogyakarta: ANDI, 2020.

[4] X. He, Z. Cai, W. WEi, Y. Zhang, L. Mou, E. Xing and P. Xie, "Towards Visual Question Answering on Pathology Images," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021.

[5] S. Lobry, D. Marcos, J. Murray and D. Tuia, "RSVQA: Visual Question Answering for Remote Sensing Data," *IEEE Transaction on Geoscience and Remote Sensing,* vol. 58, no. 12, pp. 8555-8566, 2020.

[6] A. H. Siregar, "Visual question answering (VQA) untuk objek pariwisata monas menggunakan deep learning," 2019.

[7] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Computer Vision and Image Understanding,* vol. 163, pp. 3-20, 2017.

[8] S. M. Kamel, S. I. Hassan and L. Elrefaei, "VAQA: Visual Arabic Question answering," *Arabian Journal for Science and engineering,* vol. 48, pp. 10803-10823, 2023.

[9] B. He, M. Xia, X. Yu, P. Jian, H. Meng and Z. Chen, "An Educational Robot System of Visual Question Answering for Preschoolers," in *2nd International Conference on Robotics and Automation Engineering*, 2017.

[10] A. Tewari and J. Canny, "What did Spot Hide? A Question-Answering Game for Preschool Children," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2014.

[11] M. Gupta, P. Asthana and P. Singh, "EDUVI: An Educational-Based Visual Question Answering and Image Captioning System for Enhancing the Knowledge of Primary Level Students," *Research Square,* 2023.

[12] NLLB Team, M. R. Costa-jussa, J. Cross, O. Celebi, M. Elbayad, K. Heafield, K. Heddernan and E. Kalbassi, No Language Left Behind: Scaling Human-Centered Machine Translation, 2022.

[13] Y. Koishekenov, V. Nikoulina and A. Berard, "Memory-efficient NLLB-200: Language-specific Expert Pruning of a Massively Multilingual Machine Translation Model," *arXiv preprint arXiv:2212.09811,* 2022.

[14] J. Li, D. Li, C. Xiong and S. C. H. Hoi, "Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *CoRR,* 2022.

[15] D. Li, J. Li, H. Le, G. Wang, S. Savarese and S. C. H. Hoi, "LAVIS: A One-stop Library for Language-Vision Intelligence," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Toronto, Canada, 2023.