

PERBANDINGAN KINERJA ARSITEKTUR VGG19-LSTM DAN BLIP DALAM *VISUAL QUESTION ANSWERING (VQA)* PADA CITRA MEDIS

PROPOSAL PENELITIAN

Diajukan untuk melengkapi tugas-tugas dan
memenuhi syarat-syarat guna memperoleh gelar Sarjana Komputer

Oleh:

ABDUL HAFIDH
2008107010056



**PROGRAM STUDI INFORMATIKA JURUSAN INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA, BANDA ACEH
JUNI, 2024**

PENGESAHAN PROPOSAL

PERBANDINGAN KINERJA ARSITEKTUR VGG19-LSTM DAN BLIP DALAM *VISUAL QUESTION ANSWERING* (VQA) PADA CITRA MEDIS

Oleh:

Nama : Abdul Hafidh
NPM : 2008107010056
Jurusan : Informatika

Menyetujui:

Pembimbing I

Pembimbing II

Alim Misbullah, S.Si., M.S.
NIP. 198806032019031011

Laina Farsiah, S.Si., M.S.
NIP. 198902032022032004

Mengetahui:

Koordinator Program Studi Informatika
Universitas Syiah Kuala,

Alim Misbullah, S.Si., M.S.
NIP. 198806032019031011

KATA PENGANTAR

Segala puji dan syukur kehadiran Allah SWT yang telah melimpahkan rahmat dan hidayah-Nya kepada kita semua, sehingga penulis dapat menyelesaikan penulisan tugas akhir yang berjudul **“Perbandingan Kinerja Arsitektur VGG19-LSTM dan BLIP dalam Visual Question Answering (VQA) pada Citra Medis”** yang telah dapat diselesaikan sesuai rencana. Penulis banyak mendapatkan berbagai pengarahan, bimbingan, dan bantuan dari berbagai pihak. Oleh karena itu, melalui tulisan ini penulis mengucapkan rasa terima kasih kepada:

1. Dewi Ratna Sari, SE.MM, dan Fadhli, SE.Ak, sebagai kedua orang tua penulis, senantiasa memberikan dukungan penuh terhadap aktivitas dan kegiatan yang dilakukan penulis, baik secara moral maupun material, serta menjadi motivasi terbesar bagi penulis untuk menyelesaikan tugas akhir ini.
2. Prof. Dr. Taufik Fuadi Abidin, S.Si., M.Tech, sebagai Dekan Fakultas MIPA Universitas Syiah Kuala.
3. Dr. Nizamuddin, M.Info.Sc., sebagai Ketua Jurusan Informatika Fakultas MIPA Universitas Syiah Kuala.
4. Bapak Alim Misbullah, S.Si., M.S., selaku Dosen Pembimbing I dan Ketua Program Studi Informatika, telah memberikan bimbingan dan arahan yang berharga kepada penulis, membantu penulis dalam menyelesaikan Tugas Akhir ini.
5. Ibu Laina Farsiah, S.Si., M.S., selaku Dosen Pembimbing II, memberikan bimbingan dan arahan yang sangat dibutuhkan oleh penulis, yang akhirnya membantu penulis menyelesaikan tugas akhir ini.
6. Ibu Sri Azizah Nazhifah, S.Kom., M.Sc., selaku dosen pembahas, telah membantu penulis dan kawan-kawan untuk mempersiapkan ruang lingkup tugas akhir, serta memberikan arahan yang diperlukan.
7. Bapak dr. Muhammad Ansari Adista, M.Pd.Ked., Sp.N., selaku dosen pembahas yang telah memberikan masukan dan saran yang sangat berharga bagi penulis dalam menyelesaikan tugas akhir ini.
8. Ibu Dalila Husna Yunardi, B.Sc., M.Sc., selaku dosen wali, memberikan *support* dan pedoman yang diperlukan selama penulis menempuh pendidikan.
9. Muhhamad Razan Fawwaz, Amar Suhendra, Khairul Umam Albi, Muhhamad Rudy Hidayat, Muhhamad Raja Furqan, Teuku Nabil Muhhamad Dhuha, Yoan Rifqi Candra, dan Haris Daffa, telah menemani dan memberi dukungan kepada penulis selama empat tahun perkuliahan di jurusan Informaika USK.
10. Anas Naufal Al-Kiram, Muhhamad Hanif, Daffa Mudhaffar, Muhhamad Ikhsan Fikri, dan Teuku Muhhamad Roy Adrian, teman yang selalu mengingatkan, dan membantu penulis untuk selalu berusaha dalam menyelesaikan tugas akhir di semester 8.

11. Sahabat dan teman-teman seperjuangan dari Jurusan Informatika USK 2020 lainnya, telah memberikan dukungan moral yang luar biasa selama penulis menempuh pendidikan di jurusan Informatika USK.
12. Seluruh Dosen dan Staf di Jurusan Informatika Fakultas MIPA, atas ilmu dan bimbingan yang diberikan kepada penulis selama perkuliahan, menjadi bagian tak terpisahkan dalam proses pembelajaran penulis.

Penulis mengakui adanya kekurangan dalam tulisan ini, baik dari aspek materi, metode, maupun bahasa yang digunakan. Oleh karena itu, penulis mengundang kritik dan saran yang konstruktif dari para pembaca untuk meningkatkan kualitas Tugas Akhir ini. Penulis berharap bahwa tulisan ini dapat memberikan manfaat bagi banyak orang dan berkontribusi pada kemajuan ilmu pengetahuan.

DAFTAR ISI

	<i>Halaman</i>
Halaman Judul	i
Halaman Pengesahan	ii
Kata Pengantar	iii
Daftar Isi	v
Daftar Tabel	vii
Daftar Gambar	viii
 BAB I PENDAHULUAN	 1
1.1. Latar Belakang	1
1.2. Rumusan Masalah	2
1.3. Tujuan Penelitian	2
1.4. Manfaat Penelitian	3
 BAB II TINJAUAN KEPUSTAKAAN	 4
2.1. Citra Medis	4
2.2. <i>Deep Learning</i>	5
2.2.1. <i>Artificial Neural Network</i>	6
2.2.2. Fungsi Aktivasi	6
2.2.3. Fungsi <i>loss</i>	8
2.2.4. Fungsi Optimasi	9
2.3. <i>Convolutional Neural Network</i>	10
2.4. <i>Long Short Term Memory</i>	11
2.5. <i>Transfer Learning</i>	11
2.6. Matriks Evaluasi	15
2.6.1. Akurasi	16
2.6.2. <i>Bilingual Evaluation Understudy</i>	16
2.7. <i>Visual Question Answering</i>	18
2.7.1. <i>Computer Vision</i>	19
2.7.2. <i>Natural Language Processing</i>	19
2.8. <i>Medical Visual Question Answering Dataset</i>	20
2.8.1. <i>PathVQA Dataset</i>	20
2.8.2. <i>VQA-RAD Dataset</i>	20
2.9. Penelitian Terkait	21
 BAB III METODOLOGI PENELITIAN	 23
3.1. Waktu dan Lokasi Penelitian	23
3.2. Jadwal Pelaksanaan	23
3.3. Alat dan Bahan	23
3.3.1. Perangkat Keras	23
3.3.2. Perangkat Lunak	23
3.4. Metode Penelitian	24
3.4.1. Identifikasi Masalah	25
3.4.2. Studi Literatur	25

3.4.3. Pengumpulan Data	26
3.4.4. Pemrosesan Data	27
3.4.5. Membangun Model	28
3.4.6. Melatih Model.....	28
3.4.7. Perbandingan Hasil	29
3.4.8. Membangun Sistem Medis Cerdas Berbasis Web	29
DAFTAR PUSTAKA	30

DAFTAR TABEL

	<i>Halaman</i>
Tabel 2.1 Deskripsi dari masing-masing nilai BLEU.....	17
Tabel 3.1 Jadwal pelaksanaan penelitian	23
Tabel 3.2 Proses <i>case folding</i>	27
Tabel 3.3 Proses <i>remove punctuation</i>	27
Tabel 3.4 Proses <i>stemming</i>	28
Tabel 3.5 Proses tokenisasi	28

DAFTAR GAMBAR

	<i>Halaman</i>
Gambar 2.1 Kategori pendekatan <i>deep learning</i>	5
Gambar 2.2 Perbandingan teknik optimasi Adam dengan teknik optimasi lainnya (Kingma and Ba, 2014).....	9
Gambar 2.3 Arsitektur CNN (Wardani and Leonardi, 2023).....	10
Gambar 2.4 Arsitektur LSTM (Luo et al., 2023).....	11
Gambar 2.5 Konsep penggunaan <i>transfer learning</i> (Mukhlif et al., 2023)	12
Gambar 2.6 Arsitektur VGG19 (Nguyen et al., 2022)	13
Gambar 2.7 Arsitektur BLIP (Li et al., 2022).....	13
Gambar 2.8 <i>Confusion matrix</i> (Kulkarni et al., 2020)	15
Gambar 2.9 Vanilla VQA <i>network</i> model (Srivastava et al., 2021).....	18
Gambar 3.1 Diagram alir penelitian.....	25
Gambar 3.2 Sampel data PathVQA	26
Gambar 3.3 Sampel data VQA-RAD	26

BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Citra medis, seperti *Magnetic Resonance Imaging* (MRI), *X-Ray*, *Ultrasonography* (USG), *Endoscopy*, *Computed Tomography* (CT-Scan), *Nuclear Medicine*, dan lain-lain, menjadi fokus penelitian utama dalam dunia medis (Kusuma and Kusumadewi, 2020). Profesional medis menggunakan berbagai teknik untuk mendeteksi dan menganalisis penyakit pada pasien. Dua cabang ilmu yang berkaitan dengan diagnosa penyakit, yaitu patologi dan radiologi, memainkan peran krusial dalam memahami citra medis (Sorace et al., 2012). Meskipun demikian, kesalahan analisis citra medis oleh tenaga medis terkadang bisa saja terjadi karena sifat manusia yang rentan terhadap kesalahan (Mauli, 2018).

Ketika manusia melakukan pekerjaan yang repetitif, ini dapat menyebabkan kelelahan dan berkurangnya konsentrasi, sehingga meningkatkan kemungkinan kesalahan. Hal ini berbeda dengan *Artificial Intelligence* (AI), AI tidak memiliki perasaan, sehingga AI akan tetap menghasilkan kualitas pekerjaan yang konsisten seiring berjalannya waktu (Fernando and Harsiti, 2019). Pada bidang medis AI tidak diharapkan untuk menggantikan dokter manusia dalam skala besar. Sebaliknya, AI kemungkinan akan memberdayakan praktik kedokteran dengan meningkatkan upaya dokter dan mengatasi masalah seperti kelelahan dokter (Basu et al., 2020). Oleh sebab itu, dibutuhkan sebuah sistem yang dapat membantu tenaga medis dalam menjawab permasalahan yang terdapat pada citra medis. Salah satu solusi yang dapat digunakan adalah dengan membangun sistem *Visual Question Answering* (VQA). Dalam konteks medis VQA dapat memberikan manfaat bagi dokter dan pasien yang mana dokter bisa memperoleh jawaban yang diperoleh dari sistem VQA sebagai bentuk dalam pengambilan keputusan. Sedangkan pasien bisa mengajukan pertanyaan ke sistem VQA dengan gambar medis yang ada pada dirinya untuk mengetahui kondisi kesehatannya (Nguyen et al., 2019).

Sistem VQA medis ini juga dapat berperan sebagai asisten yang memiliki pengetahuan yang luas. Sebagai contoh, pendapat tambahan atau opini kedua dari sistem VQA dapat membantu tenaga medis dalam menjawab pertanyaan berdasarkan citra medis yang diberikan dan mengurangi potensi kesalahan dalam diagnosis pada saat bersamaan (Tschandl et al., 2020). Meskipun begitu, penelitian mengenai sistem VQA medis masih terbatas karena menghadapi beberapa tantangan khusus yang harus dihadapi, seperti keanekaragaman pertanyaan yang dapat diajukan, yang memerlukan pemahaman yang mendalam tentang citra medis dan pertanyaan untuk konteks medis,

serta memiliki keterbatasan dalam interpretabilitas yakni kemampuan untuk menanyakan model tentang wilayah gambar yang spesifik (Lin et al., 2023). Akan tetapi, dengan sering dilakukannya penelitian terkait *medical visual question answering*, diharapkan dapat mengatasi tantangan yang ada, sehingga penelitian *medical visual question answering* dapat berkembang dengan baik seiring berjalannya waktu.

Penelitian ini bertujuan untuk mengembangkan sebuah sistem VQA yang mampu memberikan jawaban terhadap pertanyaan yang diajukan berdasarkan citra medis. Sistem ini akan memanfaatkan dua *dataset* yaitu PathVQA yang terkait dengan patologi, dan VQA-RAD yang fokus pada bidang radiologi. Dalam penelitian ini akan menggunakan pendekatan *deep learning* yang memanfaatkan teknik *transfer learning*. Teknik *transfer learning* ini digunakan untuk memanfaatkan model yang sudah dilatih sebelumnya pada *dataset* yang berbeda. Hasil dari penelitian ini diharapkan dapat memberikan gambaran mengenai sistem VQA untuk citra medis dengan *deep learning* dan memberikan kontribusi terkait penelitian *medical visual question answering*.

1.2 RUMUSAN MASALAH

Berdasarkan latar belakang yang telah diuraikan, permasalahan dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana membangun model VQA menggunakan citra medis dengan menerapkan teknik *transfer learning*?
2. Bagaimana membandingkan performa model VQA pada arsitektur VGG19-LSTM dan BLIP untuk *dataset* PathVQA dan VQA-RAD?
3. Bagaimana menerapkan model VQA terbaik untuk menjawab pertanyaan berdasarkan citra medis?

1.3 TUJUAN PENELITIAN

Adapun maksud dan tujuan dari penelitian ini adalah sebagai berikut:

1. Membangun model VQA menggunakan citra medis dengan menerapkan teknik *transfer learning*.
2. Membandingkan performa model VQA pada arsitektur VGG19-LSTM dan BLIP, dengan *dataset* PathVQA dan VQA-RAD.
3. Mengimplementasikan model VQA terbaik untuk menjawab pertanyaan berdasarkan citra medis.

1.4 MANFAAT PENELITIAN

Manfaat yang diinginkan dari penelitian ini adalah untuk memberikan gambaran mengenai sistem VQA untuk citra medis dengan *deep learning*. Selanjutnya, memberikan pengetahuan hasil implementasi dengan arsitektur VGG19-LSTM dan BLIP dalam mengembangkan sistem VQA pada citra medis. Lalu yang terakhir, memberikan kontribusi terkait penelitian *medical visual question answering*.

BAB II

TINJAUAN KEPUSTAKAAN

2.1 CITRA MEDIS

Citra adalah gambaran dari suatu objek yang bisa kita lihat. Citra analog tidak bisa disimpan secara langsung dalam komputer. Oleh karena itu, kita perlu mengubah citra analog menjadi citra digital agar bisa diolah oleh komputer. Citra digital adalah citra yang bisa diolah menggunakan perangkat komputer. Alasan utama citra analog tidak bisa diolah oleh komputer adalah karena citra analog tidak memiliki konsep sampling dan kuantisasi. Konsep sampling adalah suatu metode yang mengubah citra analog menjadi *grid* berbentuk M baris dan N kolom, sehingga citra menjadi lebih tersegmentasi. Semakin besar nilai M dan N, semakin halus citra digital yang dihasilkan. Sedangkan konsep kuantisasi adalah suatu metode yang mengubah intensitas dari citra analog menjadi intensitas diskrit. Baik sampling maupun kuantisasi memegang peranan penting dalam mengambil potongan citra menjadi bentuk M baris dan N kolom (proses sampling) serta menentukan nilai intensitas yang terdapat pada setiap titik tersebut (proses kuantisasi). Hasil akhir dari kedua konsep ini adalah citra yang memiliki resolusi sesuai dengan yang kita inginkan (Andono et al., 2018).

Dalam ilmu pengolahan citra, beragam jenis citra ada, dan salah satunya adalah citra *Red Green Blue* (RGB), juga disebut sebagai citra *true color*. Citra ini memiliki matriks data berukuran $m \times n \times 3$, yang menggambarkan warna merah, hijau, dan biru pada setiap piksel. Rentang nilai untuk setiap warna adalah antara 0 (nilai minimum) hingga 255 (nilai maksimum) dalam monitor komputer. Pada bagian ini, skala 0 hingga 255 dipilih berdasarkan representasi delapan digit dalam bilangan biner yang digunakan oleh komputer. Dengan demikian, lebih dari 16 juta warna dapat dihasilkan secara total. Penentuan warna pada setiap piksel didasarkan pada tingkat kecerahan merah, hijau, dan biru (Ardiansyah, 2013).

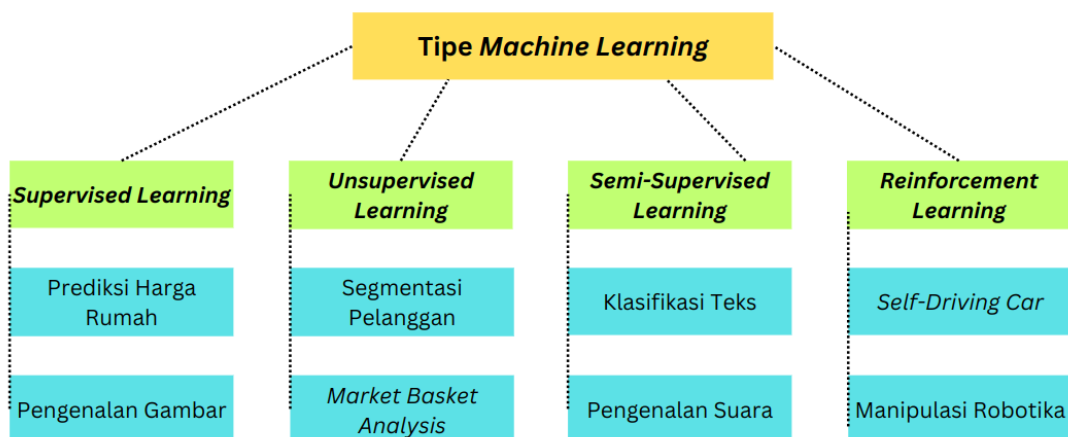
Di bidang kesehatan, pencitraan medis memegang peranan penting dalam berbagai bidang klinis seperti prosedur medis yang digunakan untuk deteksi dini, pemantauan, diagnosis, dan evaluasi pengobatan berbagai kondisi medis (Puttagunta and Ravi, 2021). Citra medis merujuk pada gambar dua dimensi yang menggambarkan struktur internal tubuh manusia, dimanfaatkan oleh profesional kesehatan untuk diagnosis penyakit. Pengolahan citra medis memiliki aplikasi luas, termasuk deteksi tumor atau kanker pada organ reproduksi wanita, identifikasi patologi pada organ-organ seperti paru-paru, hati, dan tulang, segmentasi struktur tulang dari jaringan otot, klasifikasi gigi, serta analisis gambar dari mikroskop. Berbagai teknologi digunakan dalam pencitraan medis ini, antara lain *Magnetic Resonance Imaging* (MRI), *X-Ray*,

Ultrasonography (USG), endoskopi, *Computed Tomography* (CT-Scan), dan *Nuclear Medicine* (Kusuma and Ellyana, 2018).

2.2 DEEP LEARNING

Deep learning adalah cabang ilmu *machine learning* yang berbasis pada *Artificial Neural Network* (ANN) sebagai pengembangannya. Metode dalam *machine learning* biasanya hanya mengandalkan alat seperti *Central Processing Unit* (CPU) dan *Random Access Memory* (RAM) dalam menentukan kecepatan komputasi. Sedangkan metode *deep learning* selain menggunakan CPU dan RAM, metode ini juga menggunakan *Graphics Processing Unit* (GPU) sehingga proses komputasi data yang besar dapat dilakukan dengan cepat (Ilahiyah and Nilogiri, 2018).

Deep learning memiliki pendekatan yang dapat dikategorikan seperti *supervised learning*, *unsupervised learning*, *semi-supervised learning*, dan *reinforcement learning* (Alom et al., 2019). Kategori pendekatan *deep learning* dapat dilihat pada Gambar 2.1.



Gambar 2.1. Kategori pendekatan *deep learning*

Pendekatan ini melibatkan berbagai teknik dan algoritma untuk menyelesaikan berbagai jenis masalah. Dalam mengimplementasikan model *deep learning*, beberapa *hyperparameter* yang sangat penting untuk diperhatikan adalah *learning rate*, *batch size*, jumlah *epochs*, dan arsitektur jaringan itu sendiri. *Learning rate* adalah ukuran yang menentukan seberapa besar langkah yang diambil model saat memperbarui bobot selama pelatihan. *Batch size* mengacu pada jumlah sampel data yang diproses sebelum memperbarui parameter model, sementara jumlah *epochs* menunjukkan berapa kali seluruh *dataset* akan dilalui selama pelatihan. Pemilihan *hyperparameter* sering kali dilakukan dengan fleksibilitas berdasarkan eksperimen dan kebutuhan spesifik dari proyek yang sedang dikerjakan. Tidak ada aturan tetap mengenai nilai-nilai spesifik yang harus dipilih untuk *hyperparameter*. Seorang peneliti dapat mencoba berbagai nilai

untuk menemukan kombinasi yang optimal. Metode seperti *grid search* menunjukkan bahwa nilai-nilai *hyperparameter* dapat dipilih secara bebas dalam rentang yang telah ditentukan (Bergstra and Bengio, 2012).

Grid search adalah metode pencarian *hyperparameter* yang melibatkan pencarian melalui ruang *hyperparameter* dengan mencoba setiap kombinasi yang mungkin dari nilai *hyperparameter* yang telah ditentukan sebelumnya. Dengan *grid search*, seorang peneliti membuat sebuah *grid* dari *hyperparameter* yang berbeda dan mengevaluasi model dengan setiap kombinasi yang ada di *grid* tersebut (Bergstra and Bengio, 2012). Pendekatan ini memastikan bahwa semua kombinasi yang mungkin diuji, meskipun bisa sangat memakan waktu dan sumber daya.

Pada bidang medis *deep learning* digunakan untuk memproses citra medis seperti X-rays, CT, scan MRI (*Magnetic Resonance Imaging*) dan lain-lain untuk mendiagnosa kondisi kesehatan (Kelleher, 2019).

2.2.1 Artificial Neural Network

Artificial Neural Network (ANN) atau jaringan saraf tiruan adalah jaringan saraf yang memproses informasi dengan cara yang mirip dengan otak manusia (Kristiyanti and Saputra, 2023). *Neural network* terdiri dari elemen pemrosesan sederhana yang disebut *node* yang saling terhubung, mirip dengan cara kerja neuron dalam otak manusia. Kemampuan untuk melakukan pemrosesan dalam jaringan ini disimpan dalam koneksi antara *node*, yang biasanya disebut sebagai *weight*. Nilai-nilai *weight* ini diperoleh melalui proses pembelajaran atau adaptasi yang berdasarkan pada pola data yang dipelajari oleh ANN (Gurney, 1997).

2.2.2 Fungsi Aktivasi

Fungsi aktivasi dalam konteks jaringan saraf dapat diibaratkan dengan cara tubuh manusia merespon rangsangan dari lingkungan. Ketika seseorang menerima rangsangan eksternal, tubuhnya secara otomatis meresponsnya. Sebagai contoh, ketika tangan kita digigit, tubuh kita akan merespons dengan menolak atau melepaskan gigitan tersebut. Respons tubuh ini akan semakin intens jika rangsangan yang diterima semakin kuat. Dalam konteks algoritma jaringan saraf, respons tubuh ini analoginya digantikan oleh nilai bobot dan tingkat aktivasi yang tinggi. Pada persamaan 2.1 menunjukkan persamaan dari jaringan saraf sebelum menggunakan fungsi aktivasi.

$$y = \sum_{i=1}^n x_i w_i + b \quad (2.1)$$

- y adalah nilai keluaran dari jaringan saraf.

- x_i adalah nilai *input* dari jaringan saraf.
- w_i adalah bobot dari setiap nilai *input*.
- b adalah bias dari jaringan saraf.

Ketika menerapkan fungsi aktivasi pada persamaan 2.1, maka persamaan akan berubah menjadi seperti yang ditunjukkan pada persamaan 2.2.

$$z = Act\left(\sum_{i=1}^n x_i w_i + b\right) \quad (2.2)$$

- z adalah nilai keluaran dari jaringan saraf setelah menggunakan fungsi aktivasi.
- Act adalah fungsi aktivasi.
- x_i adalah nilai *input* dari jaringan saraf.
- w_i adalah bobot dari setiap nilai *input*.
- b adalah bias dari jaringan saraf.

Persamaan 2.2 ini menunjukkan bahwa nilai keluaran z akan bergantung pada fungsi aktivasi terhadap nilai prediksi yang dihasilkan dari perkalian nilai *input* dan bobot (Kristiyanti and Saputra, 2023).

Fungsi aktivasi adalah komponen penting dalam jaringan saraf tiruan yang mengatur bagaimana nilai hasil penjumlahan terbobot dari *input* data diubah menjadi *output* yang dikeluarkan oleh neuron dalam jaringan saraf. Fungsi aktivasi ini digunakan untuk menentukan apakah suatu neuron akan meneruskan nilai kalkulasinya ke neuron berikutnya, berdasarkan suatu nilai ambang tertentu. Fungsi ini juga sering disebut sebagai fungsi transfer karena memiliki kemampuan untuk mengubah data yang dihasilkan dari hasil penjumlahan terbobot pada suatu lapisan, yang kemudian akan diteruskan ke lapisan selanjutnya. Fungsi aktivasi bisa berupa fungsi linear atau fungsi non-linear tergantung pada tugas yang ingin diselesaikan, dan fungsi aktivasi ini dapat digunakan dalam berbagai hal seperti pengenalan objek dan klasifikasi (Nwankpa et al., 2018).

Fungsi aktivasi perlu memiliki sifat diskriminatif, yang merupakan aspek yang penting karena memungkinkan penggunaan proses propagasi balik kesalahan dalam pelatihan jaringan. Salah satu fungsi aktivasi yang umum digunakan dalam konteks CNN adalah fungsi ReLU (*Rectified Linear Unit*). Fungsi ini merupakan fungsi aktivasi yang mengubah seluruh isi nilai *input* menjadi angka positif (Alzubaidi et al., 2021). Persamaan fungsi aktivasi ini dapat dilihat pada persamaan 2.3.

$$f(x)_{\text{ReLU}} = \max(0, x) \quad (2.3)$$

Salah satu fungsi aktivasi yang digunakan dalam berbagai model mutakhir seperti GPT-3, BERT, dan sebagian besar model Transformer lainnya adalah *Gaussian Error Linear Unit* (GELU). Fungsi ini merupakan fungsi aktivasi yang menimbang *input* berdasarkan persentilnya, daripada mengelompokkan *input* berdasarkan tandanya seperti ReLU. Oleh karena itu, GELU dapat dikatakan lebih mulus dibandingkan ReLU. Hal ini memungkinkan GELU untuk lebih mudah memperkirakan fungsi yang rumit daripada ReLU atau ELU (Hendrycks and Gimpel, 2016). Persamaan fungsi aktivasi ini dapat dilihat pada persamaan 2.4.

$$\text{GELU}(x) = xP(X \leq x) = x\Phi(x) = x \cdot \frac{1}{2}[1 + \text{erf}(x/\sqrt{2})] \quad (2.4)$$

- x adalah nilai *input* dari fungsi aktivasi.
- Φ adalah *Cumulative Distribution Function* (CDF) dari x .
- erf adalah fungsi kesalahan dari x .

Salah satu fungsi aktivasi yang dapat digunakan untuk mengklasifikasi lebih dari dua kelas adalah fungsi aktivasi *softmax*. Persamaan 2.5 menunjukkan persamaan dari fungsi aktivasi *softmax*.

$$f_j(Z) = \frac{e^{z_j}}{\sum_k e^{z_k}} \quad (2.5)$$

Pada persamaan 2.5, Notasi f_j menunjukkan hasil fungsi pada setiap elemen ke- j pada vektor keluaran kelas. Argumen z merupakan hipotesis yang diberikan oleh model pelatihan agar dapat diklasifikasi oleh fungsi *softmax* (Ilahiyah and Nilogiri, 2018).

2.2.3 Fungsi loss

Fungsi *loss* adalah salah satu fungsi pada *Artificial Neural Network* (ANN) untuk melakukan perhitungan nilai *error* atau kesalahan dari hasil prediksi dari suatu *output* ANN (Zhang, 2016).

Fungsi *loss* yang menghukum kesalahan probabilitas *false negative* daripada *false positive* adalah *categorical cross entropy* (Ho and Wookey, 2019). Persamaan 2.6 menunjukkan persamaan dari *categorical cross-entropy*.

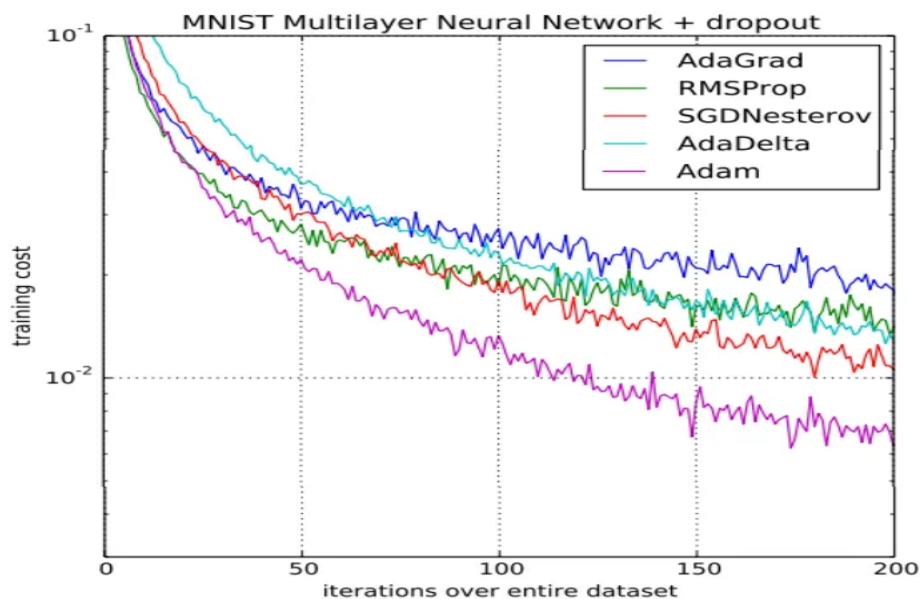
$$J_{cce} = -\frac{1}{M} \sum_{k=1}^K \sum_{m=1}^M y_m^k \log_{(\theta)}(x_m, k) \quad (2.6)$$

- M adalah jumlah sampel data yang digunakan untuk pelatihan.
- K adalah jumlah kelas yang ada pada data.
- y_m^k adalah nilai target dari sampel data ke- m pada kelas ke- k .
- x adalah *input* untuk contoh pelatihan ke- m .
- H_θ adalah bobot model *neural network* θ .

2.2.4 Fungsi Optimasi

Fungsi optimasi atau *optimization function* dapat diartikan sebagai suatu fungsi yang berperan sebagai *black box*, dimana fungsi ini menerima kesalahan sebagai input dan menghasilkan nilai bobot yang optimal untuk suatu model ANN (Li and Malik, 2017).

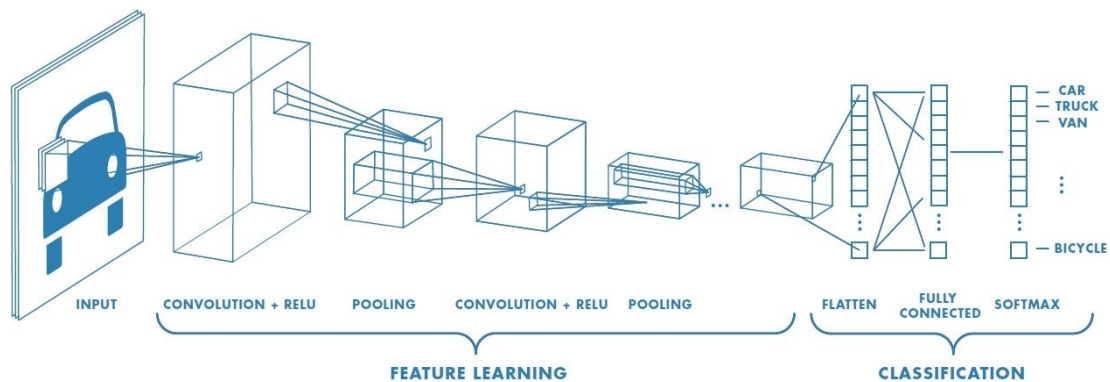
Salah satu dari fungsi optimasi yang dapat digunakan dalam pengembangan model *deep learning* adalah Adam (*Adaptive Moment Estimation*). Adam adalah teknik optimasi untuk *gradient descent*. Metode ini sangat efisien saat bekerja dengan masalah yang melibatkan banyak data atau parameter. Algoritma ini membutuhkan memori yang lebih sedikit dan efisien. Secara intuitif, Adam merupakan gabungan antara algoritma *stochastic gradient descent momentum* dan *RMSProp*. Secara eksperimen Adam adalah teknik optimasi terbaik dengan *training cost* yang rendah menurut (Kingma and Ba, 2014). Gambar 2.2 menunjukkan perbandingan teknik optimasi Adam dengan teknik optimasi lainnya.



Gambar 2.2. Perbandingan teknik optimasi Adam dengan teknik optimasi lainnya (Kingma and Ba, 2014)

2.3 CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network (CNN) adalah turunan dari algoritma *neural network* yang dikhususkan untuk memproses data yang berupa gambar. CNN adalah algoritma yang meniru proses pengolahan visual yang terjadi pada manusia. Seperti mata manusia yang berfungsi sebagai alat *input*, CNN menggunakan lapisan konvolusi yang terdiri dari miliaran neuron untuk memproses informasi visual dan menghasilkan prediksi terhadap objek yang diamati (Kristiyanti and Saputra, 2023). Algoritma CNN dirancang dengan neuron yang berfungsi mirip dengan cara area penglihatan pada otak manusia dan hewan bekerja, seperti yang dijelaskan oleh (Henningsen-Schomers and Pulvermüller, 2022). Arsitektur ini terdiri dari sejumlah lapisan, yang biasanya disebut sebagai blok-blok multi-bangunan dan dapat dilihat pada Gambar 2.3.



Gambar 2.3. Arsitektur CNN (Wardani and Leonardi, 2023)

Pada Gambar 2.3 dapat dilihat bahwa konvolusi merupakan langkah awal dalam pengolahan gambar yang digunakan untuk mengekstraksi fitur penting dari gambar *input*. Dalam konvolusi, hubungan antar piksel dipertahankan dengan cara mengoperasikan kotak kecil pada masukan untuk memahami fitur-fitur gambar. Konvolusi melibatkan operasi matematika linear yang mencakup perkalian antara matriks gambar dan filter (*kernel*) yang merupakan matriks bobot dua dimensi.

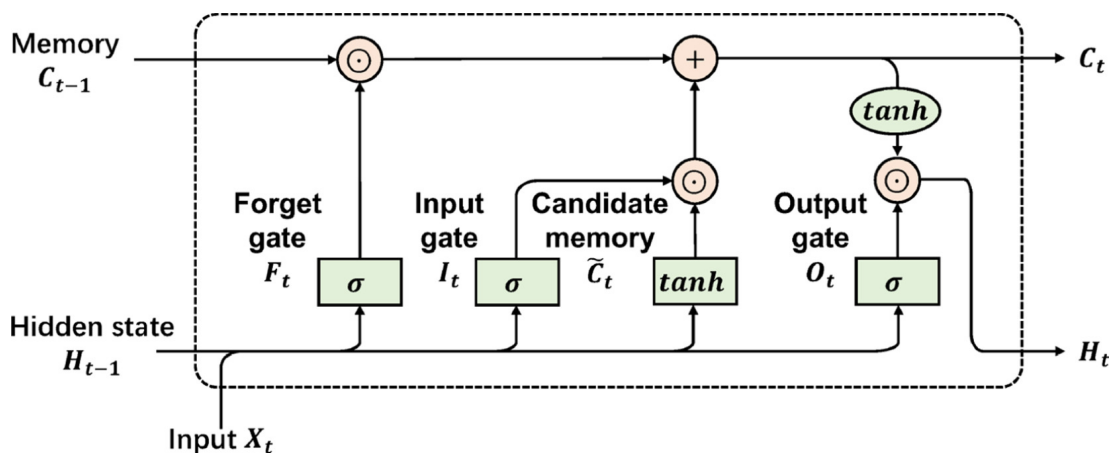
Fungsi dari lapisan *pooling* adalah untuk secara bertahap mengurangi ukuran representasi spasial gambar, sehingga mengurangi jumlah parameter yang dibutuhkan dalam jaringan. Lapisan *pooling* biasanya ditempatkan di antara lapisan-lapisan konvolusi. Lapisan ini beroperasi secara independen pada setiap peta fitur (Monedero et al., 2021).

CNN merupakan terobosan besar dalam pengenalan gambar. Mereka belajar langsung dari data gambar, menggunakan pola untuk mengklasifikasikan gambar dan menghilangkan kebutuhan ekstraksi fitur manual. Saat ini, CNN merupakan topik yang menarik dalam *machine learning*, dan telah menunjukkan kinerja yang sangat baik dalam klasifikasi (Khan et al., 2020).

2.4 LONG SHORT TERM MEMORY

Long Short Term Memory (LSTM) adalah tipe dari *Recurrent Neural Network* (RNN) yang diciptakan untuk menangani data yang bersifat *sequential* seperti data *time series*, *speech*, dan *text*. LSTM ini dikembangkan untuk mengatasi masalah *vanishing gradient* yang ada dalam RNN tradisional, yang membuatnya sulit bagi jaringan untuk mempelajari ketergantungan jangka panjang (Brownlee, 2017).

Menurut (Alom et al., 2019) LSTM adalah model jaringan saraf yang menggunakan *cell state* untuk menyimpan informasi dari *input* sebelumnya. *Cell state* memiliki tiga gerbang: *input gate* (i_t), *forget gate* (f_t), dan *output gate* (o_t) yang bisa dilihat pada Gambar 2.4.



Gambar 2.4. Arsitektur LSTM (Luo et al., 2023)

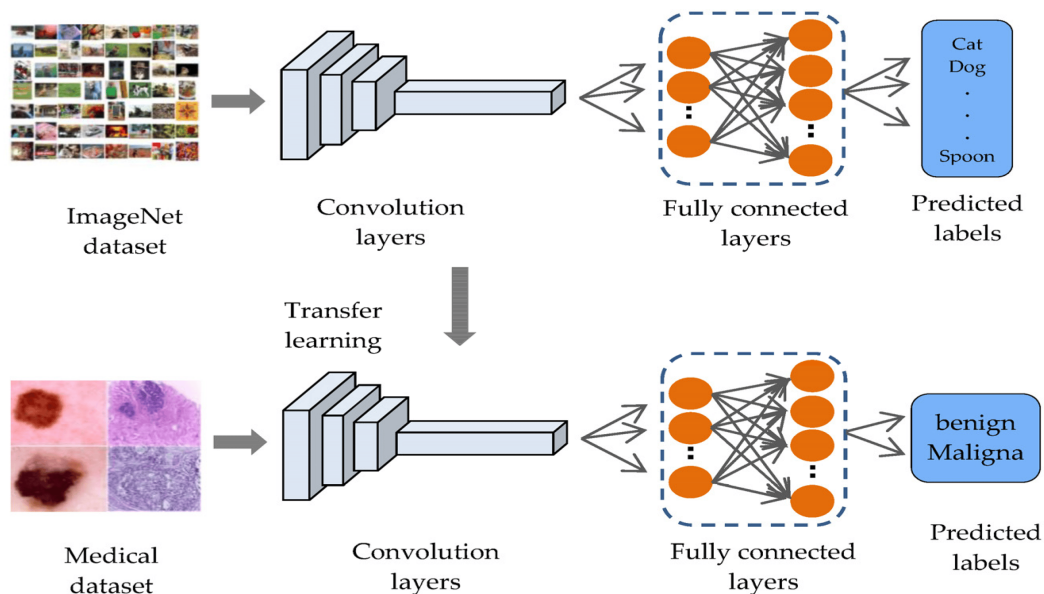
Input gate (i_t) digunakan untuk mengontrol pengaruh data yang masuk saat ini terhadap bobot unit tersebut, *forget gate* (f_t) bertujuan untuk mengendalikan pengaruh riwayat informasi pada bobot unit saat ini, *output gate* (o_t) bertujuan dalam mengendalikan ekspor nilai keadaan unit memori (Huang et al., 2021).

2.5 TRANSFER LEARNING

Transfer learning digunakan untuk meningkatkan pembelajaran dari satu domain dengan mentransfer informasi dari domain terkait. Kita dapat mengambil pengetahuan dunia nyata non-teknis untuk memahami mengapa *transfer learning* memungkinkan. Pertimbangkan contoh dua orang yang ingin belajar bermain piano. Satu orang tidak memiliki pengalaman sebelumnya dalam bermain musik, dan orang lain memiliki pengetahuan musik yang luas melalui bermain gitar. Orang dengan latar belakang musik yang luas akan dapat belajar piano dengan lebih efisien dengan mentransfer pengetahuan musik yang sudah dipelajari sebelumnya ke tugas belajar bermain piano. Satu orang dapat mengambil informasi dari tugas yang sudah dipelajari sebelumnya dan

menggunakannya secara bermanfaat untuk belajar tugas yang terkait (Pan and Yang, 2009).

Transfer learning pada *Artificial Neural Network* (ANN) bisa dibayangkan seperti seseorang yang belajar menjadi lebih mudah, cepat, dan akurat dalam memahami tugas dan konsep baru jika mereka sudah memiliki pengalaman belajar yang serupa dengan konsep baru yang ingin dipelajari. Ini mirip dengan bagaimana seorang individu dapat lebih mudah memahami fisika setelah belajar matematika atau bagaimana seseorang dapat lebih lancar mengendarai truk setelah menguasai kemampuan mengemudi mobil. Pada dasarnya *transfer learning* terjadi ketika pemahaman tentang suatu konteks dipengaruhi oleh pemahaman sebelumnya tentang konteks yang mirip (Cireşan et al., 2012). Intinya adalah bahwa *transfer learning* memungkinkan kita untuk menggunakan pengetahuan yang sudah ada untuk memecahkan masalah baru seperti yang ditunjukkan pada Gambar 2.5.

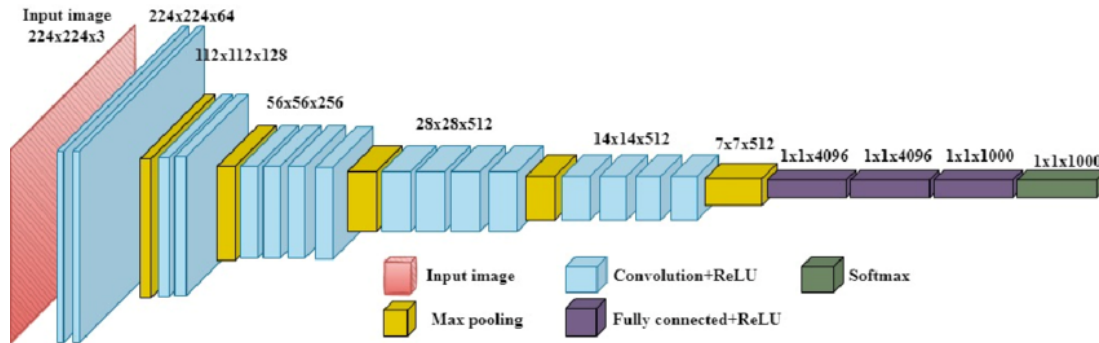


Gambar 2.5. Konsep penggunaan *transfer learning* (Mukhlif et al., 2023)

Model dasar yang digunakan dalam *transfer learning* adalah *pre-trained* model, dimana bobot di seluruh jaringan saraf tiruan sudah disesuaikan untuk data tertentu. Ini memungkinkan model untuk memiliki pemahaman yang lebih mendalam tentang fitur dasar dan fitur tingkat tinggi yang dapat mempercepat proses pelatihan. Secara konsisten, model jaringan saraf yang telah dilatih sebelumnya memberikan hasil prediksi yang lebih tepat dibandingkan dengan jaringan saraf yang dimulai dengan bobot-bobot yang diinisiasi secara acak dalam konteks masalah yang melibatkan data yang sudah memiliki label (Cireşan et al., 2012).

VGG19 merupakan salah satu model *pre-trained* yang digunakan dalam

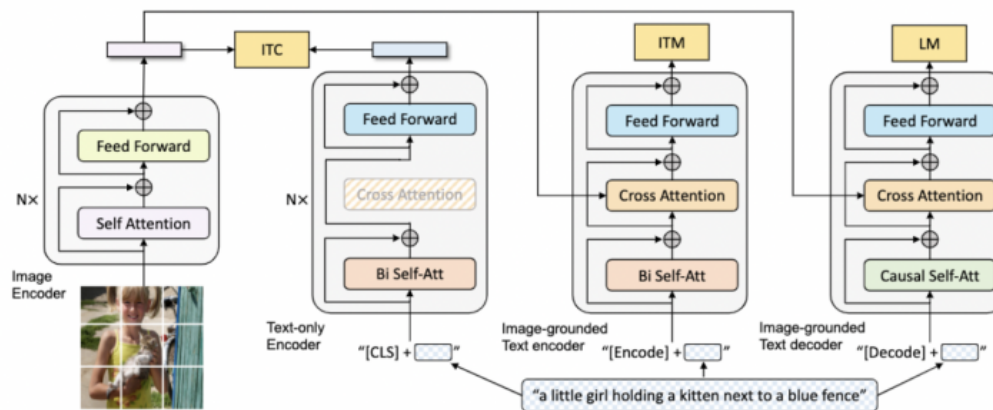
penelitian ini. VGG19 adalah model CNN yang memiliki 19 lapisan. Model ini dilatih pada *dataset* ImageNet yang memiliki 1000 kelas dan 1,2 juta gambar. Model ini memiliki 16 lapisan konvolusi dan 3 lapisan *fully connected*. Model ini memiliki 138 juta parameter dan 20 miliar operasi (Simonyan and Zisserman, 2014). Arsitektur VGG19 dapat dilihat pada Gambar 2.6.



Gambar 2.6. Arsitektur VGG19 (Nguyen et al., 2022)

Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (BLIP) adalah sebuah kerangka dari *Vision Language Pre-Training* (VLP). BLIP mampu melakukan berbagai tugas *multimodal* seperti *Visual Question Answering* (VQA), *image captioning*, dan lain-lain. Dalam hal ini, BLIP mengusulkan sebuah model *multimodal mixture of encoder-decoder* yang dapat beroperasi secara fleksibel sebagai *encoder* dan *decoder* untuk berbagai tugas penglihatan bahasa. BLIP mampu memajukan *pre-training* terunifikasi untuk transfer pembelajaran lintas tugas *vision-language*. Berikut ini dapat dilihat arsitektur pada BLIP pada Gambar 2.7.

A unified model for vision-language understanding and generation



Gambar 2.7. Arsitektur BLIP (Li et al., 2022)

BLIP memiliki tiga fungsionalitas yang dapat dilihat sebagai berikut:

1. *Unimodal encoder* yang memisahkan proses *encode* gambar dan teks. *Image encoder* didasarkan pada *Vision Transformer* (ViT), sementara *encoder* teks mirip dengan *Bidirectional Encoder Representation from Transformer* (BERT). Disini, terdapat token khusus [CLS] yang ditambahkan di awal *input* teks untuk merangkum kalimat.
2. *Image-grounded text encoder* yang bertujuan memasukkan informasi visual ke dalam *encoder* teks. Pada bagian ini dilakukan dengan menyisipkan lapisan *cross-attention* antara lapisan *self-attention* dan jaringan *feed-forward* untuk setiap blok transformer dari *encoder* teks. Pada bagian ini terdapat token khusus yaitu [Encode] yang ditambahkan ke teks, dan *embedding output* dari [Encode] ini digunakan sebagai representasi multimodal dari pasangan gambar dan teks.
3. *Image grounded text decoder* memodifikasi *encoder* teks dengan menggantikan lapisan *bidirectional self-attention* dengan lapisan *causal self-attention*. Di bagian ini, token khusus [Decode] digunakan untuk menandai awal dari sebuah urutan.

Pada gambar 2.7 dijelaskan tiga tujuan dari arsitektur BLIP. Penjelasan tiga tujuan BLIP dapat dilihat sebagai berikut:

1. *Image-Text Contrastive Loss* (ITC) bertujuan mengaktifkan *unimodal encoder*. ITC mendorong keselarasan dalam ruang fitur transformer visual dan transformer teks dengan memastikan pasangan gambar dan teks yang positif memiliki representasi yang mirip dibandingkan dengan pasangan negatif.
2. *Image-Text Matching Loss* (ITM) bertujuan dalam mengaktifkan *encoder* teks berdasarkan gambar. Bagian ini merupakan tugas klasifikasi biner dimana model memprediksi apakah pasangan gambar dan teks cocok atau tidak cocok berdasarkan fitur *multimodal* mereka.
3. *Language Modelling Loss* (LM) bertujuan mengaktifkan *decoder* teks berdasarkan gambar. Ini berfokus menghasilkan deskripsi teks yang dikondisikan pada gambar yang diberikan.

Intinya, BLIP menggunakan model serbaguna yang dapat menyatukan informasi visual ke dalam pemrosesan teks. Selama tahap pra-pelatihan, BLIP secara simultan mengoptimalkan tujuan keselarasan representasi visual dan teks, pencocokan pasangan gambar dan teks, serta pembangkitan deskripsi dalam bahasa dan gambar (Li et al., 2022).

2.6 MATRIKS EVALUASI

Matriks evaluasi yang digunakan dalam menilai kinerja model *medical VQA* dapat dikategorikan menjadi dua jenis, yaitu *classification-based metrics* dan *language-based metrics*. Pada umumnya metrik yang digunakan pada kasus klasifikasi, seperti akurasi dan *F1-score*. Metrik ini memperlakukan jawaban sebagai hasil dari klasifikasi dan menghitung pencocokan yang tepat untuk akurasi, presisi, *recall* dan lain-lain. Sedangkan metrik yang digunakan pada kasus *language-based* bertujuan untuk mengevaluasi tugas seperti penerjemah, *image captioning*, VQA, dan lain-lain. *Dataset* seperti VQA-Med-2018, VQA-Med-2019, PathVQA, VQA-Med-2020, dan VQA-Med-2021 menggunakan metrik *language-based* seperti *BLEU*, untuk mengevaluasi kinerjanya (Lin et al., 2023).

Confusion matrix adalah suatu alat pengukuran kinerja yang digunakan dalam konteks masalah klasifikasi pada pembelajaran mesin. Matriks ini terdiri dari tabel memberikan gambaran visual dan ringkasan mengenai kinerja algoritma klasifikasi. Tiap baris dalam matriks mencerminkan contoh dalam kelas aktual, sementara setiap kolom mencerminkan contoh dalam kelas yang diprediksi. *Confusion matrix* sangat bermanfaat untuk menilai efektivitas suatu model, terutama ketika terdapat ketidakseimbangan jumlah observasi antar kelas atau ketika menghadapi lebih dari dua kelas dalam suatu *dataset*. *Confusion matrix* merefleksikan nilai *True Positive* (TP) yang menunjukkan klasifikasi yang tepat pada kelas yang relevan, nilai *False Positive* (FP) yang mencerminkan klasifikasi di kelas yang seharusnya tidak relevan, nilai *False Negative* (FN) yang menunjukkan klasifikasi di kelas yang salah ketika seharusnya relevan, dan nilai *True Negative* (TN) yang mencerminkan klasifikasi yang benar di kelas yang tidak relevan. Gambar *confusion matrix* dapat dilihat pada Gambar 2.8 (Kulkarni et al., 2020).

		Actual class	
		P	N
Predicted class	P	TP	FP
	N	FN	TN

Gambar 2.8. *Confusion matrix* (Kulkarni et al., 2020)

2.6.1 Akurasi

Pada umumnya akurasi adalah metrik yang mengukur perbandingan antara jumlah prediksi yang benar dengan jumlah total kasus yang dievaluasi (Hossin and Sulaiman, 2015). Pada persamaan 2.7 menunjukkan bahwa rumus dari akurasi adalah jumlah prediksi yang benar dibagi dengan total prediksi yang dilakukan.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.7)$$

- TP (*True Positive*) adalah jumlah prediksi yang benar positif.
- TN (*True Negative*) adalah jumlah prediksi yang benar negatif.
- FP (*False Positive*) adalah jumlah prediksi yang salah positif.
- FN (*False Negative*) adalah jumlah prediksi yang salah negatif.

2.6.2 Bilingual Evaluation Understudy

Bilingual Evaluation Understudy (BLEU) adalah metrik yang digunakan untuk mengukur kesamaan pada frase (n-grams) antara dua kalimat. BLEU adalah metrik original untuk penerjemah mesin dan juga bisa digunakan untuk tugas seperti pembuatan laporan medis (Li et al., 2021).

BLEU adalah metrik yang digunakan untuk mengevaluasi kualitas sistem terjemahan mesin (Papineni et al., 2002). Ini mengukur kemiripan antara terjemahan yang dihasilkan oleh mesin dan satu atau lebih referensi terjemahan menggunakan metrik presisi yang dimodifikasi dan hukuman singkat yang dimodifikasi dan *brevity penalty* yang dimodifikasi.

$$BLEU = BP \times \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (2.8)$$

- BP adalah *brevity penalty*.
- N adalah jumlah maksimal n-gram.
- w_n adalah bobot untuk n-gram.
- p_n adalah presisi n-gram.

Berdasarkan persamaan 2.8 *Brevity Penalty* (BP) menghukum jawaban singkat, w_n adalah bobot antara 0 dan 1 untuk $\log p_n$ dan $\sum_{n=1}^N w_n = 1$, p_n adalah nilai rata-rata geometris dari presisi n-gram yang dimodifikasi, dan N adalah panjang maksimum n-gram.

BLEU memiliki beberapa variasi, seperti BLEU-1, BLEU-2, BLEU-3, dan seterusnya. Variasi ini menunjukkan jumlah n-gram yang digunakan dalam perhitungan BLEU. Semakin tinggi nilai BLEU, semakin baik kualitas terjemahan mesin tersebut. Nilai BLEU berkisar dari 0 hingga 100, di mana 100 menunjukkan terjemahan yang sempurna. Berikut adalah penjelasan mengenai berbagai variasi BLEU:

1. BLEU-1 adalah variasi dasar dari metrik BLEU yang hanya mempertimbangkan *unigram* (kata tunggal). Perhitungan BLEU-1 mengukur seberapa baik kata-kata dalam hasil terjemahan sesuai dengan kata-kata dalam kalimat referensi tanpa mempertimbangkan urutan kata.
2. BLEU-2 memperluas perhitungan dengan menggunakan *bigram* (pasangan kata). Dalam BLEU-2, selain mempertimbangkan kesesuaian *unigram*, juga diperiksa kesesuaian pasangan kata dalam hasil terjemahan dan kalimat referensi. Hal ini memberikan evaluasi yang lebih baik terhadap struktur kalimat terjemahan.
3. BLEU-3 melangkah lebih jauh dengan menggunakan *trigram* (tiga kata berturut-turut). Disini, evaluasi BLEU-3 memperhitungkan kesesuaian *trigram* antara hasil terjemahan dan kalimat referensi. Semakin panjang n-gram yang digunakan, semakin ketat evaluasi terhadap kualitas terjemahan, karena memperhatikan konteks yang lebih luas dalam kalimat.

Kesimpulannya, variasi-variasi ini menunjukkan bahwa dengan meningkatnya nilai "n" dalam n-gram, evaluasi terhadap terjemahan mesin menjadi lebih komprehensif dan memperhitungkan konteks yang lebih luas dalam kalimat (Papineni et al., 2002).

Deskripsi BLEU berbeda tergantung pada rentang nilai yang ada. Detail deskripsi untuk masing-masing rentang nilai BLEU dapat ditemukan di Tabel 2.1 (Lavie, 2010).

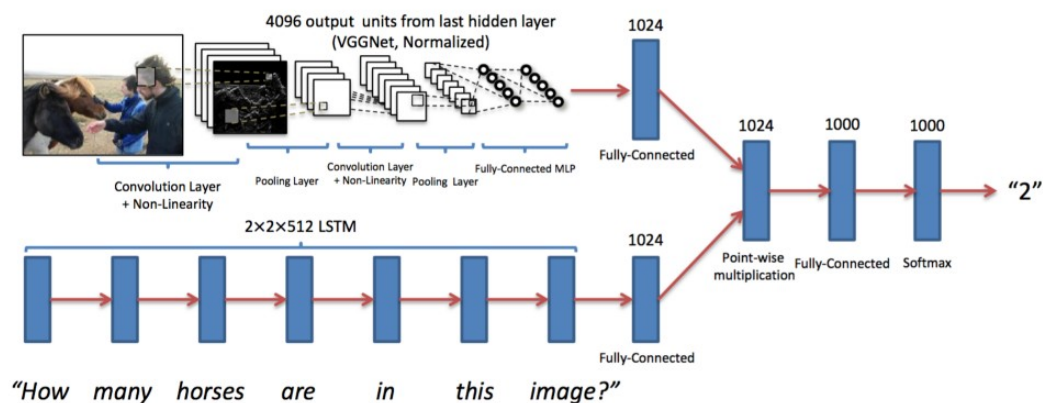
Tabel 2.1. Deskripsi dari masing-masing nilai BLEU

Nilai BLEU	Deskripsi
< 10	Hampir tidak dapat dipahami
10 - 19	Sulit untuk mendapatkan intisari kalimat
20 - 29	Inti kalimat jelas, tetapi memiliki banyak kesalahan pada tata bahasa
30 - 39	Dapat dimengerti sebagai penerjemah kalimat yang baik
40 - 49	Terjemahan kalimat berkualitas tinggi
50 - 59	Terjemahan kalimat sangat berkualitas dan memadai
> 60	Kemampuan menerjemah sering lebih bagus dari manusia

2.7 VISUAL QUESTION ANSWERING

Visual Question Answering (VQA) adalah suatu tugas tentang bagaimana menjawab pertanyaan bebas pada sebuah gambar. Karena membutuhkan pemahaman bahasa yang mendalam tentang pertanyaan dan kemampuan untuk mengaitkannya dengan berbagai objek yang ada dalam gambar. Ini adalah tugas yang ambisius dan memerlukan teknik dari baik *computer vision* dan *natural language processing* (Hildebrandt et al., 2020).

Salah satu contoh percobaan yang dilakukan oleh (Antol et al., 2015), mereka menerapkan *baseline Vanilla VQA* model yang dimana sebagai tolak ukur untuk metode *deep learning*. Model ini menggunakan CNN untuk ekstraksi fitur dan jaringan LSTM atau *recurrent network* untuk pemrosesan bahasa. Fitur-fitur ini digabungkan menggunakan suatu operasi untuk membentuk satu fitur bersama, yang digunakan untuk mengklasifikasikan salah satu jawaban seperti yang ditunjukkan dalam Gambar 2.9.



Gambar 2.9. Vanilla VQA network model (Srivastava et al., 2021)

VQA telah menarik minat besar dan mengalami perkembangan oleh para peneliti dan ilmuwan dari seluruh dunia. Tren terbaru yang diamati adalah dalam pengembangan *dataset* yang terlihat semakin mirip dengan dunia nyata dengan menggabungkan pertanyaan dan jawaban seputar kehidupan nyata. Tren terkini juga terlihat dalam pengembangan model *deep learning* yang lebih canggih dengan lebih baik memanfaatkan petunjuk visual dan petunjuk teks melalui berbagai cara. Kinerja model terbaik saat ini masih tertinggal dan hanya sekitar 60-70% saja. Oleh karena itu, masih merupakan masalah yang terbuka untuk mengembangkan model *deep learning* yang baik serta *dataset* yang lebih menantang untuk *visual question answering* (Srivastava et al., 2021).

Dalam mengembangkan sistem VQA diperlukan dua cabang ilmu seperti yang sudah dijelaskan yaitu *computer vision* dan *natural language processing* dikarenakan melibatkan gambar dan pertanyaan teks yang terkait (Teney et al., 2017).

2.7.1 Computer Vision

Computer Vision (CV) bisa diartikan sebagai sekelompok teknik yang digunakan untuk mengumpulkan, memproses, menganalisis, dan memahami data yang kompleks dan berdimensi tinggi dari lingkungan sekitar kita. Tujuannya adalah untuk menjalani eksplorasi ilmiah dan teknis yang melibatkan pemahaman dan interpretasi data visual. Secara sederhana, CV memungkinkan komputer atau sistem komputasi untuk melihat, menginterpretasi, dan merespons informasi visual dari dunia nyata. Teknologi ini memiliki berbagai aplikasi, termasuk pengenalan objek, deteksi pola, dan pengolahan citra, serta mendukung berbagai bidang seperti kecerdasan buatan, penglihatan mesin, dan analisis data visual (Jähne et al., 1999).

2.7.2 Natural Language Processing

Natural Language Processing (NLP) adalah cabang ilmu dari pembelajaran mesin yang berfokus pada teks. Metode ini memungkinkan suatu komputer untuk menganalisis, memahami, dan menghasilkan makna dari bahasa manusia dengan cerdas dan bermanfaat. Dengan memanfaatkan NLP, pengembang dapat mengorganisir dan menyusun pengetahuan untuk menjalankan tugas-tugas seperti rangkuman otomatis, terjemahan, pengenalan entitas bernama, ekstraksi hubungan, analisis sentimen, pengenalan ucapan, dan segmentasi topik (Agarwal and Saxena, 2019).

Dalam NLP *word embedding* merupakan representasi dari suatu kata. *Embedding* digunakan dalam analisis teks. Biasanya, representasi tersebut berupa vektor nilai riil yang mengkodekan makna kata sedemikian rupa sehingga kata-kata yang lebih dekat dalam ruang vektor diharapkan memiliki makna yang mirip (Teller, 2000). *Global Vectors for Word Representation* (GloVe) adalah metode untuk membuat representasi kata. GloVe adalah algoritma *unsupervised learning* yang memperoleh representasi vektor untuk kata-kata dengan melatih pada statistik kumulatif keterjadian bersama global kata-kata dari suatu korpus. Representasi yang dihasilkan menunjukkan struktur sublinear menarik dari ruang vektor kata, memungkinkan kata-kata dengan makna serupa berada lebih dekat dalam ruang vektor. Dalam hal ini, GloVe digunakan untuk membangun fitur *embedding* kata semantik dan telah digunakan dalam berbagai tugas dalam bidang NLP (Pennington et al., 2014).

2.8 MEDICAL VISUAL QUESTION ANSWERING DATASET

Medical visual question answering dataset adalah *dataset* yang digunakan untuk melakukan penelitian terkait tugas yang berkaitan dengan *medical visual question answering* (Lin et al., 2023). *Dataset* ini berisi gambar medis dan pertanyaan yang terkait dengan gambar tersebut. Pada penelitian ini akan digunakan *dataset* PathVQA dan VQA-RAD. Berikut adalah penjelasan dari kedua *dataset* tersebut.

2.8.1 PathVQA Dataset

Dataset PathVQA berisi 32.799 pertanyaan dan jawaban dari 1.670 gambar patologi yang dikumpulkan dari dua buku patologi, yaitu '*Textbook of Pathology*' dan '*Basic Pathology*'. Selain itu, terdapat 3.328 gambar lainnya yang dikumpulkan dari perpustakaan digital PEIR. Dengan demikian, total gambar pada *dataset* ini mencapai 4.998. Rata-rata setiap satu gambar memiliki 6,6 pertanyaan. Jumlah pertanyaan maksimal dan minimal adalah 14 dan 1 berturut-turut. Rata-rata jumlah kata per pertanyaan dan perjawaban adalah 9,5 dan 2,5 berturut-turut. Pada *dataset* ini terdapat 7 kategori pertanyaan, yaitu: *what*, *where*, *when*, *whose*, *how*, *how much/how many*, dan *yes/no*. Pertanyaan dari enam kategori pertama bersifat *open-ended*, sedangkan pertanyaan dari kategori terakhir bersifat *close-ended* '*yes/no*'. Jumlah jawaban '*yes*' dan '*no*' adalah 8.145 dan 8.189 berturut-turut. Pertanyaan pada *dataset* ini mencakup banyak aspek konten visual, termasuk warna, lokasi, penampilan, bentuk, dan lain-lain. Keragaman klinis ini menimbulkan tantangan yang besar bagi model AI dalam menjawab permasalahan pada gambar patologi (Xuehai et al., 2020).

2.8.2 VQA-RAD Dataset

Dataset VQA-RAD adalah *dataset* yang dibuat secara manual untuk penelitian terkait *Medical Visual Question Answering* (VQA). *Dataset* ini berisi 3.515 pasangan pertanyaan dan jawaban yang dihasilkan oleh para klinisi, serta 315 gambar radiologi yang terbagi merata antara kepala, dada, dan perut. Dengan rata-rata, setiap gambar memiliki 10 pertanyaan terkait. Setiap gambar ini terkait dengan beberapa pertanyaan, yang dikelompokkan ke dalam 11 kategori: kelainan, atribut, modalitas, sistem organ, warna, penghitungan, keberadaan objek/kondisi, ukuran, bidang, penalaran posisional, dan lain-lain. Setengah dari pertanyaan pada *dataset* ini bersifat *close-ended* (*yes/no*), sedangkan setengahnya lagi bersifat *open-ended*. *Dataset* ini merupakan kontribusi berharga dalam pengembangan model VQA di bidang radiologi, dengan pertanyaan-pertanyaan yang disusun oleh para ahli klinis, mencakup berbagai aspek interpretasi gambar medis (Lau et al., 2018).

2.9 PENELITIAN TERKAIT

Penelitian yang dilakukan oleh (Xuehai et al., 2020) berhasil mengatasi tantangan khusus dalam analisis citra patologi dengan menggabungkan metode pemrosesan gambar dan pemrosesan bahasa alami. Mereka melakukan eksperimen dengan tiga metode berbeda. Tiga metode yang digunakan melibatkan integrasi *Gated Recurrent Unit* (GRU) dan Faster R-CNN dengan melibatkan *Bilinear Attention Network*, *Convolutional Neural Network* (CNN) dan *Long Short Term Memory* (LSTM) dengan *compact bilinear pooling*, serta *Stacked Attention Network* (SAN) yang menggabungkan CNN dan LSTM. Pengkodean gambar menggunakan Faster R-CNN dan ResNet-152. Berdasarkan ketiga eksperimen ini menunjukkan bahwa metode pertama mencapai akurasi tertinggi untuk pertanyaan *close-ended* adalah 68,2% dan mendapatkan nilai BLEU-1, BLEU-2, BLEU-3 sebesar 32,4, 22,8, 17,4 berturut-turut untuk pertanyaan *open-ended*. Metode kedua mencapai akurasi sebesar 57,6% untuk pertanyaan *close-ended* dan mendapatkan BLEU-1, BLEU-2, BLEU-3, sebesar 13,3, 9,5, 6,8 berturut-turut untuk pertanyaan *open-ended*. Metode ketiga mendapatkan akurasi sebesar 59,4% untuk pertanyaan *close-ended* dan mendapatkan BLEU-1, BLEU-2, BLEU-3 sebesar 19,2, 17,9, 15,8 berturut-turut untuk pertanyaan *open-ended*. Lalu, hasil yang didapat ketika melakukan pengkodean gambar dengan Faster R-CNN mendapatkan akurasi 62,0% untuk pertanyaan *close-ended* dan mendapatkan BLEU-1, BLEU-2, BLEU-3 sebesar 24,7, 19,1, 16,5 berturut-turut untuk pertanyaan *open-ended*. Sedangkan ketika menggunakan ResNet-152, mereka mendapatkan akurasi 60,1% untuk pertanyaan *close-ended* dan BLEU-1, BLEU-2, BLEU-3 sebesar 19,9, 18,0, 16,0 berturut-turut untuk pertanyaan *open-ended* (Xuehai et al., 2020).

Penelitian lain juga dilakukan untuk meningkatkan diagnosis dengan menjawab pertanyaan klinis yang disajikan dengan gambar medis menggunakan model *Contrastive Language Image Pre Training* (CLIP) yang mengintegrasikan pembelajaran *multimodal embedding* melalui pelatihan bersama *image encoder* dan *text encoder* untuk memaksimalkan kemiripan pasangan gambar-teks yang sesuai. Metodenya melibatkan penggunaan *Vision Transformer* (ViT) untuk *image encoder*, pertanyaan dimasukkan ke dalam *transformer encoder* teks, dan penggabungan representasi visual dan teks dalam *decoder multimodal* untuk menghasilkan jawaban otomatis. Uji coba dilakukan pada dua *dataset* VQA, PathVQA dan VQA-RAD, dengan konfigurasi menggunakan *multi-head attention* sebesar 512, tingkat *dropout* pada *fully connected layers* 0,1, dan dua blok *transformer* pada *encoder* dan *decoder*. Fungsi optimasi yang digunakan adalah Adam dengan *learning rate* 0,001, *batch size* 50, dan 50 *epoch* pelatihan, dengan citra diacak dan pengolahan teks menggunakan *bert-base-uncased*. Hasilnya menunjukkan akurasi yang signifikan, dengan akurasi

tertinggi mencapai 84,63% untuk pertanyaan *close-ended* dan akurasi, BLEU-1, BLEU-2, BLEU-3 adalah sebesar 58,29% 61,78, 61,16, 59,28 secara berturut-turut untuk pertanyaan *open-ended* dalam *dataset* PathVQA. Pada *dataset* VQA-RAD akurasi yang didapatkan adalah 82,47% untuk pertanyaan *close-ended* dan akurasi, BLEU-1, BLEU-2, BLEU-3 adalah sebesar 71,49% 71,03, 70,81, 67,01 untuk pertanyaan *open-ended* (Bazi et al., 2023).

BAB III METODOLOGI PENELITIAN

3.1 WAKTU DAN LOKASI PENELITIAN

Penelitian ini akan bertempat di Ruang Lab Sistem Informasi dan *Database*. Waktu yang dibutuhkan agar penelitian ini dapat diimplementasikan adalah 4 bulan terhitung dari bulan Februari 2024 hingga Mei 2024.

3.2 JADWAL PELAKSANAAN

Untuk lebih memahami alur waktu dari penelitian ini, penulis telah menyusun sebuah jadwal yang rinci yang bisa dilihat pada Tabel 3.1.

Tabel 3.1. Jadwal pelaksanaan penelitian

No	Kegiatan	Bulan Ke -																							
		Januari				Februari				Maret				April				Mei				Juni			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	Studi literatur																								
2	Pengumpulan data																								
3	Penrosesan data																								
4	Membangun model																								
5	Melatih model																								
6	Analisis hasil																								
7	Penyusunan laporan akhir																								

3.3 ALAT DAN BAHAN

Alat dan Bahan yang akan digunakan pada penelitian ini terdiri dari beberapa perangkat keras (*hardware*) dan perangkat lunak. Lalu, data yang digunakan adalah data dari *dataset* PathVQA, dan VQA-RAD.

3.3.1 Perangkat Keras

- Laptop Lenovo Yoga C740 dengan RAM 16GB DDR4, Intel® Core™ i7-10710U. 1.10 - 4.70 GHz, *Solid State Drive* (SSD) 1TB.
- Server spesifikasi processor Intel Xeon Gold 5218, CPU 2.30GHz 64 inti, RAM 128GB, VGA 4 x NVIDIA GeForce RTX 2080 Ti GPU VRAM 12GB, dan memory 8TB.

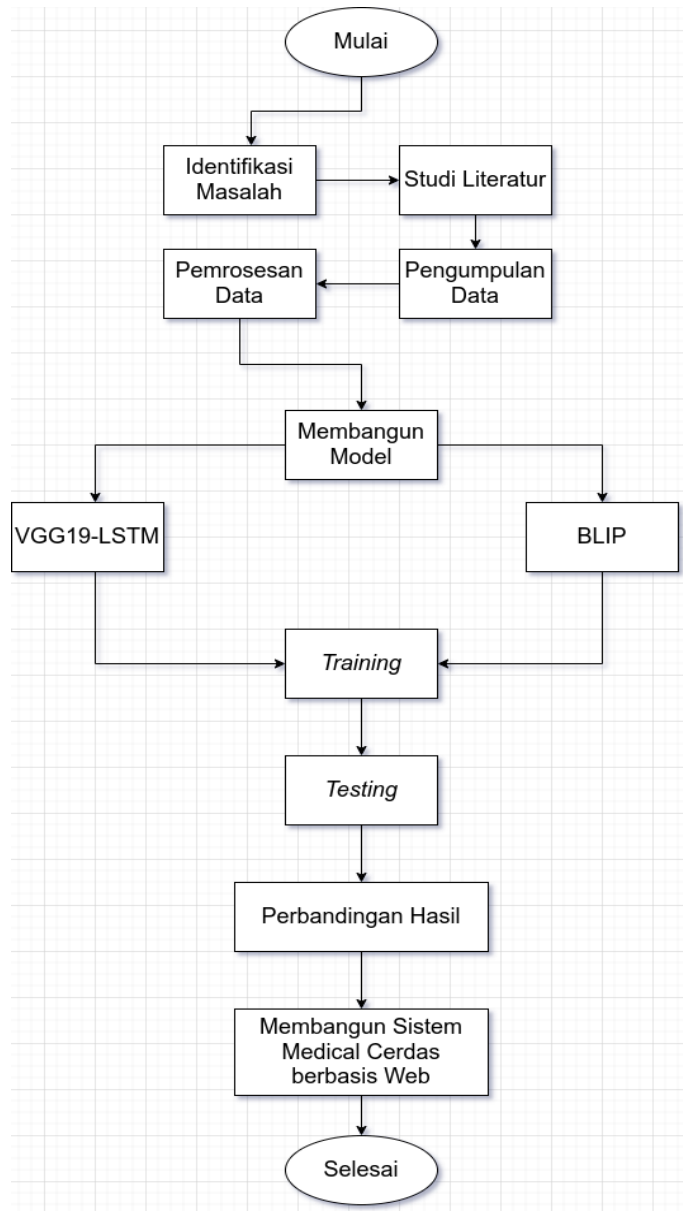
3.3.2 Perangkat Lunak

- Linux Debian Ubuntu versi 22.04 LTS
- Visual Studio Code versi 1.71.0

- c. Python versi 3.8.17
- d. PyTorch versi 2.0.1
- e. Tensorflow versi 2.11.0
- f. Transformers versi 4.31.0
- g. Huggingface Datasets versi 2.18.0
- h. FastAPI versi 0.104.1
- i. nlpaug versi 1.1.11

3.4 METODE PENELITIAN

Penelitian ini akan mengikuti beberapa tahapan yang dirancang secara sistematis untuk mencapai tujuan yang telah ditetapkan. Setiap tahap akan dilaksanakan dengan seksama untuk memastikan hasil yang akurat dan bermanfaat. Skema dari alur tahapan penelitian yang diusulkan dapat dilihat secara rinci pada Gambar 3.1. Pada Gambar 3.1 terdapat beberapa tahapan yang akan dilakukan dalam penelitian ini, yaitu identifikasi masalah, studi literatur, pengumpulan data, pemrosesan data, pembangunan model, pelatihan model, evaluasi model, dan pembangunan sistem medis cerdas berbasis web.



Gambar 3.1. Diagram alir penelitian

3.4.1 Identifikasi Masalah

Identifikasi masalah adalah proses awal dalam mengidentifikasi masalah yang akan diselesaikan dalam penelitian ini. Identifikasi masalah meliputi latar belakang masalah, rumusan masalah, tujuan penelitian serta manfaat yang dapat dihasilkan dari penelitian ini.

3.4.2 Studi Literatur

Studi literatur dilakukan untuk memperoleh informasi mengenai penelitian yang sudah dilakukan sebelumnya. Dalam tahapan ini terdiri dari banyak kegiatan yaitu




membaca referensi yang berkaitan dengan penelitian ini, membaca jurnal atau buku untuk menambah wawasan, sehingga dapat memperoleh informasi yang dibutuhkan untuk penelitian ini.

3.4.3 Pengumpulan Data

Penelitian tugas akhir ini menggunakan data dari *dataset* PathVQA dan VQA-RAD. Pengumpulan *dataset* dilakukan dengan mengunduh data dari platform Hugging Face. Sampel kedua data ini bisa dilihat pada Gambar 3.2 dan Gambar 3.3.

image image · width (px)	question string · lengths	answer string · lengths
285-426 2.7%	58-83 10.3%	23-45 8.3%
	where are liver stem cells (oval cells) located?	in the canals of hering
	what are stained here with an immunohistochemical stain for cytokeratin...	bile duct cells and canals of hering
	what do the areas of white chalky deposits represent?	foci of fat necrosis

Gambar 3.2. Sampel data PathVQA

image image · width (px)	question string · lengths	answer string · lengths
463-670 27.9%	26-37 35.2%	1-13 78.5%
	are regions of the brain infarcted?	yes
	are the lungs normal appearing?	no
	which organ system is abnormal in this image?	cardiovascular

Gambar 3.3. Sampel data VQA-RAD

3.4.4 Pemrosesan Data

Pada bagian ini, dilakukan pemrosesan data dengan tujuan menyelaraskan dan menyusun *dataset* yang diperlukan untuk model *Visual Question Answering* (VQA). Proses ini melibatkan penggabungan antara data visual, seperti gambar, dan data teks, berupa pertanyaan yang berkaitan. Pemrosesan data bertujuan untuk mengkonversi informasi kompleks dari kedua modalitas ini ke dalam format yang dapat dicerna oleh model. Dengan melakukan pemrosesan data ini, kita menciptakan representasi yang optimal untuk menyajikan informasi visual dan teks kepada model VQA. Hasilnya adalah *dataset* yang kohesif dan siap digunakan untuk melatih model VQA agar mampu memberikan jawaban yang tepat terhadap pertanyaan yang diajukan berdasarkan konteks visual yang diberikan. Adapun pemrosesan data yang dilakukan mencakup pemrosesan gambar dan teks.

1. *Case Folding*

Pada bagian ini proses *case folding* dilakukan untuk mengubah semua karakter dalam teks menjadi huruf kecil. Hal ini dilakukan untuk menghindari duplikasi kata yang sama dengan huruf besar dan kecil. Tabel 3.2 menunjukkan contoh proses *case folding*.

Tabel 3.2. Proses *case folding*

Sebelum	Sesudah
The Quick Brown Fox JUMPS	the quick brown fox jumps

2. *Remove Punctuation*

Pada bagian ini proses *remove punctuation* dilakukan untuk menghapus tanda baca dalam teks. Hal ini dilakukan untuk mengurangi kompleksitas kata dalam teks. Tabel 3.3 menunjukkan contoh proses *remove punctuation*.

Tabel 3.3. Proses *remove punctuation*

Sebelum	Sesudah
She said, 'Hello!' and waved	She said Hello and waved

3. *Stemming*

Pada bagian ini proses *stemming* dilakukan untuk mengubah kata-kata dalam teks menjadi kata dasar. Hal ini dilakukan untuk mengurangi kompleksitas kata dalam teks. Tabel 3.4 menunjukkan contoh proses *stemming-text*.

Tabel 3.4. Proses *stemming*

Sebelum	Sesudah
running quickly in the park	run quick in the park

4. Tokenisasi

Pada bagian ini proses tokenisasi dilakukan untuk memecah teks menjadi token-token yang lebih kecil. Hal ini dilakukan untuk mempermudah model VQA memahami dan memproses secara efektif. Tabel 3.5 menunjukkan contoh proses tokenisasi.

Tabel 3.5. Proses tokenisasi

Sebelum	Sesudah
The cat in the hat	[The, cat, in, the, hat]

5. Pemrosesan Gambar

Pada bagian ini proses pemrosesan gambar dilakukan menyamaratakan format gambar. Hal ini dilakukan untuk mempermudah proses pemrosesan data dan mempersiapkan data agar siap digunakan untuk melatih model VQA.

3.4.5 Membangun Model

Proses membangun model VQA ini menggunakan arsitektur VGG19-LSTM dan BLIP. Pada VGG19-LSTM, terdapat dua tahap pemrosesan, yaitu ekstraksi fitur gambar menggunakan VGG19 dan pemrosesan teks menggunakan LSTM. Sedangkan pada BLIP, terdapat dua *encoder*, yaitu *image encoder* dan *text encoder*, yang digunakan untuk mengolah data gambar dan teks. Dalam BLIP, *image encoder* menggunakan *Vision Transformer* (ViT), sedangkan *text encoder* menggunakan *transformer encoder*. Proses ini bertujuan untuk membangun model VQA yang mampu memberikan jawaban yang tepat terhadap pertanyaan yang diberikan berdasarkan konteks visual yang diberikan.

3.4.6 Melatih Model

Tahapan pada pelatihan model ini menggunakan arsitektur VGG19-LSTM dan BLIP. Pada tahap ini akan dilakukan suatu metode yang bertujuan untuk mengubah nilai atau parameter yang ada pada suatu arsitektur, Proses ini disebut sebagai *hyperparameter tuning*. *Hyperparameter* ini mengacu pada parameter yang tidak dapat diubah ketika proses pelatihan model berjalan, contohnya seperti jumlah *hidden layer*,

nilai *epoch*, *batch size*, *learning rate* pada fungsi optimasi, dan lain-lain. *Hyperparameter tuning* ini dilakukan untuk mencari kombinasi nilai yang optimal untuk parameter-parameter tersebut (Yu and Zhu, 2020).

3.4.7 Perbandingan Hasil

Setelah proses pelatihan model selesai, maka dilakukan perbandingan hasil dari kedua model yang telah dibangun dengan menggunakan metrik evaluasi yaitu akurasi dan *Bilingual Evaluation Understudy* (BLEU). Pada metrik BLEU, penulis akan menggunakan BLEU-1, BLEU-2, dan BLEU-3, bilangan pada nama BLEU menunjukkan jumlah n-gram yang digunakan. Kedua metrik ini digunakan untuk mengevaluasi kinerja dari model VQA yang telah dilatih. Perbandingan hasil ini dilakukan dengan cara menguji kedua model menggunakan data uji yang telah disiapkan. Hasil dari perbandingan ini akan digunakan untuk menentukan model mana yang lebih baik dalam memberikan jawaban dari pertanyaan yang diberikan.

3.4.8 Membangun Sistem Medis Cerdas Berbasis Web

Model yang telah siap dilatih akan di *deploy* ke dalam sistem medis cerdas berbasis web. Dalam proses membangun *website* ini meliputi pengembangan halaman tampilan yang bisa menerima dua *input* yaitu gambar dan pertanyaan dari pengguna. Setelah itu mengembangkan bagian *backend* untuk mengolah kedua *input* tersebut dan memasukkannya ke dalam model yang telah dibangun sebelumnya. Setelah model menghasilkan jawaban, maka jawaban tersebut akan dikirimkan kembali ke halaman tampilan untuk ditampilkan kepada pengguna.

DAFTAR PUSTAKA

- Agarwal, M. & Saxena, A. (2019). An overview of natural language processing. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 7(5):2811–2813.
- Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A., & Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *electronics*, 8(3):292.
- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., & Farhan, L. (2021). Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8:1–74.
- Andono, P. N., Sutojo, T., et al. (2018). *Pengolahan citra digital*. Penerbit Andi.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Ardiansyah, R. F. (2013). Pengenalan pola tanda tangan dengan menggunakan metode principal component analysis (pca). *Fakultas Ilmu Komputer Universitas Dian Nuswantoro*, 2:14.
- Basu, K., Sinha, R., Ong, A., & Basu, T. (2020). Artificial intelligence: How is it changing medical sciences and its future? *Indian journal of dermatology*, 65(5):365.
- Bazi, Y., Rahhal, M. M. A., Bashmal, L., & Zuair, M. (2023). Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380.
- Bergstra, J. & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).
- Brownlee, J. (2017). A gentle introduction to long short-term memory networks for the experts. *Machine Learning Mastery*.
- Cireşan, D. C., Meier, U., & Schmidhuber, J. (2012). Transfer learning for latin and chinese characters with deep neural networks. In *The 2012 international joint conference on neural networks (IJCNN)*, pages 1–6. IEEE.
- Fernando, D. & Harsiti, H. (2019). Studi literatur: Robotic process automation. *JSiI (Jurnal Sistem Informasi)*, 6(1):6–11.
- Gurney, K. (1997). *An Introduction to Neural Networks*. Taylor & Francis, Inc., USA.
- Hendrycks, D. & Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

- Henningsen-Schomers, M. R. & Pulvermüller, F. (2022). Modelling concrete and abstract concepts using brain-constrained deep neural networks. *Psychological research*, 86(8):2533–2559.
- Hildebrandt, M., Li, H., Koner, R., Tresp, V., & Günnemann, S. (2020). Scene graph reasoning for visual question answering. *arXiv preprint arXiv:2007.01072*.
- Ho, Y. & Wooley, S. (2019). The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE access*, 8:4806–4813.
- Hossin, M. & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.
- Huang, D., Fu, Y., Qin, N., & Gao, S. (2021). Fault diagnosis of high-speed train bogie based on lstm neural network. *Sci. Chin. Inf. Sci*, 64:1–3.
- Ilahiyah, S. & Nilogiri, A. (2018). Implementasi deep learning pada identifikasi jenis tumbuhan berdasarkan citra daun menggunakan convolutional neural network. *JUSTINDO (Jurnal Sistem Dan Teknologi Informasi Indonesia)*, 3(2):49–56.
- Jähne, B., Haussecker, H., & Geissler, P. (1999). *Handbook of computer vision and applications*, volume 2. Citeseer.
- Kelleher, J. D. (2019). *Deep learning*. MIT press.
- Khan, A., Sohail, A., Zahoor, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial intelligence review*, 53:5455–5516.
- Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kristiyanti, D. A. & Saputra, I. (2023). Machine learning untuk pemula.
- Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). 5 - foundations of data imbalance and solutions for a data democracy. In Batarseh, F. A. & Yang, R., editors, *Data Democracy*, pages 83–106. Academic Press.
- Kusuma, A. W. & Ellyana, R. L. (2018). Penerapan citra terkompresi pada segmentasi citra menggunakan algoritma k-means. *Jurnal Terapan Teknologi Informasi*, 2(1):65–74.
- Kusuma, I. W. A. W. & Kusumadewi, A. (2020). Penerapan metode contrast stretching, histogram equalization dan adaptive histogram equalization untuk meningkatkan kualitas citra medis mri. *Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer*, 11(1):1–10.
- Lau, J. J., Gayen, S., Ben Abacha, A., & Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

- Lavie, A. (2010). Evaluating the output of machine translation systems. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Tutorials*.
- Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.
- Li, K. & Malik, J. (2017). Learning to optimize neural nets. *arXiv preprint arXiv:1703.00441*.
- Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., et al. (2021). Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., & Ge, Z. (2023). Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, page 102611.
- Luo, N., Zhong, X., Su, L., Cheng, Z., Ma, W., & Hao, P. (2023). Artificial intelligence-assisted dermatology diagnosis: From unimodal to multimodal. *Computers in Biology and Medicine*, 165:107413.
- Mauli, D. (2018). Tanggung jawab hukum dokter terhadap kesalahan diagnosis penyakit kepada pasien. *Cepalo*, 2(1):33–42.
- Monedero, Í., Barbancho, J., Márquez, R., & Beltrán, J. F. (2021). Cyber-physical system for environmental monitoring based on deep learning. *Sensors*, 21(11):3655.
- Mukhlif, A. A., Al-Khateeb, B., & Mohammed, M. A. (2023). Incorporating a novel dual transfer learning approach for medical images. *Sensors*, 23(2):570.
- Nguyen, B. D., Do, T.-T., Nguyen, B. X., Do, T., Tjiputra, E., & Tran, Q. D. (2019). Overcoming data limitations in medical visual question answering. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*.
- Nguyen, T.-H., Nguyen, T.-N., & Ngo, B.-V. (2022). A vgg-19 model with transfer learning and image segmentation for classification of tomato leaf disease. *AgriEngineering*, 4(4):871–887.
- Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*.
- Pan, S. J. & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Parai, A., Deshpande, S., Iyer, A., Kumbhare, A., & Bendale, S. (2022). Predictive cancer detection and treatment using machine learning and artificial intelligence.
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.
- Puttagunta, M. & Ravi, S. (2021). Medical image analysis based on deep learning approach. *Multimedia tools and applications*, 80:24365–24398.
- Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sorace, J., Aberle, D. R., Elimam, D., Lawvere, S., Tawfik, O., & Wallace, W. D. (2012). Integrating pathology and radiology disciplines: an emerging opportunity? *BMC medicine*, 10(1):1–6.
- Srivastava, Y., Murali, V., Dubey, S. R., & Mukherjee, S. (2021). Visual question answering using deep learning: A survey and performance analysis. In *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5*, pages 75–86. Springer.
- Teller, V. (2000). Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.
- Teney, D., Wu, Q., & van den Hengel, A. (2017). Visual question answering: A tutorial. *IEEE Signal Processing Magazine*, 34(6):63–75.
- Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., et al. (2020). Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234.
- Vardhan, J. & Swetha, K. S. (2023). Detection of healthy and diseased crops in drone captured images using deep learning. *arXiv preprint arXiv:2305.13490*.
- Wardani, K. R. & Leonardi, L. (2023). Klasifikasi penyakit pada daun anggur menggunakan metode convolutional neural network. *Jurnal Tekno Insentif*, 17(2):112–126.
- Xuehai, H., Yichen, Z., Luntian, M., Eric, X., & Pengtao, X. (2020). Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.
- Yu, T. & Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689*.
- Zhang, Z. (2016). Neural networks: further insights into error function, generalized weights and others. *Annals of translational medicine*, 4(16).