Original Article

e-ISSN: 2581-0545 - https://journal.itera.ac.id/index.php/jsat/

Received 00th January 20xx
Accepted 00th Febuary 20xx
Published 00th March 20xx

Open Access

DOI: 10.35472/x0xx0000

Deteksi Gambar Kecelakaan menggunakan CLIP dengan Pendekatan *Zero-Shot*

Della Septiani [*1] , Hartiti Fadilah [2] , Josua Alfa Viando Panggabean [3] , Anissa Luthfi Alifia [4] , Syifa Firnanda [5], Pramudya Wibowo [6] , Ade Lailani [7]

Program Studi Sains Data, Institut Teknologi Sumatera, Lampung

Corresponding E-mail : della.121450109@student.itera.ac.id

Abstract: *Accident image detection is a critical step in supporting rapid response systems during emergency situations. This study employs the Contrastive Language–Image Pre-training (CLIP) model to detect accident images using a zero-shot approach, without requiring retraining on a specific dataset. The CLIP model leverages multimodal embeddings from text and images, enabling detection based on textual descriptions. The experimental results, using the ViT base patch 32 model, show that this method achieves a Top-1 accuracy of 32% and a Top-5 accuracy of 95.86%. Although the Top-1 accuracy indicates that further optimization is needed, the high Top-5 accuracy demonstrates the significant potential of CLIP for efficient accident image detection. With further development, this technology can serve as a reliable solution for various emergency response scenarios, offering flexibility and efficiency in detecting accident-related images.*

Keywords: *CLIP, Computer Vision, Crash Detection, Multimodal model, Zero-shot learning*

Abstrak: Deteksi gambar kecelakaan merupakan langkah penting dalam mendukung sistem respon cepat terhadap situasi darurat. Penelitian ini menggunakan model *Contrastive Language–Image Pre-training* (CLIP) untuk mendeteksi gambar kecelakaan dengan pendekatan zero-shot, tanpa memerlukan pelatihan ulang menggunakan dataset spesifik. Model CLIP menggabungkan embedding multi-modal dari teks dan gambar, memungkinkan deteksi berbasis deskripsi teks. Hasil eksperimen menunjukkan bahwa dengan menggunakan model CLIP ViT base patch 32 metode ini mencapai akurasi Top-1 sebesar 32% dan Top-5 sebesar 95.86%. Meskipun akurasi Top-1 masih memerlukan optimasi lebih lanjut, performa Top-5 yang tinggi mengindikasikan potensi besar CLIP dalam mendeteksi gambar kecelakaan secara efisien. Dengan pengembangan lebih lanjut, teknologi ini dapat menjadi solusi andal untuk berbagai skenario respons darurat.

Kata Kunci : CLIP, *Computer vision,* Deteksi Kecelakaan, Model Multimodal, Pembelajaran *Zero-shot*

Original Article *Journal of Science and Applicative Technology*

Pendahuluan

Dalam era digital yang semakin berkembang, teknologi kecerdasan buatan (AI) telah menjadi komponen penting dalam berbagai bidang, termasuk keselamatan publik [1]. Salah satu aplikasi yang menarik perhatian adalah deteksi gambar kecelakaan secara otomatis menggunakan teknologi AI. Sistem ini memiliki potensi untuk mempercepat respon terhadap kecelakaan dengan mengidentifikasi kejadian dalam waktu nyata dari gambar atau video yang diunggah, baik melalui kamera pengawas, media sosial, ataupun perangkat lainnya. Namun, tantangan utama dalam pengembangan sistem ini adalah keharusan memiliki dataset besar dan dianggap baik untuk melatih model deteksi, yang sering kali sulit diperoleh [2].

Untuk mengatasi masalah ini, pendekatan *zero-shot learning* menawarkan solusi yang inovatif. Dengan memanfaatkan model seperti *Contrastive Language–Image Pre Training* (CLIP), sistem dapat mengenali objek atau situasi tanpa memerlukan pelatihan ulang menggunakan dataset spesifik. CLIP memadukan kemampuan bahasa alami dengan pengenalan gambar, memungkinkan model untuk memahami hubungan antara deskripsi teks dan konten visual secara lebih fleksibel. Pendekatan ini memberikan efisiensi yang tinggi, terutama dalam situasi yang membutuhkan deteksi cepat tanpa pelatihan tambahan [3].

Penelitian ini dilakukan untuk mengamati potensi penggunaan CLIP dalam mendeteksi gambar kecelakaan menggunakan pendekatan *zero-shot learning*. Dengan pendekatan ini, diharapkan sistem mampu mengenali berbagai jenis kecelakaan tanpa memerlukan data pelatihan yang luas, sehingga mempercepat implementasi teknologi dalam dunia nyata. Hal ini dapat menjadi kontribusi penting dalam meningkatkan efektivitas penanganan kecelakaan.

Tujuan dari penelitian ini adalah mengembangkan dan menguji model deteksi gambar kecelakaan berbasis CLIP dengan pendekatan *zero-shot* untuk mengukur tingkat akurasi dan keefektifannya. Dengan demikian, penelitian ini diharapkan dapat membuka peluang baru dalam pengembangan teknologi berbasis AI untuk mendukung keselamatan dan kesejahteraan masyarakat.

Metode Penelitian

Deskripsi Data

Penelitian ini menggunakan dataset *Kaggle* dari Charan Kumar yang berjudul *Accident Detection From CCTV Footage* [4]. Dataset ini berisi berbagai rekaman CCTV yang berfokus pada deteksi kejadian kecelakaan lalu lintas melalui analisis gambar. Dataset ini awalnya berisi 3 folder yaitu latih, uji dan validasi dengan masing-masing mempunyai 2 kelas yaitu *accident* dan *non-accident.* Pada penelitian ini hanya digunakan 1 folder dataset dan tidak di split dengan komposisi 350 data gambar kecelakaan untuk kelas *Accident* dan 350 data gambar lalu lintas normal untuk kelas *Non-Accident.*

(a) (b)

Gambar 1. Sampel dari kelas *accident* (a) dan kelas non accident (b)

Selain data gambar, pada penelitian CLIP dibutuhkan multimodal yaitu data gambar dan data teks. Data teks pada penelitian ini adalah kandidat *description* pada masing-masing kelas. Pada akhirnya, kandidat description ini akan dipilih oleh model CLIP yang digunakan untuk menentukan kandidat *description* yang relevan dengan gambarnya dengan mempertimbangkan skor similarity tertinggi. Pada Tabel 1 dan Tabel 2 Berikut kandidat *description* yang di generate secara manual.

Table 1. Kandidat Deskripsi Gambar Accident

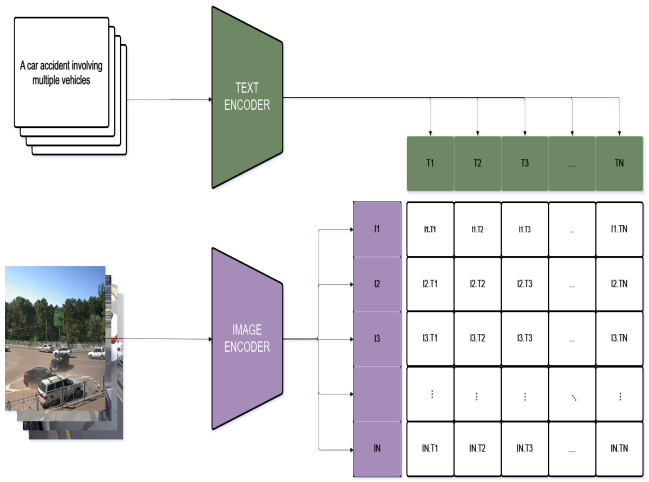
No	Kandidat Deskripsi
1	*A car accident on a road*
2	*A car accident involving multiple vehicles*
3	*A motorcycle accident involving multiple vehicles*
4	*A damaged vehicle after a collision*
5	*An overturned car on the highway*
6	*A car was hit by a truck*
7	*A car was crushed by rocks*
8	*The car crashed into a pedestrian crossing*
9	*The car hit the road divider*
10	*The truck skidded on the road*

Table 2. Kandidat Deskripsi Gambar Non Accident

No	Kandidat Deskripsi
1	*A calm street with clear traffic*
2	*A mountain landscape with no traffic*
3	*A busy street with no accidents*
4	*street view after rain with no accidents*
5	*A car parked on the side of the road*
6	*many cars parked on the side of the road*

CLIP (*Contrastive Language Image Pre-training*)

CLIP adalah model pembelajaran mesin yang melibatkan penggunaan lebih dari satu jenis data (multimodal) atau sumber informasi dalam proses pelatihan dan inferensinya [5]. CLIP dikembangkan oleh OpenAI, yang dapat memahami hubungan antara teks dan gambar [6].



Gambar 2 . Proses CLIP (*Contrastive Language–Image Pre-training*)

Dengan karakteristik metode CLIP yang menggunakan 2 jenis data yaitu gambar dan teks, diperlukan 2 encoder utama pada proses tersebut[5], yaitu :

1. Image Encoder, yaitu proses mengkonversi gambar menjadi representasi vektor.
2. Text Encoder, yaitu proses mengkonversi teks menjadi representasi vektor.

Metode CLIP ini merupakan pembelajaran kontrastif dimana metode ini menggunakan pendekatan positif (jika gambar dan teks berkaitan) dan pendekatan negatif (jika gambar dan teks tidak berkaitan) [7]. CLIP dilatih untuk memetakan gambar dan deskripsi teks ke dalam ruang vektor bersama dengan tujuan mendekatkan pasangan gambar-teks yang relevan menjadi pendekatan positif sementara pasangan yang tidak relevan menjadi lebih jauh [6]. Setelah pelatihan, CLIP dapat digunakan untuk berbagai tugas yang berbeda, karena CLIP menanamkan gambar dan teks dalam ruang yang sama, maka memungkinkan untuk menangani pengambilan gambar dan klasifikasi gambar tanpa pengambilan gambar dengan melihat kesamaan antara teks dan gambar.

Pendekatan *Zero-Shot Learning*

Zero-Shot Learning pada CLIP merujuk pada kemampuan model CLIP untuk melakukan tugas-tugas tertentu tanpa perlu pelatihan tambahan (*fine-tuning*) atau penyesuaian khusus untuk tugas tersebut[8]. Dengan kata lain, CLIP dapat mengerjakan berbagai tugas baru hanya dengan diberi deskripsi teks yang relevan tanpa perlu melihat contoh data atau gambar yang berkaitan dengan tugas itu sebelumnya[9]. CLIP bekerja dengan mengkonversi label dataset atau kandidat deskripsi dan membandingkannya dengan representasi gambar melalui kesamaan kosinus (*Cosine Similarity*) [10]. Prediksi dibuat berdasarkan skor kesamaan tertinggi antara gambar dan teks dengan fungsi *softmax* (*softmax probability*). Metode ini memungkinkan CLIP unggul dalam tugas seperti pengklasifikasian gambar dalam data multimodal secara *zero-shot* [10].

$$cosine\ similarity(v_i, t_j) = \frac{v_i \cdot t_j}{||v_i||\,||t_j||} \quad (1)$$

Keterangan:

$v_i \cdot t_j$: Hasil dot product antara vektor *v* dan *t*

$||v_i||$: Panjang vektor *v*

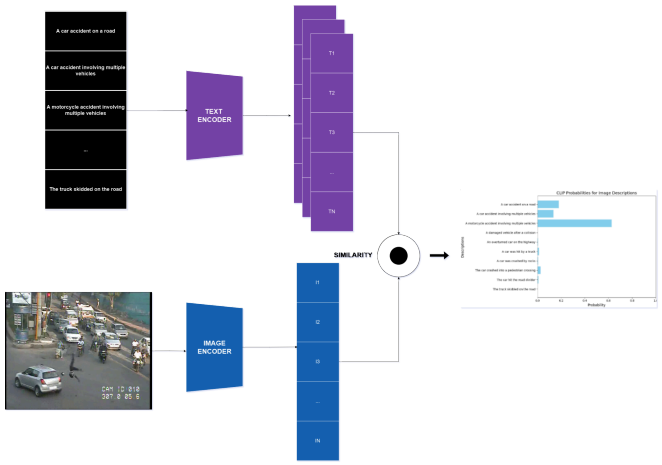
$||t_j||$: Panjang vektor *j*

$$softmax(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \quad (2)$$

Keterangan:

x_i : elemen ke-*i* dari vektor

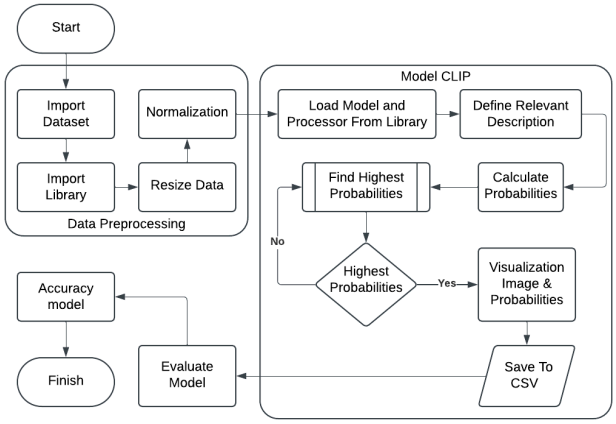
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia



Gambar 3. Proses Zero-shot pada CLIP

Diagram Alir

Pada gambar 4, merupakan gambar diagram alir dimana semua proses akan divisualisasikan dengan chart secara berurut sesuai dengan proses penelitian. Pada tahap pra-proses, dilakukan pengumpulan dataset dimana pada penelitian ini dilakukan pengunduhan dari website open source yaitu Kaggle [4]. Dilanjutkan dengan Import library dimana kita akan mengunduh dan memanggil library-library yang kita gunakan. Setelah itu dilakukan resize data untuk menyeragamkan ukuran input[11] lalu dilakukan normalisasi yaitu proses transformasi data gambar agar memiliki nilai-nilai yang seragam dalam skala tertentu[12]. Setelah pra-proses telah dilaksanakan, dilanjutkan dengan proses CLIP dimana peneliti harus memasukan model dan *processor* yang digunakan. Dalam penelitian ini digunakan 2 model untuk perbandingan utama dengan melihat efisien dalam waktu dan biaya. Model pertama yang digunakan adalah Vit base patch 32 dimana model memiliki 12 lapisan encoder serta 12 *attention heads* dan tiap inputan akan dipotong menjadi patch ukuran 32x32 piksel dan diubah menjadi embedding vektor[13]. Model kedua yang digunakan adalah Vit Large patch 14 dimana model memiliki 24 lapisan layer transformer serta 16 *multi-head attention* dan tiap inputan akan dipotong menjadi patch ukuran 14x14 piksel[14]. Dilanjutkan dengan mendefinisikan kandidat deskripsi yang relevan yaitu 10 kandidat untuk kelas *Accident* (Tabel 1) dan 6 kandidat untuk kelas *non accident* (Tabel 2). Model akan memilih mana deskripsi yang paling relevan dengan mempertimbangkan nilai terbesar dari similarity yang dihasilkan oleh model dan nilai tersebut akan disimpan ke dalam csv.



Gambar 4. Diagram alir penelitian

Hasil dan Pembahasan

Data Preprocessing

Data gambar yang digunakan dalam penelitian dibagi menjadi dua kelas, yaitu *Accident* (kecelakaan) dan *non-Accident* (bukan kecelakaan), dengan masing-masing kelas sebanyak 350 data gambar. Sebelum digunakan untuk pelatihan model, gambar-gambar tersebut melalui dua tahap *preprocessing*: *resize* dan normalisasi. *Resize* dilakukan karena ukuran gambar inputan yang beragam, sehingga perlu dilakukan penyesuaian dimensi gambar menjadi 224x224 piksel. Ukuran ini umum digunakan dan dapat membantu memaksimalkan performa model dengan memastikan konsistensi dimensi inputan. Selanjutnya, data gambar tersebut dinormalisasi ke rentang [-1, 1] dengan mean dan standar deviasi sebesar 0.5. Normalisasi ini penting untuk mengubah nilai piksel gambar yang awalnya memiliki rentang tertentu, sehingga tidak ada piksel yang mendominasi perhitungan model selama proses pelatihan. Gambar yang telah melalui tahap preprocessing kemudian disimpan di folder penyimpanan yang telah disiapkan.

Gambar 5 menunjukkan perbedaan antara gambar asli dan gambar setelah preprocessing. Gambar asli memiliki ukuran dan resolusi yang bervariasi, sementara setelah dilakukan resize, semua gambar disesuaikan menjadi ukuran 224x224 piksel untuk konsistensi input. Selain itu, normalisasi memastikan tidak ada piksel yang mendominasi perhitungan selama pelatihan model.

(a) (b)

Gambar 5. (a) data asli, (b) data setelah *preprocessing*.

Tabel 5. Perbandingan model CLIP.

Model	Average Score	Time (s)
Vit base patch32	29.58	102.64
Vit large patch14	23.79	1194.54

Setelah data gambar dinormalisasi, langkah selanjutnya adalah mendefinisikan model dan processor yang digunakan dalam penelitian ini, yaitu ViT base patch 32 dan ViT large patch 14. Dalam penerapannya, CLIP menerima dua input, gambar dan teks, dan menghasilkan output berupa skor similarity atau probability yang menunjukkan sejauh mana kecocokan keduanya. CLIP memiliki dua encoder utama: Vision Encoder berbasis Vision Transformer (ViT), yang mengkonversi gambar menjadi representasi vektor, dan Text Encoder berbasis Transformer, yang mengonversi teks menjadi vektor yang dapat dibandingkan dengan representasi gambar. Model ViT yang digunakan di sini adalah varian "base" dengan 12 lapisan dan patch ukuran 32x32 piksel, yang memungkinkan pemrosesan gambar melalui mekanisme attention dan varian “large” dengan 24 layer serta 16 *multi-head attention* dengan ukuran patch 14x14 piksel. Pada tabel 5, didapatkan hasil average skor dari 2 model yang dibandingkan dan berasal dari proses perhitungan kesamaan (similarity score) antara embedding gambar dan embedding teks. Terlihat model pertama unggul di average score dan waktu eksekusinya. Dalam penelitian ini, dilanjutkan hanya dengan model pertama.

Selain gambar, dibutuhkan data kategori yang berupa deskripsi gambar yang berfungsi untuk memberikan label atau kategori yang relevan dengan setiap gambar dalam dataset. Deskripsi gambar seperti “*A car accident on a road*”, “*A calm street with clear traffic*”, dan lainnya. Data deskripsi ini sangat penting dalam pendekatan *zero-shot* yang diterapkan oleh CLIP, karena model membutuhkan teks deskripsi untuk menghitung kesamaan antara gambar dan label. Beberapa deskripsi yang digunakan untuk gambar kategori kecelakaan (*Accident*) mencakup: “*A car accident involving multiple vehicles*” dan “*An overturned car on the highway*”, sementara untuk kategori non-kecelakaan (*Non-Accident*), deskripsi seperti “*A calm street with clear traffic*” dan “*A busy street with no accidents*” digunakan. Deskripsi ini membantu model dalam menentukan relevansi antara gambar dan teks untuk mencapai hasil yang akurat.

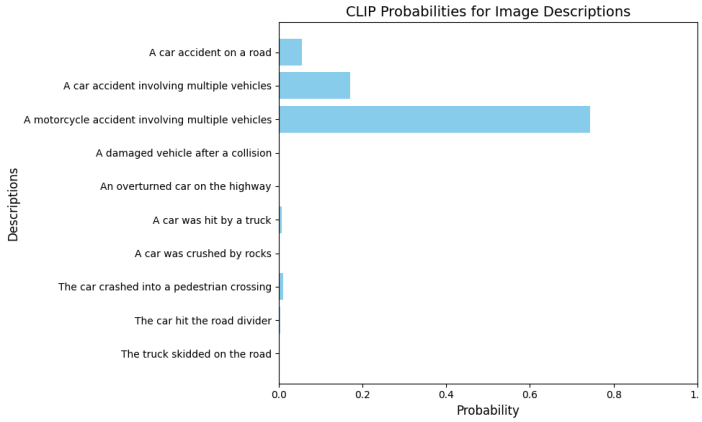
Kandidat Kemiripan (*Similarity*) Tertinggi

Proses deteksi dimulai dengan pemrosesan gambar dan perhitungan kesamaan antara gambar dan deskripsi teks yang relevan. Deskripsi teks yang digunakan mencakup berbagai kondisi yang menggambarkan kategori gambar, seperti "*A car accident on a road*" untuk gambar kecelakaan dan "*A calm street with clear traffic*" untuk gambar non-kecelakaan. Kesamaan antara gambar dan deskripsi dihitung menggunakan *cosine similarity*.

Sebagai contoh, untuk **Gambar 6**, deskripsi dengan probabilitas tertinggi adalah *"A car accident on a road"*, yang menunjukkan bahwa gambar tersebut paling cocok dengan kategori *Accident*. Begitu pula dengan **Gambar 8**, deskripsi dengan probabilitas tertinggi adalah *"A calm street with clear traffic"*, yang menunjukkan gambar tersebut paling cocok dengan kategori *Non-Accident*. Hasil probabilitas tertinggi ini diperoleh dari perhitungan kesamaan antara representasi gambar dan teks dalam ruang laten CLIP, dengan probabilitas yang lebih tinggi menandakan tingkat kecocokan yang lebih besar antara gambar dan deskripsi teks.

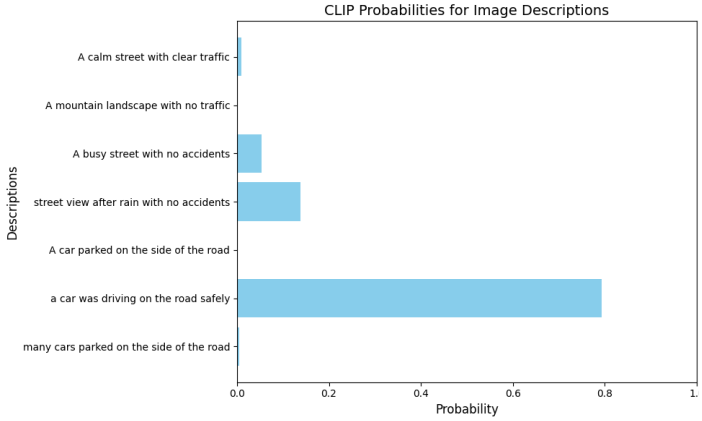
Gambar 6. Sampel gambar input *Accident*

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia



Gambar 7. Hasil probabilitas softmax untuk **gambar 6.**

Gambar 8. Sampel gambar input *Non-Accident*



Gambar 9. Hasil probabilitas softmax untuk **gambar 8**

Cosine similarity mengukur hubungan linear antara dua vektor, dalam hal ini antara representasi gambar dan teks dalam ruang laten CLIP. Nilai cosine similarity berkisar antara -1 hingga 1, dengan nilai yang lebih tinggi menunjukkan kesamaan yang lebih besar antara gambar dan teks. Nilai kemiripan ini bersifat relatif pada satu deskripsi, tidak mempertimbangkan kemiripan gambar dengan deskripsi lainnya.

Di sisi lain, ***softmax probability*** mengubah skor similarity menjadi probabilitas [0, 1], yang memudahkan dalam pelatihan dan evaluasi model dengan menggunakan fungsi loss. Probabilitas ini menunjukkan keyakinan model terhadap kesesuaian deskripsi tertentu dengan gambar, sambil mempertimbangkan hubungannya dengan semua deskripsi yang ada.

Tabel 3. Perbandingan probabilitas Cosine dan Softmax untuk Gambar 10

Deskripsi	Cosine	Probability Softmax
A car accident on a road	0.2811	0.1524
A car accident involving multiple vehicles	0.2806	0.0617
A motorcycle accident involving multiple vehicles	0.2935	0.0725
A damaged vehicle after a collision	0.2436	0.0002
An overturned car on the highway	0.2544	0.0143
A car was hit by a truck	0.2642	0.1214
A car was crushed by rocks	0.2561	0.0149
The car crashed into a pedestrian crossing	0.2630	0.5484
The car hit the road divider	0.2535	0.0009
The truck skidded on the road	0.2178	0.0013

Sebagai contoh, Tabel 3 menunjukkan hasil perhitungan *cosine similarity* dan probabilitas softmax untuk **Gambar 10**. Deskripsi *"The car crashed into a pedestrian crossing"* memiliki probabilitas tertinggi (0.5484), menandakan bahwa gambar tersebut paling cocok dengan deskripsi ini dalam konteks kecelakaan. Sebaliknya, deskripsi lainnya, seperti *"A damaged vehicle after a collision"*, memiliki skor yang lebih rendah baik dalam cosine similarity (0.2436) maupun probabilitas (0.0002), menunjukkan kecocokan yang lebih rendah dengan gambar tersebut.

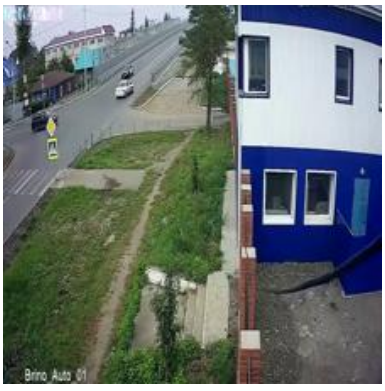
Contoh lainnya, Tabel 4 menunjukkan hasil perhitungan *cosine similarity* dan probabilitas softmax untuk **Gambar 11**. Deskripsi "*A car was driving on the road safely*" memiliki probabilitas tertinggi (0.6219), menandakan bahwa gambar tersebut paling cocok dengan deskripsi ini dalam konteks

non-kecelakaan. Sebaliknya, deskripsi lainnya, seperti "*A mountain landscape with no traffic*", memiliki skor yang lebih rendah baik dalam *cosine similarity* (0.1604) maupun probabilitas softmax (0.0001), menunjukkan kecocokan yang lebih rendah dengan gambar tersebut.

Gambar 10. Data *Accident*

Tabel 4. Perbandingan probabilitas Cosine dan Softmax untuk Gambar 10

Deskripsi	Cosine	Probability Softmax
A calm street with clear traffic	0.2180	0.0241
A mountain landscape with no traffic	0.1604	0.0001
A busy street with no accidents	0.2406	0.2307
Street view after rain with no accidents	0.2197	0.0285
A car parked on the side of the road	0.2262	0.0547
A car was driving on the road safely	0.2505	0.6219
Many cars parked on the side of the road	0.2231	0.0400



Gambar 11. Data Non-*Accident*

Setelah menghitung skor *similarity* dan probabilitas softmax, deskripsi dengan probabilitas tertinggi dianggap paling sesuai untuk merepresentasikan gambar tersebut. Hasil akhir dari proses ini disimpan dalam format CSV dengan tiga kolom: *image path, Best Description,* dan *Highest Probability*. Ini memungkinkan untuk melakukan evaluasi lebih lanjut dan analisis terhadap hasil deteksi kategori gambar berdasarkan deskripsi yang diberikan.

Tabel 5. Nilai Akurasi model CLIP Vit base patch 32

No	Top	Accuracy (%)
1	Top 1	32.00
2	Top 5	95.86

Hasil evaluasi model CLIP dengan pendekatan *zero-shot classification* pada Tabel 5, menunjukkan bahwa akurasi Top-1 mencapai 32.00%, sedangkan akurasi Top-5 mencapai 95.86%. Akurasi Top-1 mencerminkan seberapa sering prediksi teratas model sesuai dengan label sebenarnya, dalam hal ini model masih kesulitan untuk memilih prediksi yang benar sebagai jawaban utama. Sementara itu, akurasi Top-5 menunjukkan bahwa dalam sebagian besar kasus (hampir semua), label sebenarnya muncul di antara 5 prediksi teratas yang diberikan oleh model. Hal ini mengindikasikan bahwa model CLIP cukup baik dalam memahami hubungan antara teks dan gambar secara umum.

Kesimpulan

Proses pengolahan gambar untuk klasifikasi "*Accident*" dan "*Non Accident*" menggunakan pendekatan berbasis model CLIP. Setelah melalui tahap preprocessing, seperti resizing dan normalisasi data gambar, hasil ini menunjukkan bahwa model CLIP ViT base patch 32 dapat mencapai akurasi Top 1 sebesar 32% dan akurasi Top 5 sebesar 95,86%. Walaupun akurasi Top 1 masih memerlukan optimasi lebih lanjut, hasil Top 5 menunjukkan bahwa model dapat secara konsisten mencakup jawaban yang tepat dalam lima prediksi teratas. Pada penelitian berikutnya dapat diperbaiki dengan meningkatkan akurasi melalui optimasi dataset atau fine-tuning model. Lalu penambahan lebih banyak data dari berbagai deteksi kecelakaan juga dapat meningkatkan generalisasi model untuk aplikasi dunia nyata.

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Konflik Kepentingan

Penulis menyatakan tidak ada konflik kepentingan dalam penulisan naskah ini.

Ucapan Terima kasih

Dengan penuh rasa syukur dan penghargaan, kami ingin menyampaikan ucapan terima kasih yang mendalam kepada berbagai pihak yang telah berkontribusi dalam penelitian dan penulisan artikel tentang Deteksi Gambar Kecelakaan Menggunakan CLIP dengan Pendekatan Zero-Shot" ini. Penelitian ini tidak akan tercapai tanpa bantuan dan dukungan dari banyak pihak. Terima kasih kami sampaikan kepada Institut Teknologi Sumatera dan Program Studi Sains Data atas dukungan fasilitas penelitian yang memadai.

References

[1] S. M. Prasetiyo, A. Rahmayani, and A. Melania, “Artificial Intelligence dalam Kesehatan dan Keselamatan Kerja di Bidang Kelistrikan,” OKTAL: Jurnal Ilmu Komputer dan Science, vol. 2, no. 8, pp. 2214–2216, Aug. 2023. [Online]. Available: https://journal.mediapublikasi.id/index.php/oktal

[2] A. Rezky, A. Bagir, D. Pamerean, and F. Makhrus, “Deteksi Kecelakaan Lalu Lintas Otomatis Pada Rekaman CCTV Indonesia Menggunakan Deep Learning,” Buletin Pagelaran Mahasiswa Nasional Bidang Teknologi Informasi dan Komunikasi, vol. 1, no. 1, pp. 1–5, Oct. 2023.

[3] T. Penulis, “Deteksi Kecelakaan Berbasis Perubahan Pola Berkendara Menggunakan Parameter Arah dan Kecepatan Kendaraan,” M.S. thesis, Universitas (nama universitas), 2023.

[4] C. Kay, "Accident Detection from CCTV Footage," Kaggle, 2021. https://www.kaggle.com/datasets/ckay16/accident-detection-from-cctv-footage.

[5] A. Radford **et al.**, "CLIP (Contrastive Language–Image Pre-training)," in Proc. Int. Conf. Mach. Learn. (ICML), 2021.

[6] OpenAI, "CLIP: Connecting text and images," 2021. [Online]. Available: https://openai.com/research/clip.

[7] Z. Tu **et al.**, "A Closer Look at the Robustness of Contrastive Language-Image Pre-Training (CLIP)," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023.

[8] X. Wei **et al.**, "iCLIP: Bridging Image Classification and Contrastive Language-Image Pre-training for Visual Recognition," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2023.

[9] F. Pourpanah et al., "A Review of Generalized Zero-Shot Learning Methods," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, pp. 4051-4070, 1 April 2023, doi: 10.1109/TPAMI.2022.3191696 buat ieee

[10] Han **et al.**, "Contrastive Embedding for Generalized Zero-Shot Learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2021.

[11] N. M. Abdi, R. K. Setiawan, dan I. A. Triwahyu, "Peningkatan kualitas citra digital menggunakan metode super resolusi pada domain spasial," Jurnal Media Elektrik, vol. 11, no. 2, pp. 135-139, 2021. [Online]. Tersedia: https://media.neliti.com.

[12] S. Harjoko dan A. S. Wardana, "Pengolahan citra digital untuk peningkatan kualitas gambar dengan metode histogram equalization," Jurnal Informatika, vol. 16, no. 1, pp. 55-63, 2020. [Online]. Tersedia: https://core.ac.uk.

[13] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929.

[14] OpenAI, "CLIP ViT-Large-Patch14 (336px): A Vision Transformer Model," Hugging Face Model Repository, 2021. [Online]. Available: https://huggingface.co/openai/clip-vit-large-patch14-336.