Original Article

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

Evaluasi Performa Model CLIP dalam Klasifikasi Gambar-Teks Spesies, Perilaku, Lokasi, dan Waktu pada Ternak dengan Pendekatan *Zero-Shot Learning*

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

Open Access

DOI: 10.35472/x0xx0000

Wahyudiyanto[*a], Nadilla Andhara Putri [b], Yunaena Maratul Kirom [c], Revaldo Dafa Fahmindo [d], Shula Talitha Ardhya Putri [e], Ibnu Farhan Al-Ghifari [f]

[a, b, c, d, e, f] *Program Studi Sains Data, Institut Teknologi Sumatera, Lampung 3565*

** Corresponding E-mail: wahyudiyanto.121450040@student.itera.ac.id*

Abstract: The development of Artificial Intelligence (AI) has introduced multimodal models like CLIP, capable of image-text classification using a zero-shot learning approach. This study aims to evaluate the performance of four CLIP models—RN50, ViT-B/16, ViT-L/14, and ViT-B/32—in classifying livestock species, behavior, location, and time. The dataset comprises 279 livestock images processed with augmentation techniques such as rotation, flipping, and cropping. Predictions are made by calculating cosine similarity between image and text features. Results show that ViT-L/14 achieved the highest accuracy in species (79%), behavior (62%), and time (77%), but performed poorly in location (43%). ViT-B/16 demonstrated balanced performance across categories, while RN50 and ViT-B/32 excelled in species and location. Confidence analysis indicates that RN50 and ViT-B/16 produced more consistent predictions compared to other models. This study highlights the potential of CLIP models, particularly ViT-L/14, for text and image-based classification in livestock management. Further improvements through fine-tuning are recommended to enhance accuracy and reduce bias in certain categories.

Keywords: *CLIP, image-text classification, livestock, animal behavior, zero-shot learning*

Abstrak: Perkembangan Artificial Intelligence (AI) telah menghadirkan model multimodal seperti CLIP yang mampu melakukan klasifikasi gambar-teks dengan pendekatan *zero-shot learning*. Penelitian ini bertujuan mengevaluasi performa empat model CLIP—RN50, ViT-B/16, ViT-L/14, dan ViT-B/32 dalam klasifikasi spesies ternak, perilaku, lokasi, dan waktu. Dataset terdiri atas 279 gambar ternak yang diproses dengan teknik augmentasi seperti rotasi, flipping, dan cropping. Prediksi dilakukan dengan menghitung kesamaan kosinus antara fitur gambar dan teks. Hasil menunjukkan bahwa ViT-L/14 memiliki akurasi tertinggi pada kategori spesies (79%), perilaku (62%), dan waktu (77%), meskipun rendah pada kategori lokasi (43%). ViT-B/16 menunjukkan performa seimbang di berbagai kategori, sementara RN50 dan ViT-B/32 unggul pada spesies dan lokasi. Analisis confidence mengindikasikan bahwa RN50 dan ViT-B/16 memiliki konsistensi prediksi lebih baik dibanding model lainnya. Penelitian ini membuktikan potensi model CLIP, khususnya ViT-L/14, dalam klasifikasi berbasis teks dan gambar di peternakan. Penyempurnaan lebih lanjut melalui *fine-tuning* diperlukan untuk meningkatkan akurasi dan mengurangi bias pada kategori tertentu.

Kata Kunci: CLIP, klasifikasi gambar-teks, hewan ternak, perilaku hewan, zero-shot learning

Copyright © 2022 Journal of Science and Applicative Technology.
Content from this work may be used under the terms of the Creative Commons Attribution-NonCommercial 4.0 International Licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI. Published under licence by Journal of Science and Aplicative Technology (JSAT).

Pendahuluan

Kemajuan *Artificial Intelligence* (AI) telah mendorong pengembangan model multimodal seperti CLIP (*Contrastive Language-Image Pre-Training*), yang unggul dalam *zero-shot learning* [1]. Model ini dapat menyelesaikan tugas tanpa pelatihan tambahan, cukup dengan memanfaatkan representasi multimodal yang dipelajari selama pelatihan awal [2]. CLIP telah menunjukkan kinerja luar biasa dalam berbagai tugas, termasuk klasifikasi gambar berbasis teks, pencocokan gambar-tekstual dan transfer pembelajaran ke domain baru tanpa memerlukan pelabelan manual [3].

Salah satu aplikasi menarik dari CLIP adalah dalam klasifikasi perilaku hewan ternak, seperti makan, istirahat dan bergerak, yang relevan dengan manajemen peternakan modern. Kemampuan *zero-shot*-nya memungkinkan deteksi perilaku spesifik pada domain yang belum pernah dilatih sebelumnya [4].

Namun, penerapan CLIP dalam klasifikasi hewan ternak masih belum banyak dieksplorasi. Padahal, teknologi ini berpotensi besar dalam meningkatkan kesejahteraan hewan dan efisiensi peternakan. Pendekatan fine-tuning pada dataset spesifik telah terbukti dapat meningkatkan akurasi CLIP, terutama untuk domain seperti peternakan, di mana variasi spesies dan perilaku hewan berbeda dari data pelatihan awal model [5].

Studi ini diangkat dari penelitian terdahulu yang berfokus pada pengembangan model berbasis data teks dan pemahaman makna hubungan antara teks dan gambar. Dataset yang digunakan diambil dari studi terdahulu yaitu *"Zero-shot animal behaviour classification with image-text foundation models"* oleh Dussert et al. (2024), yang mengklasifikasikan spesies dan perilaku berbagai hewan [6].

Untuk itu, Penelitian ini akan mengevaluasi performa empat varian model CLIP yaitu RN50, ViT-L/14, ViT-B/16, dan ViT-B/32 dalam mengklasifikasikan hewan ternak. Sebagai pembaharuan, penelitian ini menambahkan kelas baru yang sebelumnya hanya "spesies" dan "perilaku", kemudian ditambahkan "tempat" dan "waktu" agar pemantauan ternak menjadi lebih spesifik. Selain itu teks yang digunakan dalam prediksi gambar tidak didefinisikan secara langsung, melainkan diambil dari kombinasi 11 sub -kelas yang nantinya akan membentuk satu teks utuh yang mendeskripsikan gambar.

Metode

Dataset yang digunakan dalam penelitian ini diambil dari penelitian *"Zero-shot animal behavior classification with image-text foundation models*" [5], yang menggunakan *dataset CREA Mont-Blanc camera trap*. Dataset asli berisi 131.537 gambar yang dikumpulkan antara 2017-2020, dengan 24.580 kejadian. Fokus penelitian ini pada tiga spesies ternak: *cow* (71 gambar), *goat* (92 gambar), dan *sheep* (116 gambar), dengan total 279 gambar. Dataset ini dipilih dengan jumlah terbatas karena komputasi tinggi yang dibutuhkan dalam pemrosesan gambar.

Augmentasi Gambar

Augmentasi gambar adalah teknik yang digunakan untuk memperluas variasi data dengan memodifikasi gambar tanpa mengubah label aslinya [8]. Dalam penelitian ini, augmentasi dilakukan untuk meningkatkan robustness (ketahanan) model CLIP terhadap perbedaan kondisi visual.

1. Rotasi acak : Gambar diputar -30 - 30 derajat
2. Penyesuaian Kecerahan dan Kontras: Mengubah kecerahan dan kontras dalam rentang 0,5 sampai 1,5
3. Flip Horizontal: Membalik gambar secara horizontal untuk memperkenalkan simetri spasial.

Setelah augmentasi, gambar dinormalisasi menggunakan preprocessing bawaan CLIP, termasuk konversi ke format RGB dan normalisasi nilai piksel agar sesuai dengan input yang diharapkan model. Tahapan ini memastikan model dapat menangani variasi gambar dengan lebih baik dalam klasifikasi spesies, perilaku, lokasi, dan waktu ternak.

Kombinasi Teks Kelas untuk Prediksi Gambar-Teks

Dalam penelitian ini, model CLIP digunakan untuk memprediksi label kelas gambar berdasarkan kombinasi kelas yang meliputi spesies (animal), perilaku (behavior), lokasi (place), dan waktu (time). Di setiap kelas, terdapat sub-kelas lagi yang dapat dilihat pada tabel berikut.

Tabel 1. Kelas prediksi

Animals	*Behaviors*	*Places*	*Times*
Cow	*Eating*	*In the field*	*Day*
Goat	*Moving*	*On the hill*	*Night*
Sheep	*Resting*	*In the forest*	

Untuk menghasilkan label teks yang digunakan dalam memprediksi gambar, terdapat dua langkah utama:

1. Tokenisasi dan Ekstraksi Fitur Teks :

   Setiap sub-kelas teks, seperti "cow", "eating", "in the field", dan "day", diubah menjadi representasi numerik melalui proses tokenisasi. Proses tokenisasi dilakukan menggunakan tokenizer bawaan CLIP, yang mengubah teks menjadi token-token yang dapat dipahami oleh model. Selanjutnya, model CLIP mengekstrak fitur embedding dari token tersebut untuk membentuk vektor representasi semantik.

2. Kombinasi Teks untuk Prediksi :

   Dengan 4 kelas utama dan 11 sub-kelas yang terdiri dari 3 spesies, 3 perilaku, 3 lokasi, dan 2 waktu, kombinasi teks yang dihasilkan adalah 54 kombinasi unik. Kombinasi ini diperoleh dengan menggabungkan setiap sub-kelas dari empat kategori menggunakan permutasi terstruktur.

   Contoh kombinasi teks yang dihasilkan :

   - *"The cow is eating in the field in the day time"*
   - *"The goat is moving on the hill at night time"*
   - *"The sheep is resting in the forest in the day time"*

   Dengan kombinasi ini, model CLIP menghitung kesamaan kosinus antara fitur gambar yang diinput dan representasi teks yang dihasilkan. Teks dengan nilai kesamaan tertinggi dipilih sebagai prediksi terbaik.

Prediksi Berdasarkan Kesamaan Kosinus

Prediksi gambar dilakukan dengan menghitung kesamaan kosinus antara fitur gambar dan fitur teks yang diekstrak oleh model CLIP. Kesamaan kosinus digunakan karena mampu mengukur kemiripan antara dua vektor dalam ruang berdimensi tinggi, yang pada dasarnya mewakili hubungan semantik antara gambar dan teks [7].

Kesamaan kosinus antara dua vektor A (fitur gambar) dan B (fitur teks) dinyatakan sebagai berikut :

$$\text{cosine similarity}\,(\mathrm{A},\mathrm{B}) = \frac{\mathrm{A}\cdot\mathrm{B}}{||\mathrm{A}||\,||\mathrm{B}||} \qquad (1)$$

Keterangan :

- A·B: Dot product antara dua vektor A dan B. Ini mengukur hasil perkalian linear elemen- elemen pada kedua vektor.
- ||A||: Norma L2 atau panjang vektor A, dihitung sebagai || A|| = ΣA
- ||B||: Norma L2 atau panjang vektor B.
- Hasil akhirnya berupa nilai antara -1 dan 1, di mana:
  - 1 menunjukkan kemiripan maksimum.
  - 0 menunjukkan tidak ada hubungan.
  - -1 menunjukkan hubungan berlawanan.

Karena vektor hasil CLIP sudah dinormalisasi, panjang vektor ||A|| dan ||B|| selalu bernilai 1. Sehingga persamaan kesamaan kosinus menjadi lebih sederhana:

$$\text{cosine similarity}(\mathrm{A},\mathrm{B}) = \mathrm{A}\cdot\mathrm{B} \qquad (2)$$

Contrastive Language-Image Pretraining (CLIP)

Contrastive Language-Image Pre-Training (CLIP) adalah model multimodal yang dikembangkan untuk memahami hubungan antara teks dan gambar menggunakan pendekatan *contrastive learning*. CLIP dilatih pada data pasangan gambar-teks dalam skala besar yang diperoleh dari internet. Berbeda dengan model tradisional yang memerlukan pelabelan spesifik, CLIP mampu belajar representasi visual dan tekstual secara simultan dengan memaksimalkan kesamaan vektor antara pasangan yang relevan dan meminimalkan kesamaan untuk pasangan yang tidak sesuai [9][4].

Arsitektur CLIP terdiri dari encoder gambar dan encoder teks. Encoder gambar menggunakan arsitektur seperti ResNet atau Vision Transformer (ViT) untuk mengubah gambar menjadi vektor berdimensi tinggi, sedangkan encoder teks berbasis Transformer memproses teks menjadi vektor dengan dimensi yang sama. Kedua vektor dari gambar dan teks ini kemudian dipetakan ke dalam ruang representasi yang sama sehingga perhitungan kesamaan kosinus dapat digunakan untuk menentukan relevansi antar pasangan gambar-teks [10].

Keunggulan utama CLIP adalah kemampuannya dalam zero-shot learning, yaitu melakukan klasifikasi pada kategori baru yang tidak pernah dilihat selama pelatihan hanya dengan menyediakan deskripsi tekstualnya. Kemampuan ini memungkinkan CLIP untuk menyelesaikan berbagai tugas seperti klasifikasi gambar, pengambilan kembali gambar berbasis teks (*image retrieval*), dan pengenalan objek tanpa perlu dilatih ulang untuk setiap domain baru [4]. Selain itu, CLIP terbukti lebih fleksibel dibandingkan model vision klasik karena memanfaatkan supervisi bahasa alami dalam jumlah besar dari internet, bukan dataset berlabel yang spesifik [2].

Copyright © 2022 Journal of Science and Applicative Technology
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

RN50 (ResNet-50)

RN50 adalah arsitektur berbasis Residual Network (ResNet-50) yang digunakan sebagai encoder gambar dalam model CLIP. ResNet-50 merupakan jaringan konvolusional (CNN) dengan 50 lapisan yang memanfaatkan residual blocks untuk memecahkan masalah *vanishing gradient* pada jaringan yang sangat dalam [11]. Keunggulan utama RN50 terletak pada kemampuannya untuk mengekstraksi fitur visual secara efisien, terutama untuk objek dengan struktur yang jelas dan sederhana. Dengan pendekatan hierarkis dalam mendeteksi fitur dari tingkat rendah hingga tinggi, RN50 memiliki performa yang stabil dalam klasifikasi gambar di berbagai domain [9]. Namun, dibandingkan dengan Vision Transformer (ViT), RN50 cenderung kurang fleksibel dalam menangkap hubungan global antar piksel, sehingga performanya sedikit lebih rendah pada data dengan konteks visual yang kompleks [12].

ViT-B/16 (Vision Transformer - Base, Patch 16)

ViT-B/16 adalah arsitektur Vision Transformer yang digunakan sebagai encoder gambar dalam model CLIP. Model ini membagi gambar menjadi patch berukuran 16x16 piksel, diubah menjadi embedding linear, dan diproses melalui lapisan Transformer untuk menangkap hubungan global antar-patch melalui mekanisme self-attention. Dengan total 12 lapisan Transformer dan sekitar 86 juta parameter, ViT-B/16 mampu memahami konteks visual secara menyeluruh, menjadikannya efektif dalam menangani gambar dengan detail kompleks [6][13]. Namun, ViT-B/16 memerlukan jumlah data pelatihan yang besar untuk mencapai performa optimal, karena model Transformer cenderung overfitting pada dataset kecil [14]. Dalam CLIP, ViT-B/16 menunjukkan performa yang seimbang di berbagai tugas klasifikasi gambar-teks, seperti spesies, perilaku, lokasi, dan waktu, berkat kemampuan representasinya yang mendalam dan fleksibel [15].

ViT-L/14 (Vision Transformer - Large, Patch 14)

ViT-L/14 adalah varian Vision Transformer dengan ukuran lebih besar dibandingkan ViT-B/16, yang digunakan sebagai encoder gambar dalam model CLIP. Model ini membagi gambar menjadi patch berukuran 14x14 piksel, sehingga mampu menangkap detail visual yang lebih halus dibandingkan varian dengan patch lebih besar. Dengan 24 lapisan Transformer, 16 kepala *self-attention* per lapisan, dan sekitar 307 juta parameter, ViT-L/14 memiliki kapasitas yang lebih tinggi untuk memahami representasi visual kompleks dalam gambar [6][13]. Keunggulan ViT-L/14 terletak pada kemampuannya dalam menangkap informasi spasial global dan lokal secara lebih baik, sehingga menghasilkan representasi fitur berkualitas tinggi. Namun, ukuran model yang besar memerlukan sumber daya komputasi lebih banyak serta dataset pelatihan dalam skala besar [14]. Dalam model CLIP, ViT-L/14 terbukti unggul dalam tugas zero-shot learning, terutama pada klasifikasi yang membutuhkan analisis detail visual dan konteks yang kompleks [2].

ViT-B/32 (Vision Transformer - Base, Patch 32)

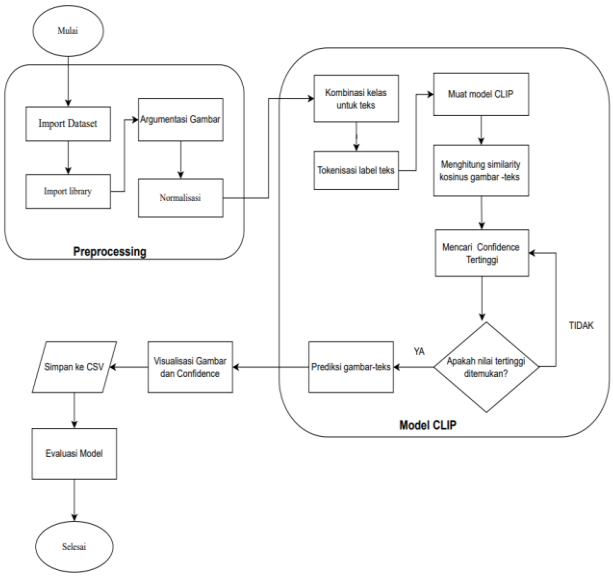
ViT-B/32 adalah varian Vision Transformer dengan arsitektur serupa ViT-B/16 namun menggunakan patch berukuran 32x32 piksel. Dengan ukuran patch yang lebih besar, ViT-B/32 memiliki efisiensi komputasi yang lebih baik dibandingkan ViT-B/16 dan ViT-L/14, namun dengan resolusi fitur yang lebih rendah karena detail visual yang halus cenderung diabaikan [6][13]. Model ini memiliki 12 lapisan Transformer dan sekitar 86 juta parameter, sama seperti ViT-B/16, tetapi lebih cepat dalam proses inferensi karena jumlah patch yang lebih sedikit. Keunggulan ViT-B/32 adalah kemampuannya untuk menangani dataset skala besar dengan sumber daya komputasi yang lebih ringan, meskipun performanya dapat menurun pada tugas yang membutuhkan analisis visual dengan detail kompleks [14]. Dalam implementasi CLIP, ViT-B/32 menunjukkan performa yang baik untuk tugas zero-shot learning pada gambar dengan struktur visual sederhana dan kategori yang lebih jelas, menjadikannya pilihan yang efisien dalam berbagai aplikasi klasifikasi gambar berbasis teks [2].

Evaluasi

Evaluasi dilakukan dengan membandingkan prediksi model dengan label *ground truth*. Akurasi dihitung untuk setiap kategori secara terpisah, yaitu hewan, perilaku, tempat, dan waktu. Akurasi keseluruhan juga dihitung dengan mempertimbangkan kecocokan prediksi pada semua kategori. Hasil akurasi dan prediksi dari keempat model CLIP—RN50, ViT-B/16, ViT-L/14, dan ViT-B/32 kemudian dibandingkan untuk menilai performa masing-masing model dalam tugas klasifikasi gambar ternak.

Hasil confidence score juga divisualisasikan untuk analisis lebih lanjut, yang menggambarkan performa model pada masing-masing kategori. Perbandingan akurasi dan hasil prediksi keempat model dilakukan untuk mendapatkan wawasan mengenai model mana yang memberikan performa terbaik pada tugas klasifikasi ini.

Diagram Alir



Gambar 1. Diagram alir

Hasil dan Pembahasan

Sebelum mendapatkan hasil yang diinginkan ada beberapa tahapan dalam proses pengolahan data untuk model CLIP ini.

Hasil Kombinasi Teks untuk Model CLIP

Dalam penelitian ini, pelabelan teks digunakan untuk menguji performa model CLIP dalam menghubungkan gambar dengan deskripsi teks. Label teks dibuat dari kombinasi empat kategori yaitu hewan (*cow, goat, sheep*), perilaku (*eating, moving, resting*), tempat *(in the field, in the hill, in the forest*), dan waktu (*day, night*). Dengan menggunakan kombinasi *cartesian product* menghasilkan 54 kombinasi teks.

Tabel 2. Hasil Kombinasi Teks Prediksi

No	Teks Prediksi
1	*The cow is eating in the field in the day time*
2	*The cow is eating in the field in the night time*
…	…
53	*The sheep is resting in the forrest in the day time*
54	*The sheep is resting in the forrest in the night time*

Hasil *Preprocessing* dan *Augmentasi* Gambar

Pada tahap *preprocessing* dan augmentasi gambar bertujuan untuk memastikan bahwa dataset gambar memiliki format dan variasi yang optimal untuk dilakukan pelatihan model CLIP selanjutnya. Augmentasi dilakukan dengan transformasi acak seperti *horizontal flip,* rotasi, dan *resized crop*. Sehingga menciptakan variasi dari gambar asli dan meningkatkan kemampuan model dalam mengenali objek meskipun dalam posisi atau kondisi yang berbeda.

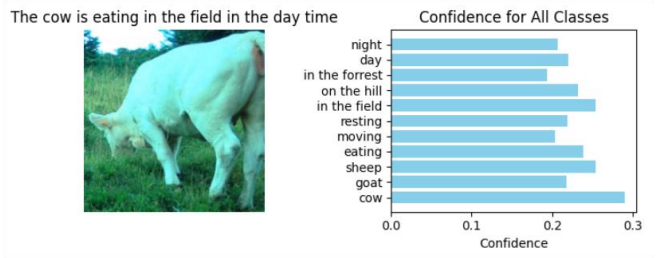
Gambar 2. Hasil Preprocessing dan Augmentasi Gambar

Pada gambar diatas hasil dari *preprocessing* dan augmentasi gambar, terlihat jelas gambar pertama adalah gambar asli yang tidak mengalami perubahan, sedangkan gambar kedua dan ketiga hasil dari *preprocessing* dengan augmentasi. Gambar yang dihasilkan terpotong sebagian akibat proses random resized crop.

Ekstraksi Fitur Teks-Gambar dan Perhitungan Persamaan Kosinus

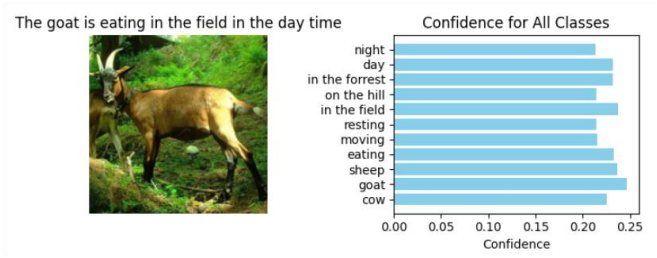
Pada tahap ini CLIP mengekstrak fitur teks dan gambar sebagai vektor numerik, lalu menghitung kesamaan kosinus antara keduanya (*confidence*) . Kategori dengan nilai kesamaan tertinggi akan menjadi prediksi terbaik, karena menunjukkan kemiripan terbesar antara gambar dan teks deskripsi.

Copyright © 2022 Journal of Science and Applicative Technology
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia



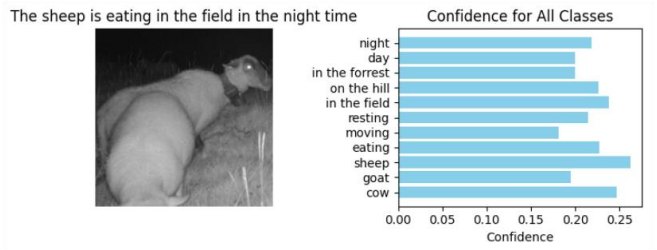
Gambar 3. Hasil nilai confidence untuk sapi (*cow*)

Pada gambar diatas, model dengan yakin memprediksi cow, eating, in the field dan day sebagai nilai *confidence* tertinggi. Nilai tertinggi pada kelas ini kemudian dijadikan sebagai teks prediksi untuk gambar 3.



Gambar 4. Hasil nilai confidence untuk kambing (*goat*)

Pada gambar 4 prediksi yang paling tinggi adalah kambing (*goat*), namun terlihat ada keraguan dalam mengidentifikasi jenis hewan kambing dengan hewan lain. Ini mungkin terjadi karena karakteristik kambing yang hampir mirip dengan sapi dan domba. Model juga kurang tepat dalam mengidentifikasi tempat, yang harusnya *in the forrest*, namun diprediksi *in the field*. Nilai confidence pada tiap kelas kemudian digabung dan dijaikan teks prediksi untuk gambar 4.



Gambar 5. Hasil nilai confidence untuk domba (*sheep*)

Pada gambar 5 model berhasil memprediksi domba (*sheep*) dengan confidence mencapai 0.25+, Selain itu eating, int the field dan night memiliki nilai confidence tertinggi, sehingga kelas ini dijadikan teks prediksi untuk gambar 5. Dapat dilihat juga, model sangat baik dalam membedakan waktu *day* dan *night*.

Hasil Klasifikasi Setiap Model

- **RN50**

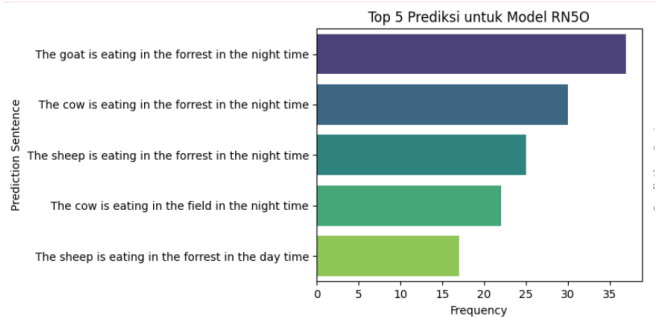
Gambar 6. Hasil Prediksi Model RN50

Model CLIP RN50 menunjukkan performa baik dalam mengenali spesies ternak dan perilakunya. Dari 279 gambar, model menghasilkan 31 prediksi yang 100% benar, dengan 24 teks bergambar dan 30 teks tanpa gambar.

Tabel 3. Akurasi Prediksi Model RN50

Kategori	Akurasi
Cow	0.767
Goat	0.828
Sheep	0.86
Eating	0.419
Moving	0.369
Resting	0.771
In the Field	0.624
On the Hill	0.785
In the Forest	0.738
Day	0.326
Night	0.326

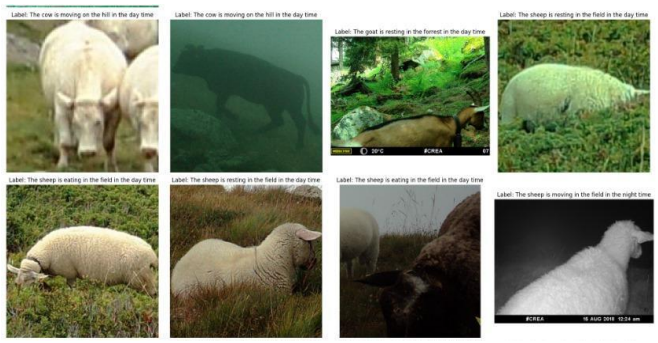
Model RN50 menunjukkan performa yang baik pada kategori animal, dengan akurasi 0.77 (*cow*), 0.83 (*goat*), dan 0.86 (*sheep*). Namun, pada kategori behavior, akurasi rendah di 0.42 (*eating*), 0.37 (*moving*), dan lebih tinggi di 0.77 (*resting*). Untuk kategori *place*, model mencatat akurasi 0.62 (*in the field*), 0.78 (*on the hill*), dan 0.74 (*in the forest*). Sementara itu, akurasi kategori time hanya mencapai 0.33 untuk *day* dan *night*, menunjukkan kelemahan dalam menangkap informasi waktu.



Gambar 7. Frekuensi Prediksi Model RN50

Berdasarkan frekuensi prediksi, Model RN50 menunjukkan kecenderungan prediksi pada aktivitas malam hari dan lingkungan hutan. Hal ini terlihat dari dominasi prediksi yang berfokus pada perilaku makan di malam hari.

- **ViT-B/32**



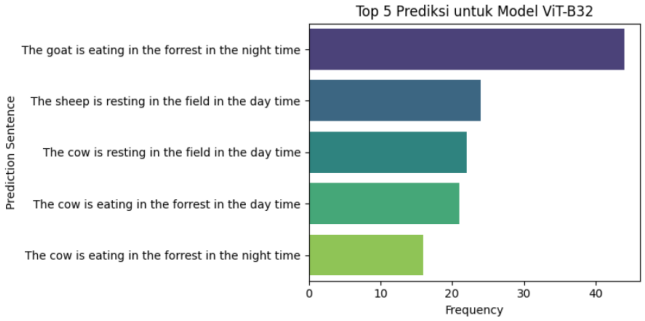
Gambar 8. Hasil Prediksi Model ViT-B/32

Model CLIP ViT-B/32 menunjukkan peningkatan akurasi dibandingkan RN50. Dari 279 gambar, model ini menghasilkan 34 prediksi yang 100% benar, dengan 25 teks bergambar dan 29 teks tanpa gambar.

Tabel 4. Akurasi Prediksi Model ViT-B/32

Kategori	Akurasi
Cow	0.735
Goat	0.835
Sheep	0.835
Eating	0.513
Moving	0.394
Resting	0.674
In the Field	0.685
On the Hill	0.785
In the Forest	0.778
Day	0.563
Night	0.563

Model ViT-B/32 menunjukkan performa stabil dengan akurasi tertinggi pada kategori place, yaitu 0.68 *(in the field*), 0.78 *(on the hill*), dan 0.78 *(in the forest*). Pada kategori animal, akurasi mencapai 0.73 (*cow*), 0.84 (*goat*), dan 0.84 (*sheep*). Namun, kategori behavior masih lemah, dengan akurasi 0.51 (*eating*), 0.39 (*moving*), dan 0.67 (*resting*). Kategori time menunjukkan akurasi 0.56 untuk *day* dan *night*, menunjukkan performa lebih baik dibanding RN50.



Gambar 9. Frekuensi Prediksi Model VIT-B32

Berdasarkan frekuensi prediksi, Model ViT-B/32 memberikan hasil yang lebih seimbang antara waktu siang dan malam, dengan fokus pada aktivitas makan di berbagai lokasi. Model ini menunjukkan keyakinan yang lebih stabil dibandingkan RN50, dengan distribusi confidence yang lebih terpusat di sekitar 0.22, khususnya pada kategori tempat.

- **ViT-L/14**

Gambar 10. Hasil Prediksi Model ViT-L/14

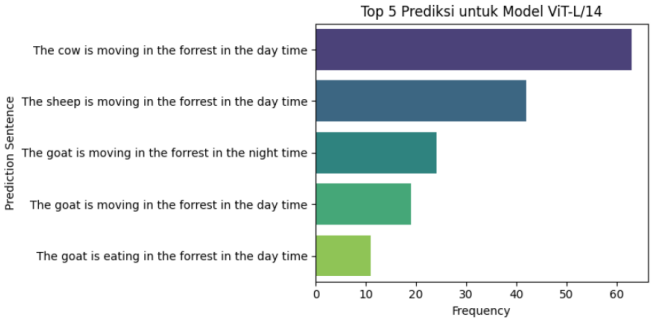
Copyright © 2022 Journal of Science and Applicative Technology
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Model CLIP ViT-L/14, dengan konfigurasi lebih besar dan kapasitas yang tinggi, menunjukkan performa terbaik dalam presisi. Dari 279 gambar, model ini menghasilkan 70 prediksi yang 100% benar, dengan 32 teks bergambar dan 22 teks tanpa gambar.

Tabel 5. Akurasi Prediksi Model ViT-L/14

Kategori	Akurasi
Cow	0.835
Goat	0.832
Sheep	0.91
Eating	0.735
Moving	0.656
Resting	0.857
In the Field	0.548
On the Hill	0.803
In the Forest	0.502
Day	0.774
Night	0.774

Model ViT-L/14 mencatat akurasi tertinggi di kategori animal, yaitu 0.84 (cow), 0.83 (*goat*), dan 0.91 (*sheep*). Pada kategori behavior, model ini unggul dengan akurasi 0.73 (*eating*), 0.66 (*moving*), dan 0.86 (*resting*). Namun, akurasi kategori place lebih rendah dibanding model lainnya, yakni 0.55 *(in the field*), 0.80 (*on the hill)*, dan 0.50 *(in the forest*). Kategori time menunjukkan akurasi tertinggi sebesar 0.77 untuk *day* dan *night*.



Gambar 11. Frekuensi Prediksi Model ViT-L/14

Model ViT-L/14 unggul dalam menangkap prediksi aktivitas yang terjadi pada siang hari, terutama di lokasi hutan. Prediksi model lebih variatif dengan frekuensi tinggi pada perilaku bergerak dan makan. Confidence model memiliki rentang yang lebih lebar dibandingkan ViT-B/32, namun hal ini sejalan dengan kompleksitas dan presisi prediksi yang lebih baik, terutama pada kategori spesies ternak dan perilaku.

- **ViT-B/16**

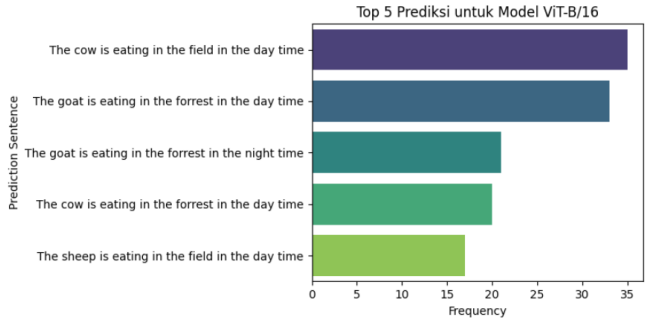
Gambar 12. Hasil Prediksi Model ViT-B/16

Model CLIP ViT-B/16 menunjukkan peningkatan detail pada gambar, dengan ukuran patch yang lebih kecil (16 piksel) dibandingkan dengan ViT-B/32. Dari 279 gambar, model ini menghasilkan 32 prediksi benar, dengan 23 teks bergambar dan 31 teks tidak bergambar.

Tabel 6. Akurasi Prediksi Model ViT-B/16

Kategori	Akurasi
Cow	0.735
Goat	0.835
Sheep	0.835
Eating	0.513
Moving	0.394
Resting	0.674
In the Field	0.685
On the Hill	0.785
In the Forest	0.778
Day	0.563
Night	0.563

Model ViT-B/16 menunjukkan performa paling seimbang di seluruh kategori. Akurasi kategori animal mencapai 0.79 (*cow*), 0.85 (*goat*), dan 0.87 (*sheep*). Pada kategori *behavior*, model mencatat akurasi 0.43 (*eating*), 0.39 (*moving*), dan 0.80 (*resting*). Kategori place memiliki akurasi 0.66 *(in the field*), 0.84 *(on the hill*), dan 0.73 (*in the forest*). Akurasi kategori time adalah 0.72 untuk *day* dan *night*, menunjukkan performa yang stabil dan konsisten.

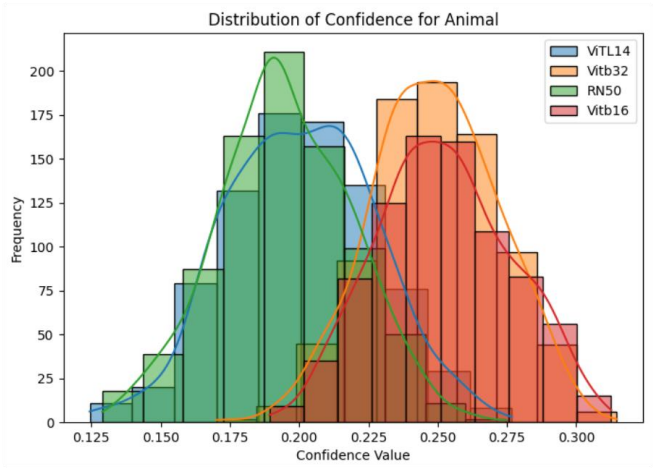


Gambar 13. Frekuensi Prediksi Model ViT-B/16

Model ViT-B/16 menunjukkan performa paling seimbang di semua kategori, dengan dominasi prediksi pada aktivitas siang hari di berbagai lokasi. Distribusi confidence lebih stabil dan terkonsentrasi di sekitar 0.25, mencerminkan keyakinan yang konsisten dalam memberikan hasil prediksi. Model ini mampu menangkap detail lingkungan dengan baik, menunjukkan peningkatan performa dibandingkan RN50 dan ViT-B/32.

Distribusi Confidence

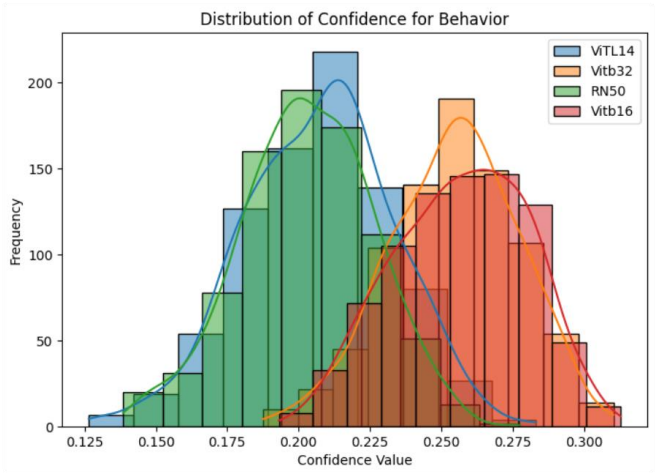
- **Hewan (*Animal*)**



Gambar 14. Distribusi confidence animal

Pada kategori animal, ViT-B/16 dan RN50 memiliki confidence tertinggi di kisaran 0.25–0.30, menunjukkan keyakinan tinggi dalam mengenali spesies ternak. Sementara itu, ViT-L/14 memiliki distribusi lebih lebar di 0.2–0.25, menandakan variasi prediksi yang lebih besar. ViT-B/32 memiliki confidence terendah di kisaran 0.175–0.225, menunjukkan keyakinan yang lebih lemah.

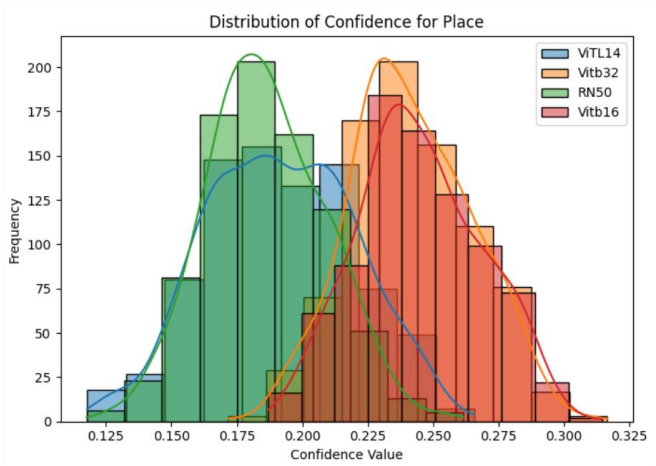
- **Perilaku (*Behavior)***



Gambar 15. Distribusi confidence behavior

Distribusi confidence pada behavior menunjukkan pola serupa. ViT-B/16 dan RN50 memiliki confidence tertinggi sekitar 0.25–0.30, menandakan performa stabil. ViT-L/14 memiliki distribusi lebih lebar di 0.2–0.25, sementara ViT-B/32 kembali memiliki confidence terendah, mencerminkan prediksi yang kurang yakin.

- **Tempat *(Place)***

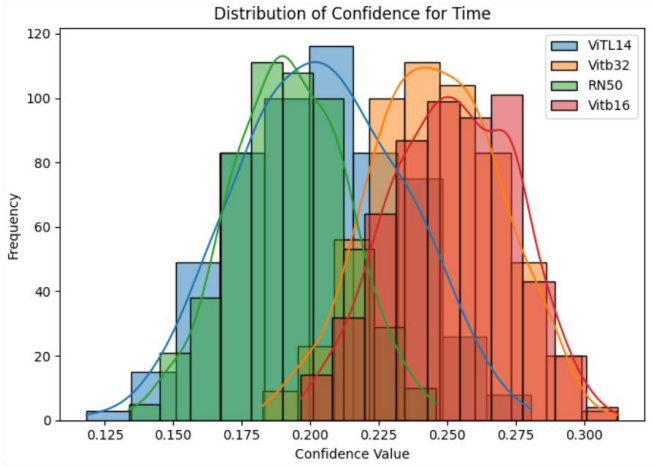


Gambar 16. Distribusi confidence place

Pada kategori place, ViT-B/16 dan RN50 tetap dominan dengan confidence tertinggi di sekitar 0.25–0.30. ViT-L/14 lebih menyebar di 0.175–0.225, menunjukkan variasi yang lebih besar. ViT-B/32 memiliki puncak di 0.2–0.225, menunjukkan performa yang masih lebih rendah dibanding model lainnya.

Copyright © 2022 Journal of Science and Applicative Technology
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Waktu *(Time)*



Gambar 17. Distribusi confidence time

Pada kategori time, ViT-B/16 dan RN50 memiliki confidence tertinggi di 0.25–0.30, menunjukkan keyakinan yang konsisten. ViT-L/14 memiliki distribusi lebih lebar di 0.2–0.25, sementara ViT-B/32 memiliki confidence terendah di 0.2–0.225, menunjukkan performa yang lebih lemah.

Analisis Perbandingan ke-4 Model

- **Prediksi**

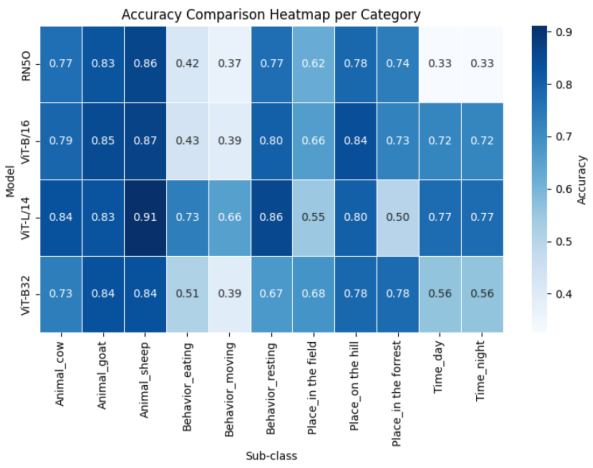
Hasil prediksi yang didapatkan adalah berupa gambar dan teks yang sesuai. Untuk menghasilkan akurasi yang tepat sesuai label *ground truth,* diperlukan model yang baik. Dari 279 gambar yang diprediksi dan 54 kombinasi teks yang berbeda, berikut adalah perbandingan hasil prediksi dari keempat model

Tabel 7. Hasil prediksi 100% tiap model

No	Model	Hasil Prediksi 100%	Teks Bergambar	Teks Tidak Bergambar
1	RN50	31	24	30
2	Vit-B/32	34	25	29
3	ViT-L/14	70	32	22
4	ViT-B/16	32	23	31

Model ViT-L/14 menunjukkan performa terbaik dengan 70 prediksi 100% benar, jauh lebih unggul dibandingkan model lainnya. Model berbasis Vision Transformer (ViT) dengan ukuran lebih besar lebih efektif dalam memahami hubungan gambar-teks yang kompleks dibandingkan model kecil seperti RN50.

- **Akurasi Sub-kelas**

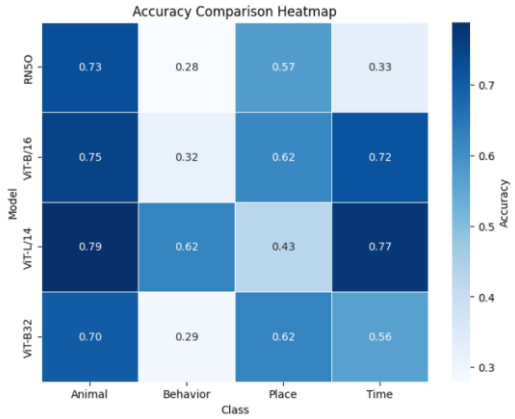


Gambar 18. Perbandingan akurasi *sub-class* model CLIP

Dari tabel di atas, dapat dilihat bahwa setiap model kesulitan untuk membedakan *cow* dan *goat*, kemungkinan disebabkan oleh karakteristik visual seperti warna, bentuk, dan postur yang serupa. Hal ini terlihat dari akurasi yang cenderung lebih rendah pada kedua kelas tersebut dibandingkan dengan *sheep*, yang memiliki ciri khas lebih jelas.

ViT-B/16 menunjukan hasil yang konsisten, terutama dalam class *sheep* (0,94), *in the forest* (0,89) dan *day/night (*0,94). Model ViT-B/16 dan ViT-L/14 memiliki tingkat akurasi yang hampir sama dalam sebagian besar class, namun ViT-L/14 lebih unggul dalam beberapa class seperti *Bahavior_resting* (0,89) dan *day/night*(0,92). RN50 cenderung memiliki hasil yang lebih rendah dibandingkan dengan model ViT, khususnya pada class *eating*(0,35) dan *moving* (0,33).

- **Akurasi Kelas Utama**

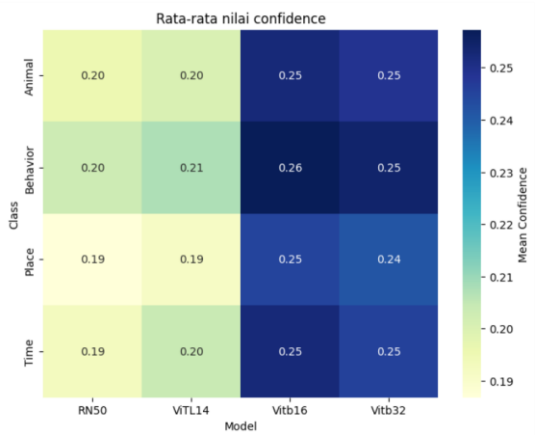


Gambar 19. Perbandingan akurasi *class* model CLIP

ViT-L/14 unggul di kategori *animal* (0.79), *behavior* (0.62), dan *time* (0.77), meskipun rendah pada *place* (0.43). ViT-B/16 menunjukkan performa paling stabil dengan akurasi baik di semua kategori, khususnya *place* (0.62) dan *time* (0.72). ViT-B/32 menonjol di kategori *place* (0.62) tetapi lemah pada *behavior*(0.29). Sementara itu, RN50 hanya unggul di *animal* (0.73) dengan performa rendah di kategori lainnya.

Secara keseluruhan model ViT-L/14 unggul dalam beberapa kategori. Model ini memiliki kapasitas yang besar dan mampu menangkap detail yang lebih halus dalam data visual karena menggunakan ukuran patch yang lebih kecil (14x14 piksel) sehingga memungkinkan ViT-L/14 untuk menganalisis dengan resolusi lebih tinggi. Penelitian sebelumnya dalam pengembangan CLIP menunjukkan bahwa kombinasi pretraining berbasis teks dan gambar dengan model CLIP L/14 unggul memiliki akurasi (59.2%) dibandingkan model lain RN50 (56.4%). [3]

- **Perbandingan Nilai Confidence**



Gambar 20. Perbandingan nilai *confidence class* model CLIP

Confidence model menunjukkan bahwa ViT-B/16 dan ViT-B/32 memiliki tingkat kepercayaan yang tinggi dan stabil pada berbagai kategori, terutama *Place* dan *Time*. Namun, jika dilihat dari akurasi, ViT-L/14 justru unggul pada kategori *Animal* (0.79), *Behavior* (0.62), dan *Time* (0.77). Meskipun *confidence*-nya moderat, model ini mampu menangkap detail lebih halus berkat kapasitas besar dan ukuran patch yang kecil (14x14). Sebaliknya, RN50 menunjukkan confidence rendah secara keseluruhan, kecuali pada *Animal*, di mana akurasinya mencapai 0.73.

Nilai *confidence* mencerminkan seberapa yakin model terhadap prediksinya, tetapi tidak selalu sejalan dengan akurasi. ViT-L/14 lebih akurat meskipun *confidence*-nya tidak selalu paling tinggi, sedangkan ViT-B/16 menunjukkan performa paling stabil di semua kategori.

Kesimpulan

Penelitian ini mengevaluasi empat model CLIP (RN50, ViT-B/16, ViT-L/14, dan ViT-B/32) dalam klasifikasi gambar-teks spesies, perilaku, lokasi, dan waktu ternak menggunakan pendekatan zero-shot learning. RN50 memiliki akurasi terbaik pada kategori spesies (animal) dengan nilai tertinggi 0.86 tetapi lemah dalam memprediksi perilaku dan waktu. ViT-B/32 unggul pada kategori lokasi (place) dengan akurasi 0.68–0.78 namun masih rendah pada kategori perilaku. ViT-L/14 menunjukkan performa tertinggi secara keseluruhan dengan akurasi mencapai 0.91 pada kategori spesies dan 0.73–0.86 pada perilaku, meskipun akurasi lokasi lebih rendah. Sementara itu, ViT-B/16 menunjukkan performa paling seimbang di semua kategori dengan akurasi tinggi pada spesies (0.79–0.87), lokasi (0.66–0.84), dan waktu (0.72), menjadikannya model paling stabil untuk klasifikasi ini.

Untuk meningkatkan akurasi model, penelitian selanjutnya dapat mempertimbangkan fine-tuning model CLIP menggunakan dataset spesifik terkait ternak, yang terbukti efektif dalam mengatasi kelemahan pada kategori tertentu seperti perilaku dan lokasi. Selain itu, memperluas dataset dengan jumlah gambar yang lebih besar dan kondisi visual yang bervariasi akan membantu model dalam meningkatkan kemampuan generalisasi. Eksplorasi metode tambahan seperti *few-shot learning* atau model ensemble juga dapat dilakukan untuk mengombinasikan keunggulan dari setiap model yang dievaluasi.

Konflik Kepentingan

Tidak terdapat konflik kepentingan dalam penelitian dan penulisan.

Daftar Pustaka

[1] V. Vidit, M. Engilberge, and M. Salzmann, "CLIP the Gap: A single domain generalization approach for object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 3219–3229.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G.

Copyright © 2022 Journal of Science and Applicative Technology
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, PMLR, Jul. 2021, pp. 8748–8763.

[3] M. Li, R. Xu, S. Wang, L. Zhou, X. Lin, C. Zhu, M. Zeng, H. Ji, and S.-F. Chang, "CLIP-Event: Connecting text and images with event structures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 16420–16429.

[4] H. Wang, Y. Li, H. Yao, and X. Li, "CLIPN for zero-shot OOD detection: Teaching CLIP to say no," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2023, pp. 1802–1812.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] G. Dussert, V. Miele, C. van Reeth, A. Delestrade, S. Dray, and S. Chamaillé-Jammes, "Zero-shot animal behavior classification with vision-language foundation models," *Working paper or preprint*, Nov. 2024

[7] R. D. Indrapurasih, M. A. Bijaksana, and I. L. Sardi, "Implementasi dan Analisis Kesamaan Semantik Antar Kata Bahasa Indonesia Menggunakan Metode GloVe," *e-Proceeding of Engineering*, vol. 5, no. 3, pp. 7699-7706, Dec. 2018.

[8] J. Sanjaya and M. Ayub, "Augmentasi Data Pengenalan Citra Mobil Menggunakan Pendekatan Random Crop, Rotate, dan Mixup," *JuTISI*, vol. 6, no. 2, Aug. 2020.

[9] OpenAI, "CLIP: Connecting text and images," *OpenAI*, Jan. 5, 2021. [Online]. Available: https://openai.com/index/clip/.

[10] GeeksforGeeks, "CLIP: Contrastive Language-Image Pretraining," *GeeksforGeeks*, [Online]. Available: https://www.geeksforgeeks.org/clip-contrastive-language-image-pretraining/.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] X. Chen, C.-J. Hsieh, and B. Gong, "When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations," *arXiv preprint*, arXiv:2106.01548 [cs.CV], 2021.

[13] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *ACM Computing Surveys*, vol. 54, no. 10s, Jan. 2022.

[14] G. Boesch, "Vision Transformers (ViT) in Image Recognition: Full Guide," *viso.ai*, Nov. 25, 2023. [Online]. Available: https://viso.ai/deep-learning/vision-transformer-vit/.

[15] OpenAI, "CLIP ViT-B/16 Model," *Hugging Face*, 2021. [Online]. Available: https://huggingface.co/openai/clip-vit-base-patch16..