Nama : Ketut Satria Wibisana
NIM : 1103213148
Kelas : TK-45-G04

Laporan Simulasi pada Computer Vision

1. Pendahuluan

Simulasi ini bertujuan untuk menerapkan konsep-konsep computer vision dengan memanfaatkan model-model multimodal berbasis Vision-Language seperti CLIP, BLIP, DePlot, dan lain-lain. Melalui eksperimen ini, berbagai aplikasi dalam domain seperti Visual Question Answering (VQA), Image Captioning, dan Transfer Learning diterapkan menggunakan data gambar dan teks. Hasilnya diharapkan dapat memperluas pemahaman dalam penggunaan model berbasis vision dan language untuk berbagai kasus aplikasi dalam kehidupan nyata.

2. Tujuan Simulasi

- Mengimplementasikan VQA (Visual Question Answering): Menerapkan teknik untuk menjawab pertanyaan berdasarkan gambar yang diberikan.
- Image Captioning: Menghasilkan deskripsi teks dari gambar yang diberikan.
- Penerapan Transfer Learning dalam Vision-Language Models: Menggunakan model pre-trained dan melakukan fine-tuning untuk aplikasi spesifik seperti VQA dan Captioning**.**

3. Metode Simulasi

3.1 Persiapan Lingkungan Kerja

Perangkat dan Software yang Digunakan:

- Python (versi 3.10)
- Libraries: transformers, torch, PIL, requests
- Model yang digunakan: Salesforce/blip-vqa-capfilt-large, dandelin/vilt-b32-finetuned-vqa, openai/clip-vit-base-patch32, dan lainnya.

3.2 Visual Question Answering (VQA)

Pada tahap pertama, dilakukan penerapan model BLIP-VQA untuk menjawab pertanyaan berdasarkan gambar.

- **Langkah Implementasi:**
    1. Mengimpor model dan processor yang diperlukan menggunakan library transformers.

2. Memuat gambar yang akan dianalisis menggunakan PIL.
3. Menyiapkan pertanyaan yang relevan dengan gambar, seperti "Is there an elephant?".
4. Menggunakan pipeline visual-question-answering untuk memperoleh jawaban berdasarkan gambar dan pertanyaan yang diberikan.

Kode simulasi:

```
from PIL import Image
from transformers import pipeline
vqa_pipeline = pipeline(
   "visual-question-answering", model="Salesforce/blip-vqa-capfilt-large"
)
image = Image.open("elephant.jpg")
question = "Is there an elephant?"
result = vqa_pipeline(image, question, top_k=1)
print(result)  # Output: [{'answer': 'yes'}]
```

3.3 Image Captioning

Pada tahap kedua, digunakan model untuk menghasilkan deskripsi gambar.

- **Langkah Implementasi:**
  1. **Menggunakan model Salesforce/blip-image-captioning-large untuk melakukan captioning pada gambar.**
  2. **Gambar dimuat menggunakan URL atau dari file lokal, lalu diproses untuk menghasilkan deskripsi.**

Kode Simulasi:

```
from transformers import BlipProcessor, BlipForConditionalGeneration
import requests
from PIL import Image
processor = BlipProcessor.from_pretrained("Salesforce/blip-image-captioning-large")
model = BlipForConditionalGeneration.from_pretrained("Salesforce/blip-image-captioning-large")
```

```
img_url = "https://storage.googleapis.com/sfr-vision-language-research/BLIP/demo.jpg"
raw_image = Image.open(requests.get(img_url, stream=True).raw).convert("RGB")
inputs = processor(raw_image, return_tensors="pt")
out = model.generate(**inputs)
caption = processor.decode(out[0], skip_special_tokens=True)
print(caption)  # Output: 'a photography of a woman and her dog on the beach'
```

3.4 Transfer Learning pada Vision-Language Models

Pada tahap ini, dilakukan eksperimen transfer learning dengan model microsoft/git-base-coco untuk penerapan captioning pada gambar yang diunggah.

- **Langkah Implementasi:**
    1. **Menggunakan model dan processor git-base-coco untuk memproses gambar dan menghasilkan caption.**
    2. **Gambar dimuat menggunakan URL dan caption dihasilkan berdasarkan gambar tersebut.**

Kode Simulasi:

```
from transformers import AutoProcessor, AutoModelForCausalLM
import requests
from PIL import Image
processor = AutoProcessor.from_pretrained("microsoft/git-base-coco")
model = AutoModelForCausalLM.from_pretrained("microsoft/git-base-coco")
url = "http://images.cocodataset.org/val2017/000000039769.jpg"
image = Image.open(requests.get(url, stream=True).raw)
pixel_values = processor(images=image, return_tensors="pt").pixel_values
generated_ids = model.generate(pixel_values=pixel_values, max_length=50)
generated_caption = processor.batch_decode(generated_ids,
skip_special_tokens=True)[0]
print(generated_caption)  # Output: 'two cats sleeping on a pink blanket next to remotes.'
```

4. Hasil Simulasi

1. **Visual Question Answering:**
    - Pertanyaan: "Is there an elephant?"
    - Jawaban: "yes" (dari model BLIP-VQA)
2. **Image Captioning:**
    - Deskripsi Gambar: "A photography of a woman and her dog on the beach."
3. **Transfer Learning:**
    - Gambar yang diunggah menghasilkan caption: "Two cats sleeping on a pink blanket next to remotes."

5. Analisis dan Diskusi

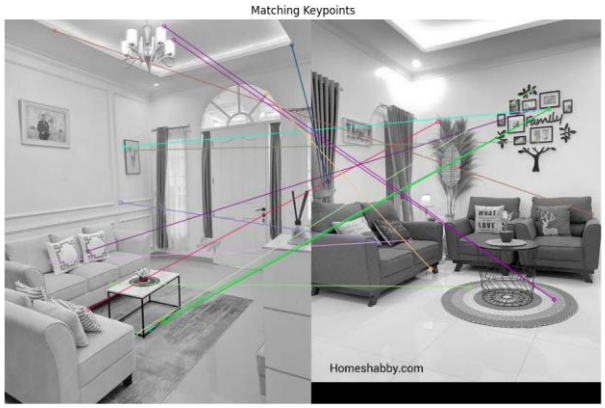
- **Kekuatan Model:**
    - Model-model seperti BLIP dan CLIP menunjukkan kemampuan yang sangat baik dalam memahami hubungan antara gambar dan teks.
    - VQA (Visual Question Answering) dapat memberikan jawaban yang relevan dan akurat sesuai dengan gambar yang diberikan.
    - Image Captioning dapat menghasilkan deskripsi yang cukup mendetail tentang gambar yang dianalisis.
- **Keterbatasan:**
    - Model-model ini mungkin masih terbatas pada konteks atau data yang dilatih, yang dapat mempengaruhi akurasi pada kasus yang lebih kompleks atau gambar yang tidak umum.
    - Dalam beberapa kasus, jawaban yang dihasilkan bisa sangat sederhana dan tidak memberikan konteks yang lebih dalam.

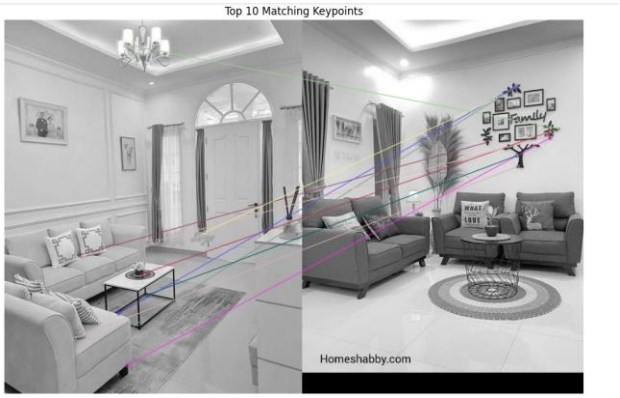
6. Kesimpulan

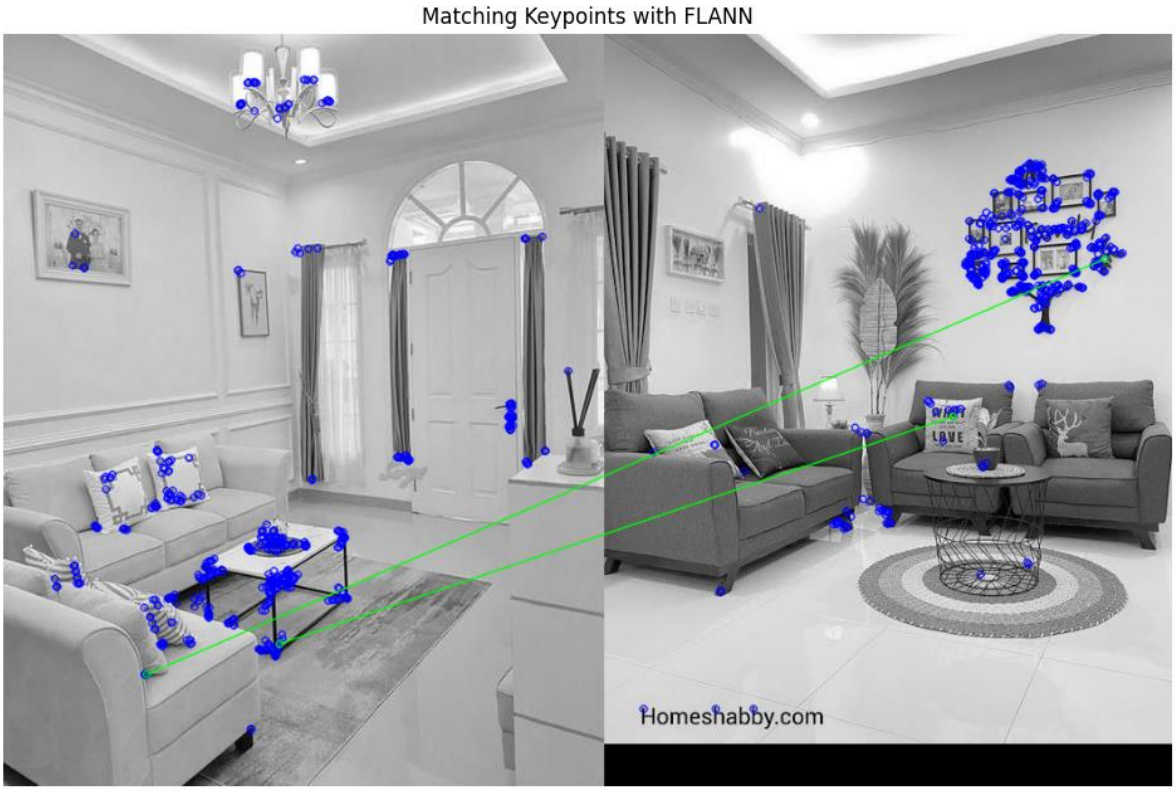
Simulasi ini berhasil menunjukkan bagaimana teknologi multimodal dapat digunakan untuk memahami gambar dan teks secara bersamaan dalam berbagai aplikasi. Penggunaan model seperti BLIP, CLIP, dan DePlot memperlihatkan potensi besar dalam memahami hubungan antara teks dan gambar di berbagai tugas seperti VQA dan Image Captioning. Selain itu, transfer learning terbukti sangat berguna dalam meningkatkan kemampuan model untuk menangani tugas-tugas spesifik.

Video Link YT: https://youtu.be/Yb3wDnt3Tak

Dokumentasi Output Pengerjaan:



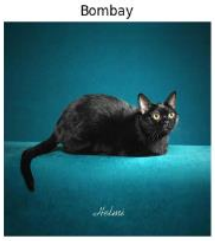





0
200
400
600
800
1000


newfoundland


Russian Blue


Bombay


Egyptian Mau


basset hound


scottish terrier


scottish terrier


keeshond


Birman


Persian


havanese


Maine Coon


miniature pinscher


yorkshire terrier


yorkshire terrier

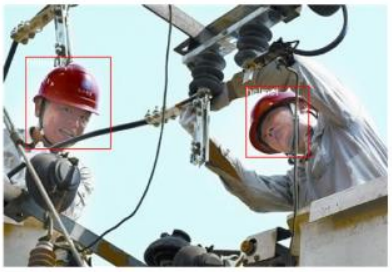

label: Maine Coon
predicted: Maine Coon
label: english setter
predicted: english setter
label: Egyptian Mau
predicted: Egyptian Mau
label: german shorthaired
predicted: german shorthaired
label: samoyed
predicted: samoyed
label: Bengal
predicted: Bengal
label: english cocker spaniel
predicted: english cocker spaniel
label: Egyptian Mau
predicted: Egyptian Mau
label: Bombay
predicted: Bombay
label: Bombay
predicted: Bombay
label: basset hound
predicted: basset hound
label: american bulldog
predicted: american bulldog
label: Bombay
predicted: Bombay
label: english setter
predicted: english setter
label: havanese
predicted: havanese


江苏省苏中建设集团股份有限公司
承建 上海青浦汇金路1#地块 工程
环保安康建时
诚信守法筑苏
helmet
head


湖北新闻网

新华网江西频道
www.jx.xinhuanet.com

东莞时间网






helmet
股东在线