Jaringan Saraf Konvolusional (CNN), Transformer Visi (ViT), dan Model Multimodal: Pendekatan Terpadu dalam Penglihatan Komputer

Chapter 2-4

Abstrak

Penglihatan komputer telah mengalami perkembangan signifikan dengan diperkenalkannya Jaringan Saraf Konvolusional (CNN) dan Transformer Visi (ViT). Selain itu, model multimodal yang mengintegrasikan berbagai jenis data semakin menjadi fokus penelitian. Artikel ini membahas konsep dasar, perkembangan terkini, dan aplikasi dari CNN, ViT, dan model multimodal dalam penglihatan komputer.

1. Pendahuluan

Penglihatan komputer bertujuan untuk memungkinkan mesin memahami dan menafsirkan data visual. Jaringan Saraf Konvolusional (CNN) telah menjadi arsitektur dominan dalam tugas-tugas penglihatan komputer, seperti klasifikasi citra dan deteksi objek. Namun, dengan munculnya Transformer Visi (ViT), pendekatan baru yang memanfaatkan mekanisme perhatian (attention) telah diperkenalkan, menawarkan kemampuan untuk menangkap dependensi jarak jauh dalam data visual. Selain itu, model multimodal yang menggabungkan informasi dari berbagai sumber data, seperti teks dan gambar, telah menunjukkan potensi besar dalam meningkatkan pemahaman konteksual.

2. Jaringan Saraf Konvolusional (CNN)

2.1 Definisi dan Arsitektur

CNN adalah jenis jaringan saraf tiruan yang dirancang untuk memproses data yang memiliki grid pattern, seperti citra [1]. Arsitektur CNN terdiri dari lapisan konvolusi, lapisan pooling, dan lapisan fully connected. Lapisan konvolusi berfungsi untuk mengekstraksi fitur lokal dari citra, sementara lapisan pooling mengurangi dimensi data dan kompleksitas komputasi [1].

2.2 Aplikasi CNN

CNN telah berhasil diterapkan dalam berbagai tugas penglihatan komputer, termasuk klasifikasi citra, deteksi objek, dan segmentasi citra [1]. Keberhasilan CNN dalam tugas-tugas ini disebabkan oleh kemampuannya dalam menangkap fitur hierarkis dari data visual [1].

3. Transformer Visi (ViT)

3.1 Pendahuluan

Transformer Visi (ViT) adalah adaptasi dari arsitektur Transformer yang awalnya dikembangkan untuk pemrosesan bahasa alami, diterapkan pada data citra [2]. ViT membagi citra menjadi patch-patch kecil dan memprosesnya sebagai urutan, memungkinkan model untuk menangkap hubungan global dalam citra [2].

3.2 Keunggulan ViT

ViT memiliki kemampuan untuk menangkap dependensi jarak jauh dalam data visual, yang sulit dicapai oleh CNN [2]. Selain itu, ViT dapat dilatih pada data skala besar dan menunjukkan kinerja yang kompetitif dengan arsitektur CNN tradisional dalam berbagai tugas penglihatan komputer [2].

4. Model Multimodal

4.1 Definisi

Model multimodal adalah model yang dirancang untuk memproses dan mengintegrasikan informasi dari berbagai modalitas data, seperti teks, gambar, dan audio [3]. Pendekatan ini memungkinkan model untuk memahami konteks yang lebih kaya dan membuat prediksi yang lebih akurat [3].

4.2 Aplikasi Model Multimodal

Model multimodal telah diterapkan dalam berbagai aplikasi, termasuk pencarian gambar berbasis teks, analisis sentimen dengan mempertimbangkan ekspresi wajah, dan sistem tanya jawab yang menggabungkan informasi visual dan tekstual [3].

5. Tantangan dan Peluang

5.1 Tantangan

Meskipun memiliki potensi besar, penerapan CNN, ViT, dan model multimodal menghadapi beberapa tantangan, seperti kebutuhan akan data pelatihan dalam jumlah besar, kompleksitas komputasi yang tinggi, dan integrasi informasi dari berbagai modalitas yang berbeda [4].

5.2 Peluang

Dengan kemajuan dalam teknologi komputasi dan ketersediaan data yang semakin meningkat, terdapat peluang besar untuk mengembangkan model yang lebih efisien dan efektif [4]. Selain itu, integrasi model-model ini dalam aplikasi dunia nyata, seperti kendaraan otonom dan sistem kesehatan cerdas, dapat membawa dampak positif yang signifikan [4].

6. Kesimpulan

Jaringan Saraf Konvolusional, Transformer Visi, dan model multimodal merupakan komponen kunci dalam pengembangan sistem penglihatan komputer modern. Dengan memahami konsep dan aplikasi dari masing-masing model, serta tantangan yang dihadapi, peneliti dan praktisi dapat mengembangkan solusi yang lebih baik untuk berbagai permasalahan dalam domain ini.

Daftar Pustaka

[1] Dosovitskiy, A., et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021. [Online]. Available: https://arxiv.org/abs/2010.11929.

[2] Liu, Z., et al., "A ConvNet for the 2020s," *CVPR*, 2022. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Liu_A_ConvNet_for_the_2020s_CVPR_2022_paper.html.

[3] Radford, A., et al., "Learning Transferable Visual Models From Natural Language Supervision," *ICML*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020.

[4] Akbari, H., et al., "VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text," *NeurIPS*, 2021. [Online]. Available: https://arxiv.org/abs/2104.11178.

[5] Carion, N., et al., "End-to-End Object Detection With Transformers," *ECCV*, 2020. [Online]. Available: https://arxiv.org/abs/2005.12872.