ABSTRAK

Vision-language model biasanya dilatih ke data berupa pasangan citra dan teksnya. Model ini dapat digunakan untuk image-text retrieval. Model akan mengembalikan citra yang sesuai dengan query teks dan sebaliknya. Model ini biasanya dilatih pada dataset bahasa Inggris. Kebanyakan penelitian menggunakan encoder multimodal untuk teks berbahasa Inggris, tetapi terbatas pada bahasa lainnya. Penelitian ini bertujuan untuk melatih model berbasis Transformer ke dataset berbahasa Indonesia yang besar. Penulis melatih ke dataset yang sangat besar agar dapat digunakan untuk downstream task atau untuk transfer learning pelatihan model pada dataset yang lebih sedikit. Dataset diambil dari COCO dan Flickr yang diterjemah. Penerjemahan dilakukan dengan Google Translate yang serupa dengan data image captioning pada NusaCrowd. Model yang digunakan, yaitu CLIP, ALIGN, dan BLIP yang merupakan model dengan performa baik. Penulis juga membangun model seperti CLIP berbasis DeIT dan BeIT sebagai image encoder dan IndoBERT serta ModernBERT sebagai text encoder. Citra diubah ukurannya menjadi ukuran tetap dan diaugmentasi secara acak. Teks dibersihkan dari tanda baca dan diubah menjadi huruf kecil. Model berhasil meraih performa yang baik pada data test. Model terbaik, yaitu BeIT-IndoBERT meraih R@10 pada COCO test set sebesar 72,20 untuk image-to-text dan 65,21 untuk text-to-image retrieval. Penelitian ini dapat dikembangkan dengan menggunakan dataset yang dianotasi oleh manusia. Data yang diambil dari situs web berbahasa Indonesia juga dapat digunakan di penelitian selanjutnya. Model-model yang dibangun dapat digunakan untuk task downstream, yaitu image captioning dan visual question answering.

Kata Kunci: Image-Text Retrieval, Multimodal, Transformer, Vision-Language, Retrieval