Multimodal Transformer for Indonesian Image Text Retrieval

Muhammad Naufal Hawari
School of Computing
Telkom University
Bandung, Indonesia
naufalhawari@student.telkomuniversity.ac.id

Gamma Kosala
School of Computing
Telkom University
Bandung, Indonesia
gammakosala@telkomuniversity.ac.id

Rifki Wijaya
Center of Excellence Artificial Intelligence for Learning and Optimization
Telkom University
Bandung, Indonesia
rifkiwijaya@telkomuniversity.ac.id

***Abstract*—Vision-language model massively trained on image-text pairs. The model can be utilized for image-text retrieval. It retrieves most similar image given a text query, and vice versa. The model trained on abundance of English image captions. Most research pretrained multimodal encoder for English caption dataset but limited to other language. This research aimed to train Transformer-based model to large Indonesian image-text pairs for image-text retrieval. We trained to large dataset, so it can further be utilized for downstream tasks or transfer the learned parameter to a smaller dataset. Dataset were acquired from translated COCO and Flickr dataset. The translation was done by Google Translate platform as in NusaCrowd image captioning catalogue. The employed pretrained models in this research were CLIP, ALIGN, and BLIP as the top multimodal model benchmark. We also constructed CLIP-like architecture utilizing ViT, DeIT, and BeIT as image encoder with IndoBERT and ModernBERT as text encoder. Images were resized and augmented randomly. Texts were cleaned from punctuation and lowercased. Models performed well and achieved good retrieval performance on the test set. The highest R@10 on COCO test set was by BeIT-IndoBERT model with 72.20% R@10 for image-to-text and 65.21% R@10 for text-to-image retrieval. This research can be improved by employing on more refined dataset with human-annotated image captions by Indonesian annotators. Dataset scraped from Indonesian online documents can also be leveraged in future research. These models can be further trained on downstream tasks, including image captioning and visual question answering.**

Keywords—Image-Text Retrieval, Multimodal, Transformer, Vision-Language, Retrieval

I. INTRODUCTION

Multimodal retrieval model has evolved to retrieve context from various modal. One type of multimodal retrieval is utilizing image and text, known as image-text retrieval. The trained model retrieves the most similar text given an image query, and vice versa. Its applications are useful for image search engine. It was also employed for retrieving information to Large Language Model (LLM) context.

Image-text retrieval has been well-developed by researchers on English language as high resource language. A. Li et al. proposed Visual N-Gram [1]. Qi et al. proposed ImageBERT [2]. G. Li et al. developed Unicoder-VL [3]. Gan et al. proposed VILLA [4]. Chen and Li developed UNITER [5]. John et al. proposed GPO [6]. X. Li et al. proposed Oscar [7]. Yu et al. proposed ERNIE-ViL [8]. Radford et al. developed CLIP [9]. Jia et al. proposed ALIGN [10]. Li et al. proposed BLIP [11]. These models were pretrained to large English dataset.

However, image-text retrieval has not developed for low resource language. One of them was Indonesian language (Bahasa). There was not study on Indonesian image-text retrieval. The only study on large image-text dataset was Indonesian image captioning. An Indonesian image captioning study by Mahadi et al. utilizing RNN with Attention model on Google translated COCO and Flickr30k. In total, the dataset were more than 500.000 image-text pairs in Indonesian [12].

In this research, we trained Transformer-based models on large Indonesian image-text pairs dataset for Indonesian image-text retrieval task. Transformer-based models were employed because it can extract feature in parallel. The models were trained to learn the similarity between a text and an image with contrastive loss. The similarity score can be leveraged for image-text retrieval. The dataset were acquired from translated Microsoft COCO and Flickr30k. Dataset were translated by Google Translate as it was utilized by Mahadi et al. [12] and this translation method was also employed as one of dataset catalogue in Indonesian dataset collection, NusaCrowd [13].

II. LITERATURE REVIEW

Image-text retrieval has been well-developed by researchers on English language as high resource language. The dataset for training were pairs of image-text. The dataset were collected with refined captions by annotators like in COCO and Flickr30k dataset. It also can be acquired by collecting images from web pages and extracts surrounding text or alt html tag of the web pages like in Google Conceptual Captions [14] or ImageBERT data collection [2].

Image-text model mostly has two main building blocks. It included image encoder and text encoder. The encoder mostly constructed based on Convolutional Neural Network (CNN). Recurrent Neural Network (RNN), and Transformer. Some of them also performed nondeep learning model. A. Li et al. proposed Visual N-Gram [1]. They used N-Gram embedding for the text encoder and CNN as image encoder. Visual N-Gram was trained with naive N-Gram loss. Qi et al. proposed ImageBERT with BERT (Bidirectional Encoder Representation from Transformer) as text encoder and Faster-RCNN as image encoder [2]. The Faster-RCNN extracted features from several regions. G. Li et al. developed Unicoder-VL [3]. This research performed masked scene graph knowledge modeling. Mask were applied to text and fed to graph parser. Gan et al. proposed VILLA [4]. VILLA was trained by adding adversarial perturbation on both image and text extracteed features. Chen and Li developed UNITER [5]. They performed RCNN as image encoder and Transformer as text encoder. It uniquely performed word region alignment training objective. John et al. proposed GPO [6]. GPO