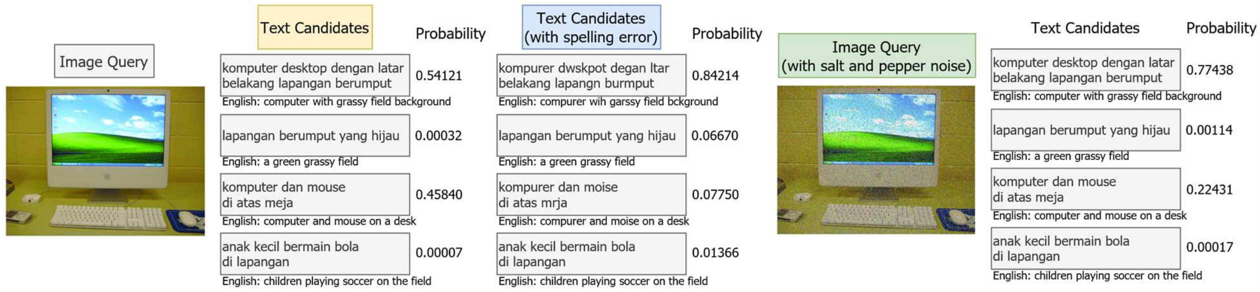

Fig. 7. Image-Text Retrieval Inference Examples

Fig. 7 shows inference examples of image-text retrieval with our model. We observed the model can return higher probability to relevant text candidates under normal condition. Noise in image affect changes of text probability but the model still ranked relevant texts higher. It shows spelling error affected quite high changes in probability. In spelling error example, the model gave similar probability to second and third text. Although the second text got higher probability than the third, the gap between second and third text probability should be higher in spelling error example.

Fig. 8 text-image retrieval examples with our model. We tried to query an image from a text query given three images. We witnessed the model return highest probability to the relevant image given a text query. The model accurately gave higher probability to second image among the three images. We also observed images with noise did not change much the image candidates probability. Spelling error on the text query affected much lower probability changes compared to image-text retrieval inference example.

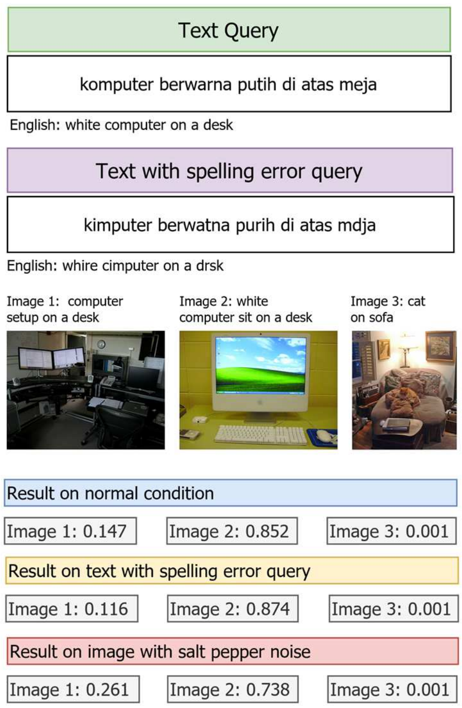

Fig. 8. Text-Image Retrieval Inference Examples

V. Conclusion

This research is the first research to developed image-text retrieval in Indonesian. We employed large dataset and multimodal Transformer with contrastive loss objective. The best model achieved more than 65% of Recall@10 both image to text and text to image task on COCO and Flickr30k test set. We also observed the model has potential as initial model for transfer learning to smaller dataset.

English pretrained model mostly got trained on millions of image. Our research employed around 500,000 images. We should have more dataset to improve our model. So, it can achieve metrics similar to English pretrained models in future research. There were more image-text pairs dataset in NusaCrowd. They were English dataset from web pages translated to Indonesian by machine translation system. Even though it was not as refined as COCO or Flickr30k set, these dataset can be considered to be employed in the future. It would be much better to has refined dataset. The refined dataset can be acquired either with images carefully annotated by Indonesian locals or scraped online documents with careful filtering.

References

[1] A. Li, A. Jabri, A. Joulin, and L. van der Maaten, "Learning Visual N-Grams From Web Data," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017.

[2] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data," *CoRR*, vol. abs/2001.07966, 2020, [Online]. Available: https://arxiv.org/abs/2001.07966

[3] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 11336–11344, Apr. 2020, doi: 10.1609/aaai.v34i07.6795.

[4] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, "Large-Scale Adversarial Training for Vision-and-Language Representation Learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Curran Associates, Inc., 2020, pp. 6616–6628. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/49562478de4c54fafd4ec46fdb297de5-Paper.pdf

[5] L. and Y. L. and E. K. A. and A. F. and G. Z. and C. Y. and L. J. Chen Yen-Chun and Li, "UNITER: UNiversal Image-TExt Representation Learning," in *Computer Vision – ECCV 2020*, H. and B. T. and F. J.-M. Vedaldi Andrea and Bischof, Ed., Cham: Springer International Publishing, 2020, pp. 104–120.

[6] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the Best Pooling Strategy for Visual Semantic Embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 15789–15798.

[7] X. Li *et al.*, "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks," 2020, pp. 121–137. doi: 10.1007/978-3-030-58577-8_8.

[8] F. Yu *et al.*, "ERNIE-ViL: Knowledge Enhanced Vision-Language Representations through Scene Graphs," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, pp. 3208–3216, May 2021, doi: 10.1609/aaai.v35i4.16431.

[9] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, Mar.

2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[10] C. Jia *et al.*, “Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, Jun. 2021, pp. 4904–4916. [Online]. Available: https://proceedings.mlr.press/v139/jia21b.html

[11] J. Li, D. Li, C. Xiong, and S. Hoi, “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation,” in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., in Proceedings of Machine Learning Research, vol. 162. PMLR, Nov. 2022, pp. 12888–12900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[12] M. R. S. Mahadi, A. Arifianto, and K. N. Ramadhani, “Adaptive Attention Generation for Indonesian Image Captioning,” in *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 2020, pp. 1–6. doi: 10.1109/ICoICT49345.2020.9166244.

[13] S. Cahyawijaya *et al.*, “NusaCrowd: Open Source Initiative for Indonesian NLP Resources,” 2023. [Online]. Available: https://arxiv.org/abs/2212.09648

[14] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, “Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3557–3567. doi: 10.1109/CVPR46437.2021.00356.

[15] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., Cham: Springer International Publishing, 2014, pp. 740–755.

[16] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models,” *Int J Comput Vis*, vol. 123, no. 1, pp. 74–93, 2017, doi: 10.1007/s11263-016-0965-7.

[17] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ArXiv*, vol. abs/2010.11929, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:225039882

[18] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, “Training data-efficient image transformers & distillation through attention,” in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., in Proceedings of Machine Learning Research, vol. 139. PMLR, Jun. 2021, pp. 10347–10357. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a.html

[19] H. Bao, L. Dong, and F. Wei, “BEiT: BERT Pre-Training of Image Transformers,” *CoRR*, vol. abs/2106.08254, 2021, [Online]. Available: https://arxiv.org/abs/2106.08254

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[21] B. Wilie *et al.*, “IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding,” in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020.

[22] B. Warner *et al.*, “Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference,” 2024. [Online]. Available: https://arxiv.org/abs/2412.13663