IMPLEMENTASI *VIDEO CAPTIONING* MENGGUNAKAN *OBJECT RELATIONAL GRAPH* DENGAN PENDEKATAN *NON-AUTOREGRESSIVE*

SKRIPSI

Diajukan untuk Memenuhi Sebagian dari
Syarat Memperoleh Gelar Sarjana Komputer
Program Studi Ilmu Komputer

Oleh
Muhammad Ilham Malik
1902563

PROGRAM STUDI ILMU KOMPUTER
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA
BANDUNG
2022

IMPLEMENTASI *VIDEO CAPTIONING* MENGGUNAKAN *OBJECT RELATIONAL GRAPH* DENGAN PENDEKATAN NON-AUTOREGRESSIVE

Oleh

Muhammad Ilham Malik

NIM 1902563

Diajukan untuk memenuhi salah satu syarat memperoleh gelar Sarjana Komputer
pada Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

Universitas Pendidikan Indonesia

Agustus 2023

MUHAMMAD ILHAM MALIK

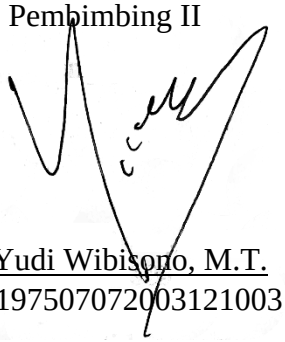
IMPLEMENTASI *VIDEO CAPTIONING* MENGGUNAKAN *OBJECT RELATIONAL GRAPH* DENGAN PENDEKATAN NON-AUTOREGRESSIVE

disetujui dan disahkan oleh pembimbing:

Pembimbing I

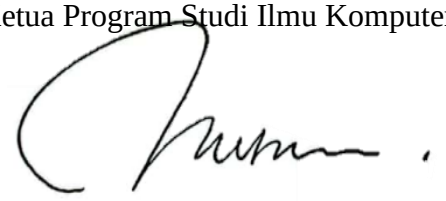
Yaya Wihardi, M.Kom
NIP: 198903252015041001

Pembimbing II



Dr. Yudi Wibisono, M.T.
NIP: 197507072003121003

Mengetahui
Ketua Program Studi Ilmu Komputer



Dr. Muhamad Nursalman, M.T.
NIP. 197909292006041002

SURAT PERNYATAAN

Dengan ini saya menyatakan bahwa skripsi/tesis/disertasi dengan judul "Implementasi *Video Captioning* Menggunakan *Object Relational Graph* dengan Pendekatan *Non-Autoregressive*" ini beserta seluruh isinya adalah benar-benar karya saya sendiri. Saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika ilmu yang berlaku dalam masyarakat keilmuan. Atas pernyataan ini, saya siap menanggung risiko/sanksi apabila di kemudian hari ditemukan adanya pelanggaran etika keilmuan atau ada klaim dari pihak lain terhadap keaslian karya saya ini.

Bandung, Juli 2023

Yang membuat pernyataan

Muhammad Ilham Malik

1902563

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada kehadirat Allah SWT. karena atas kehendak-Nya serta karunia-Nya penyusunan skripsi yang berjudul "Implementasi *Video Captioning* Menggunakan *Object Relational Graph* dengan Pendekatan *Non-Autoregressive*" ini dapat diselesaikan tepat waktu. Penyusunan skripsi ini ditujukan untuk memenuhi dan melengkapi salah satu syarat untuk memperoleh gelar sarjana komputer atas jenjang S1 pada Program Studi Ilmu Komputer Universitas Pendidikan Indonesia.

Penulis menyadari bahwa selama penyusunan skripsi ini tidak terlepas dari peran, dukungan, dan bantuan dari berbagai pihak baik secara langsung atau tidak langsung. Oleh karena itu, penulis sampaikan terima kasih yang sebesar-besarnya kepada:

1. Kedua orang tua yang senantiasa berdoa demi kelancaran dalam penyusunan skripsi ini serta dukungan secara moral ataupun materi.
2. Bapak Yaya Wihardi, M.Kom., dan Bapak Dr. Yudi Wibisono, M.T., sebagai pembimbing, atas segala ilmu, motivasi dan waktu yang telah dicurahkan kepada penulis serta sumber daya komputer yang telah dipinjamkan kepada penulis.
3. Ibu Rosa Ariani Sukamto, M.T., atas segala ilmu dasar pemrograman hingga pemrograman tingkat lanjut, motivasi, dan serta kesediannya untuk berdiskusi.
4. Ibu Dr. Rani Megasari, M.Kom., dan Bapak Dr. Muhammad Nursalman, M.T., dan segenap dosen Prodi Ilmu Komputer atas ilmu yang sangat bermanfaat yang telah ditanamkan kepada penulis.
5. Rekan-rekan mahasiswa Prodi Ilmu Komputer 2019 yang telah bersama berjuang menimba ilmu.
6. Semua pihak yang telah membantu penulis dalam menyelesaikan skripsi ini yang tidak bisa disebutkan satu-persatu.

Akhirnya, penulis menyampaikan permohonan maaf atas segala ketidaksempurnaan. Penulis sangat mengharapkan setiap saran dan kritik yang

membangun agar skripsi ini dapat bermanfaat bagi kemajuan ilmu pengetahuan dan teknologi.

Bandung, Juli 2023

Penulis

IMPLEMENTASI *VIDEO CAPTIONING* MENGGUNAKAN *OBJECT RELATIONAL GRAPH* DENGAN PENDEKATAN *NON-AUTOREGRESSIVE*

Oleh

Muhammad Ilham Malik – muh.ilham.malik@upi.edu

1902563

ABSTRAK

Kemampuan video captioning dalam menghasilkan deskripsi singkat yang menjelaskan isi dari video secara detail dengan waktu inference yang rendah merupakan hal yang penting. Namun, metode-metode yang ada saat ini memiliki kekurangan pada kedua aspek tersebut. Pada penelitian ini, kami mengajukan sebuah model video captioning yang bernama *Object Relational Graph* dengan pendekatan *Non-autoregressive Coarse-to-Fine* (ORG-NACF) untuk mengatasi masalah video captioning pada kedua aspek tersebut. Modul ORG digunakan untuk memperoleh informasi objek secara detail dan mempelajari hubungan antar objek. Modul NACF bersama sequential cross attention digunakan untuk menyelesaikan masalah waktu *inference* yang tinggi dan menjaga kualitas caption ketika pembangkitan *caption*. Evaluasi dengan Dataset MSR-VTT menunjukan hasil unjuk kerja Model ORG-NACF yang setara bahkan melebihi dari model *state-of-the-ar*t pada beberapa metrik serta memiliki kelebihan lain yaitu waktu *inference* pembangkitan *caption* yang lebih rendah. Hasil ini menunjukan bahwa Model ORG-NACF mampu membangkitkan *caption* yang deskriptif dengan waktu *inference* yang lebih rendah dibandingkan dengan metode yang ada saat ini.

Kata Kunci: *Video Captioning, Transformer, Object Detection, Graph Convolutional Network, Convolutional Neural Network.*

IMPLEMENTASI *VIDEO CAPTIONING* MENGGUNAKAN *OBJECT RELATIONAL GRAPH* DENGAN PENDEKATAN *NON-AUTOREGRESSIVE*

Arranged by

Muhammad Ilham Malik – muh.ilham.malik@upi.edu

1902563

ABSTRACT

The ability of video captioning to generate a detailed caption that explains the content of the video with low inference is important. However, existing methods have limitations in both aspects. In this paper, we propose a video captioning model Object Relational Graph with Non-autoregressive Coarse to Fine (ORG-NACF) approach to tackle the video captioning problem in both aspects. The ORG module is used to obtain detailed object information and learn the relationship between the objects. The NACF module along with sequential cross attention is used to solve the problem of high inference time and maintain caption quality during caption generation. Experimental evaluation on benchmark MSR-VTT dataset shows that the performance of the ORG-NACF model is competitive and even exceeds the state-of-the-art model on several metrics and has the advantage of faster inference time. This model achieved 7 times more faster inference time than the baseline model. These results show that the ORG-NACF Model is able to generate descriptive and detailed captions with lower inference time compared to existing methods.

Kata Kunci: *Video Captioning, Transformer, Object Detection, Graph Convolutional Network, Convolutional Neural Network.*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR PUSTAKA

Aafaq, N., Akhtar, N., Liu, W., Gilani, S. Z., & Mian, A. (2019). Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 12479–12488. https://doi.org/10.1109/CVPR.2019.01277

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *ArXiv Preprint ArXiv:1409.0473*.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.

Chen, D., & Dolan, W. B. (2011). Collecting highly parallel data for paraphrase evaluation. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 190–200.

Chen, S., & Jiang, Y.-G. (2021). Motion guided region message passing for video captioning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1543–1552.

Chen, S., & Jiang, Y.-G. (2019). Motion guided spatial attention for video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 8191–8198.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., & Batra, D. (2017). Visual dialog. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 326–335.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Gella, S., Lewis, M., & Rohrbach, M. (2018). A dataset for telling the stories of social media videos. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 968–974.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."

Ghazvininejad, M., Levy, O., Liu, Y., & Zettlemoyer, L. (2019). Mask-predict: Parallel decoding of conditional masked language models. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 6112–6121. https://doi.org/10.18653/v1/d19-1633

Girshick, R. (2015). Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision*, 1440–1448.

Hara, K., Kataoka, H., & Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6546–6555.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Hori, C., Hori, T., Lee, T. Y., Zhang, Z., Harsham, B., Hershey, J. R., Marks, T. K., & Sumi, K. (2017). Attention-Based Multimodal Fusion for Video Description. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 4203–4212. https://doi.org/10.1109/ICCV.2017.450

Hubel, D. H. (1959). Single unit activity in striate cortex of unrestrained cats. *The Journal of Physiology*, *147*(2), 226.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International Conference on*

Machine Learning, 448–456.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 74–81.

Liu, A. A., Xu, N., Wong, Y., Li, J., Su, Y. T., & Kankanhalli, M. (2017). Hierarchical & multimodal video captioning: Discovering and transferring multimodal knowledge for vision to language. *Computer Vision and Image Understanding*, *163*, 113–125. https://doi.org/10.1016/j.cviu.2017.04.013

Liu, F., Ren, X., Wu, X., Yang, B., Ge, S., & Sun, X. (2021). O2NA: An Object-Oriented Non-Autoregressive Approach for Controllable Video Captioning. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 281–292. https://doi.org/10.18653/v1/2021.findings-acl.24

Namjoshi, M., & Khurana, K. (2021). A mask-rcnn based object detection and captioning framework for industrial videos. *International Journal of Advanced Technology and Engineering Exploration*, *8*(84), 1466–1478. https://doi.org/10.19101/IJATEE.2021.874394

Nguyen, V.-Q., Suganuma, M., & Okatani, T. (2022). Grit: Faster and better image captioning transformer using dual visual features. *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, 167–184.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *ArXiv Preprint ArXiv:1511.08458*.

Pan, Y., Mei, T., Yao, T., Li, H., & Rui, Y. (2016). Jointly modeling embedding and translation to bridge video and language. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 4594–4602. https://doi.org/10.1109/CVPR.2016.497

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual*

Meeting of the Association for Computational Linguistics, 311–318.

Pei, W., Zhang, J., Wang, X., Ke, L., Shen, X., & Tai, Y.-W. (2019). Memory-attended recurrent network for video captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8347–8356.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520.

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3202–3212.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv Preprint ArXiv:1910.01108*.

Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural Comput*, *9*(8), 1735–1780.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Torabi, A., Pal, C., Larochelle, H., & Courville, A. (2015). Using descriptive video services to create a large data source for video annotation research. *ArXiv Preprint ArXiv:1503.01070*.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 4489–4497. https://doi.org/10.1109/ICCV.2015.510

Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural language processing with transformers*. " O'Reilly Media, Inc."

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4566–4575.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2015). Translating videos to natural language using deep recurrent neural networks. *NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 1494–1504. https://doi.org/10.3115/v1/n15-1173

Voykinska, V., Azenkot, S., Wu, S., & Leshed, G. (2016). How blind people interact with visual content on social networking services. *Proceedings of the 19th Acm Conference on Computer-Supported Cooperative Work & Social Computing*, 1584–1595.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., & Wang, W. Y. (2019). Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4581–4591.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual

transformations for deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1492–1500.

Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5288–5296.

Xu, J., Yao, T., Zhang, Y., & Mei, T. (2017). Learning multimodal attention LSTM networks for video captioning. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, 537–545. https://doi.org/10.1145/3123266.3123448

Yang, B., Zou, Y., Liu, F., & Zhang, C. (2021). Non-autoregressive coarse-to-fine video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(4), 3119–3127.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing videos by exploiting temporal structure. *Proceedings of the IEEE International Conference on Computer Vision*, 4507–4515.

Zhang, J., & Peng, Y. (2019). Object-aware aggregation with bidirectional temporal graph for video captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8327–8336.

Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., & Zha, Z. (2020). Object relational graph with teacher-recommended learning for video captioning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 13275–13285. https://doi.org/10.1109/CVPR42600.2020.01329