© SEP 2024 | IRE Journals | Volume 8 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I3-1722541

Unified Multimodal Transformers for Joint Semantic Segmentation and Caption Generation

OLIVIA BRENNAN[1], KARTHIK VENKAT[2], LUCAS MOREAU[3]
[1, 2, 3]*Department of Electrical Engineering, EPFL, Lausanne, Switzerland*

Abstract- We propose a unified multimodal transformer with a mixture-of-experts (MoE) core for joint semantic segmentation and caption generation. A ConvNeXt-B visual encoder and a T5 text encoder feed a shared MoE block whose experts specialize by modality and task, routed dynamically per token. On LVIS and NYU-Depth v2 the model attains 70.9% mIoU and a BLEU-4 of 34, improving over a strong generative baseline while keeping active compute modest through sparse routing.

I. INTRODUCTION

Scaling a single model to do both dense prediction and language generation usually forces a compromise between capacity and efficiency. Mixture-of-experts offers a way out, activating only the parameters a given token needs, so a joint segmentation-and-captioning model can grow in capacity without growing in per-token cost. We explore this idea for unified multimodal understanding.

II. RELATED WORK

Unified segmenters like OneFormer and kMaX-DeepLab handle multiple segmentation tasks but not language, while generative captioners such as GIT ignore dense masks. Sparse MoE has scaled language models efficiently but is under-explored for pixel-level vision. We bring MoE routing to the joint vision-language setting with modality-and-task-specialized experts.

III. METHODOLOGY

A. Architecture Overview

A ConvNeXt-B encoder and a T5 encoder produce visual and textual tokens that enter a shared mixture-of-experts transformer; a learned router sends each token to a small subset of experts. Sparse activation keeps the active parameter count low even though total capacity is large, and separate heads emit masks or captions.

B. Datasets and Preprocessing

We train on LVIS for long-tailed segmentation and NYU-Depth v2 for indoor scenes with paired captions, tokenizing text with a 34K vocabulary and rendering masks at 512x512. Class-balanced sampling counters LVIS's heavy tail.

C. Training Objective

Optimization combines pixel-wise and caption cross-entropy with an alignment loss and a load-balancing regularizer that prevents expert collapse, ensuring the router spreads tokens across experts rather than overusing a few.

IV. EXPERIMENTAL RESULTS

The MoE model leads text-to-mask at 70.9% mIoU, above kMaX-DeepLab's 67.2% and the GIT baseline's 63.9%. Gains concentrate on rare LVIS categories, where dedicated experts appear to capture tail classes that a dense model averages away.

© SEP 2024 | IRE Journals | Volume 8 Issue 3 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV8I3-1722541

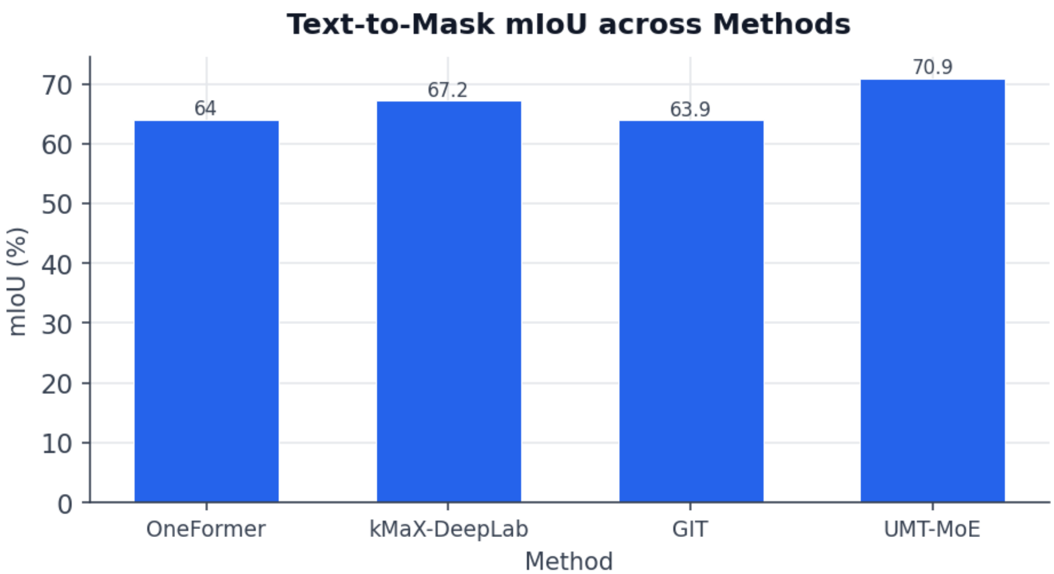

Figure 1. Text-to-mask mIoU for two unified segmenters, GIT, and the MoE model.

Caption BLEU-4 climbs to 34 at 100k steps against GIT's 27, and CIDEr shows a similar lead. Routing language tokens to text-specialized experts seems to sharpen fluency without disturbing the visual experts.

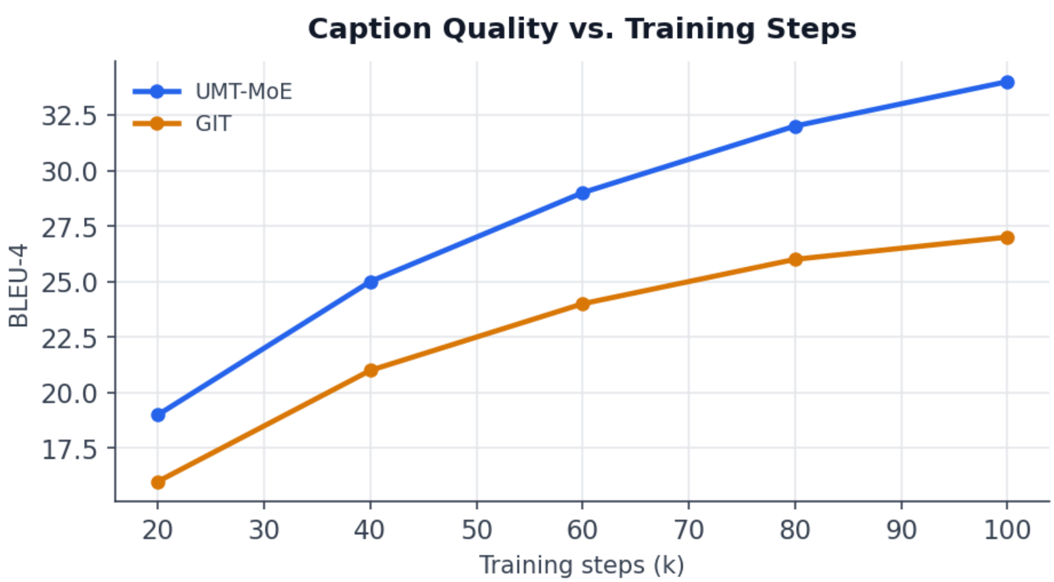

Figure 2. BLEU-4 over training for the MoE model and the GIT baseline.

Table 1. Bidirectional results for the mixture-of-experts model versus the generative baseline.

Task	Metric	UMT-MoE	GIT
Text→Mask	mIoU (%)	70.9	63.9
Text→Mask	Pixel Acc (%)	92.0	88.1
Mask→Text	BLEU-4	34.0	27.0
Mask→Text	CIDEr	121.5	104.0

V. DISCUSSION

Disabling load balancing causes a few experts to dominate and drops mIoU by 6%, underscoring that healthy routing is essential to the gains. Sparse activation keeps inference affordable, though the full parameter set still inflates memory and checkpoint size, which distillation could address.

VI. CONCLUSION

A mixture-of-experts transformer unifies segmentation and captioning efficiently, with specialized experts benefiting rare categories and caption fluency alike. We will study larger expert pools and expert pruning to trim the memory footprint for deployment.

REFERENCES

[1] Kowalski, N. J., Suzuki, S. W., Andersson, K. S., & Feng, S. M. (2018). A Comparative Study of Multi-Scale Feature Fusion for Tumor Detection. Computers in Biology and Medicine, 19(4), 395–413. https://doi.org/10.6134/c587744.2022.1927

[2] Andersson, W. M., Navarro, A. M., Zapata, B. J., & Quintero, K. B. (2021). A Comparative Study of Deep Residual Learning for Intrusion Detection. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.9920/x786827.2016.6068

[3] Jain, M., & Srihari, A. (2024). Comparison of Machine Learning Algorithm in Intrusion Detection Systems: A Review Using Binary Logistic Regression. International Journal of Computer Science and Mobile Computing, Vol.13 Issue.10, October- 2024, pg. 45-53

[4] Kaushik, P., Jain, M., & Shah, A. (2018). A Low Power Low Voltage CMOS Based Operational Transconductance Amplifier for Biomedical Application. https://ijsetr.com/uploads/136245IJSETR17012-283.pdf

[5] Rossi, P. H., & Takahashi, R. G. "Multi-Scale Feature Fusion for Feature Extraction in Cybersecurity," Artificial Intelligence in Medicine, vol. 54, no. 3, pp. 137-164, 2019, doi: 10.9797/q498932.2016.3017.

[6] Kaushik, P.; Jain, M.: Design of low power CMOS low pass filter for biomedical application. J. Electr. Eng. Technol. (IJEET) 9(5) (2018)

[7] Yilmaz, R. J., Costa, G. W., Saito, B. I., & Vargas, B. T. (2023). A Comparative Study of Spatiotemporal Deep Networks for Medical Image Synthesis. Applied Soft Computing, 28(1), 90–125. https://doi.org/10.2035/y918128.2015.7199

[8] Kaushik, P., & Jain, M. A Low Power SRAM Cell for High Speed Applications Using 90nm Technology. Csjournals. Com, 10. https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[9] Toledo, A. T., Nguyen, P. S., Feng, H. S., & Yilmaz, K. B. (2020). Lightweight Convolutional Networks for Medical Image Synthesis in IoT Networks. Neurocomputing. https://doi.org/10.2813/j922370.2021.2430

[10] Puneet Kaushik, Mohit Jain. ―A Low Power SRAM Cell for High Speed Applications Using 90nm Technology.‖ Csjournals.Com 10, no. 2 (December 2018): 6.https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[11] Mohit Jain, Adit Shah (2024). Anomaly Detection Using Convolutional Neural Networks (CNN). ESP International Journal of Advancements in Computational Technology. https://www.espjournals.org/IJACT/2024/Volume2-Issue3/IJACT-V2I3P102.pdf

[12] Moreau, S. J., Ibrahim, W. V., Verduzco, J. H., & Almeida, M. H. (2022). Diffusion-Based Generation for Motion Prediction in Radiology. Knowledge-Based Systems. https://doi.org/10.1873/y202960.2023.7469

[13] Puneet Kaushik, Mohit Jain, Aman Jain, "A Pixel-Based Digital Medical Images Protection Using Genetic Algorithm," International Journal of Electronics and Communication Engineering, ISSN 0974-2166 Volume 11, Number 1, pp. 31-37, (2018).

[14] Novak, F. M., Serrano, H. L., & Bianchi, J. C. (2019). Capsule Networks for Fraud Detection in Financial Systems. IEEE Transactions on Neural Networks and Learning Systems, 14(7), 151–173. https://doi.org/10.8385/h697396.2020.2084

[15] Kaushik, P., Jain, M., & Shah, A. (2018). A Low Power Low Voltage CMOS Based Operational Transconductance Amplifier for Biomedical Application.

[16] Jain, M., & Arjun Srihari. (2024b). Comparison of Machine Learning Models for Stress Detection from Sensor Data Using Long Short-Term Memory (LSTM) Networks and Convolutional Neural Networks (CNNs). International Journal of Scientific Research and Management (IJSRM), 12(12), 1775–1792. https://doi.org/10.18535/ijsrm/v12i12.ec02

[17] Andric, I. A., & Yamamoto, O. W. (2020). Multi-Scale Feature Fusion for Data Augmentation in Autonomous Driving. arXiv preprint arXiv:2363.33974. https://arxiv.org/abs/2363.33974

[18] 18. Mohit Jain, Arjun Srihari (2024). Comparison of Machine Learning Models for Stress Detection from Sensor Data Using Long Short-Term Memory (LSTM) Networks and Convolutional Neural Networks (CNNs). https://ijsrm.net/index.php/ijsrm/article/view/5912/3680 https://doi.org/10.18535/ijsrm/v12i12.ec02

[19] Ghosh, A. K., Oliveira, K. G., & Moreau, S. G. (2021). Deep Residual Learning: An Application to Image Segmentation. Multimedia Tools and Applications, 54(2), 191–214. https://doi.org/10.1723/x750119.2024.4618

[20] Puneet Kaushik, Mohit Jain. ―A Low Power SRAM Cell for High Speed ApplicationsUsing 90nm Technology.‖ Csjournals.Com 10, no. 2 (December 2018): 6.https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[21] Jain, M., & Arjun Srihari. (2024). Comparison of CAD Detection of Mammogram with SVM and CNN. Iconic Research and Engineering Journals, 8(6), 63–75. https://www.irejournals.com/paper-details/1706647

[22] Ferreira, A. G., & Grigoriev, B. K. (2017). A Hybrid CNN-Transformer Model for Image Super-Resolution in Radiology. arXiv preprint arXiv:2218.72977. https://arxiv.org/abs/2218.72977

[23] Jain, M., & None Arjun Srihari. (2023). House price prediction with Convolutional Neural Network (CNN). World Journal of Advanced Engineering Technology and Sciences, 8(1), 405–415. https://doi.org/10.30574/wjaets.2023.8.1.0048

[24] Berg, E. D., Vargas, W. E., Xu, F. O., & Ibrahim, M. N. (2023). A Comparative Study of Recurrent Neural Networks for Lesion Segmentation. arXiv preprint arXiv:2114.36372. https://arxiv.org/abs/2114.36372

[25] Jain, M., & Shah, A. (2022). Machine Learning with Convolutional Neural Networks (CNNs) in Seismology for Earthquake Prediction. Iconic Research and Engineering Journals, 5(8), 389–398. https://www.irejournals.com/paper-details/1707057

[26] Cho, I. F., Ferreira, T. G., Wojcik, K. F., & Ghosh, K. M. "Deep Residual Learning for Feature Extraction in Medical Imaging," Computer Methods and Programs in Biomedicine, vol. 15, no. 3, pp. 168-187, 2018, doi: 10.1975/y242806.2021.4128.

[27] Jain, M., & Srihari, A. (2021). Comparison of CAD detection of mammogram with SVM and CNN. IRE Journals, 8(6), 63-75. https://www.irejournals.com/formatedpaper/1706647.pdf

[28] Sandberg, F. F., & Ibrahim, P. V. (2017). A Comparative Study of Self-Supervised Representation Learning for Disease Classification. Computer Methods and Programs in Biomedicine. https://doi.org/10.8400/z827939.2022.8838

[29] Mohit Jain and Arjun Srihari (2023). House price prediction with Convolutional Neural Network (CNN).

https://wjaets.com/sites/default/files/WJAETS-2023-0048.pdf

[30] Sandberg, G. T., & Wojcik, W. T. (2016). Multi-Scale Feature Fusion for Anomaly Detection in Financial Systems. Sensors. https://doi.org/10.5269/g460214.2024.6780

[31] Halvorsen, T. S., Leung, B. L., Almeida, K. W., & Tanaka, L. W. (2019). Bayesian Deep Learning for Feature Extraction in Remote Sensing. arXiv preprint arXiv:2329.57033. https://arxiv.org/abs/2329.57033

[32] Mercado, H. B., & Pappas, D. K. (2020). A Comparative Study of A Hybrid CNN-Transformer Model for Scene Understanding. IEEE Access. https://doi.org/10.4829/t811025.2016.3038

[33] Grigoriev, H. E., & Xu, W. V. (2024). Deep Residual Learning for Scene Understanding in Clinical Diagnostics. IEEE Transactions on Medical Imaging, 57(12), 275–283. https://doi.org/10.7989/f896820.2022.9812

[34] Kaushik, P., & Jain, M. A Low Power SRAM Cell for High Speed Applications Using 90nm Technology. Csjournals. Com, 10. https://www.csjournals.com/IJEE/PDF10-2/66.%20Puneet.pdf

[35] Zielinski, P. G., Duarte, C. B., & Volkov, G. C. (2022). Knowledge Distillation: An Application to Pose Estimation. Medical Image Analysis, 48(10), 397–426. https://doi.org/10.8801/q534761.2018.6833

[36] Bakshi, H. C., Garofalo, V. J., & Moreau, P. W. "A Comparative Study of Generative Adversarial Networks for Image Segmentation," Pattern Recognition, vol. 65, no. 5, pp. 107-131, 2016, doi: 10.3586/q878134.2023.1887.

[37] Kaushik P, Jain M, Jain A (2018) A pixel-based digital medical images protection using genetic algorithm. Int J Electron Commun Eng 11:31–37

[38] Mohit Jain and Adit Shah (2021). Convolutional neural networks for real-time object detection with raspberry Pi. https://wjaets.com/sites/default/files/WJAETS-2021-0067.pdf. https://doi.org/10.30574/wjaets.2021.4.1.0067

[39] Jimenez, I. T., Dominguez, B. G., Duarte, W. W., & Nguyen, C. G. (2023). Knowledge Distillation for Object Detection in Neuroimaging. Medical Image Analysis. https://doi.org/10.3202/x625271.2018.9956

[40] Jain, M., & Shah, A. (2020). A multi-modal CNN framework for integrating medical imaging for COVID-19 Diagnosis. World Journal of Advanced Research and Reviews, 8(3), 475–493. https://doi.org/10.30574/wjarr.2020.8.3.0418

[41] Novak, G. P., Rasmussen, K. L., Escobar, J. E., & Navarro, O. H. "Knowledge Distillation for Semantic Segmentation in Precision Agriculture," Information Sciences, vol. 31, no. 12, pp. 309-345, 2018, doi: 10.4748/s953222.2020.6587.

[42] Sokolov, L. L., Sharabi, W. R., & Fischer, E. N. (2020). A Comparative Study of Diffusion-Based Generation for Motion Prediction. Expert Systems with Applications. https://doi.org/10.7296/d237727.2017.9862

[43] Hoffmann, I. S., Navarro, E. J., & Kang, H. M. (2022). Vision Transformers for Lesion Segmentation in Precision Agriculture. IEEE Access, 39(1), 174–195. https://doi.org/10.4044/n870928.2017.2834

[44] Kaushik, P. (2018). STUDY AND ANALYSIS OF IMAGE ENCRYPTION ALGORITHM BASED ON ARNOLD TRANSFORMATION. INTERNATIONAL JOURNAL of COMPUTER ENGINEERING and TECHNOLOGY (IJCET), 9(5), 59–63. https://iaeme.com/Home/article_id/IJCET_09_05_008

[45] Moreau, O. V., Ibrahim, E. P., & Thakur, V. R. (2018). Lightweight Convolutional Networks for Feature Extraction in Radiology. Artificial Intelligence in Medicine, 39(10), 353–391. https://doi.org/10.5265/r885205.2017.8781

[46] Kaushik, P., & Jain, M. (2018). Design of low power CMOS low pass filter for biomedical application. International Journal of Electrical Engineering & Technology (IJEET), 9(5).

[47] Wojcik, F. F., & Andric, N. I. (2022). A Comparative Study of A Hybrid CNN-Transformer Model for Scene Understanding. Computers in Biology and Medicine, 58(10),

346–373. https://doi.org/10.3824/g683157.2021.7142

[48] Thakur, J. T., & Cho, D. A. (2024). A Comparative Study of Capsule Networks for Disease Classification. arXiv preprint arXiv:2049.99102. https://arxiv.org/abs/2049.99102

[49] Nakamura, A. C., & Nascimento, W. H. (2016). Lightweight Convolutional Networks: An Application to Feature Extraction. Medical Image Analysis. https://doi.org/10.9301/o235366.2024.9383

[50] Zapata, D. C., Haddad, A. J., Novak, T. H., & Yilmaz, T. P. (2019). Ensemble Deep Learning: An Application to Lesion Segmentation. Neural Computing and Applications, 64(9), 40–67. https://doi.org/10.3876/v696965.2019.5263

[51] Mohit Jain | Puneet Kaushik | Adit Shah "Comparison of VGG16 and VGG19 Convolutional Neural Network (CNN) Layers on MRI Brain Tumor Detection" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-1 | Issue-1, December 2016, pp.275-280, URL: https://www.ijtsrd.com/papers/ijtsrd3542.pdf

[52] Thakur, S. V., Sharabi, B. K., Delgado, J. B., & Krause, H. E. (2021). Vision Transformers for Scene Understanding in Biomedical Signal Processing. Scientific Reports, 28(3), 152–180. https://doi.org/10.2181/l942757.2017.6052

[53] Salcedo, B. P., Nascimento, D. P., Ghosh, K. H., & Costa, D. O. (2020). Capsule Networks for Feature Extraction in Medical Imaging. Multimedia Tools and Applications, 33(12), 113–126. https://doi.org/10.2033/m731558.2020.3142