Liminal efficiency for AI general models: investigating compression techniques for maximized performance in the glaucoma's diagnosis and prognosis*

Volodymyr Kysil[1,†], Olga Drachuk[2,*,†], Anatolii Korol[2,*,†], Valentyna Hnenna[2,*,†] and Elena Zaitseva[3,†]

[1] *Khmelnytskyi National University, Institutska str., 11, Khmelnytskyi, 29016, Ukraine*
[2] *National Pirogov Memorial Medical University, Pirogova str., 56, Vinnytsya, 21018, Ukraine*
[3] *Zilina University, Univerzitná 8215, 010 26 Žilina, Slovakia*

Abstract

Deep learning models have demonstrated remarkable performance across the glaucoma's diagnosis and prognosis; however, deploying them in resource-constrained environments poses significant challenges. This research explores the balance between compression and accuracy preservation in specialist convolutional neural networks (CNNs) intended for CPU-based execution with minimal storage requirements. By employing pruning, knowledge distillation, quantization, and weight sharing, it is aimed to achieve maximal compression without compromising essential task performance. Resulting findings provide insights into the efficiency limits of model compression and its implications for real-world deployment. Additionally, the applicability of these compression techniques to Transformer-based architectures is examined throughout the work, which pose unique challenges due to their reliance on attention mechanisms.

Keywords

deep learning, glaucoma's diagnosis and prognosis, medical imaging optic nerve segmentation, fundus image analysis, model compression, pruning, knowledge distillation, quantization, weight sharing, transformer models, convolutional neural networks (CNN), low-resource deployment, edge computing, self-attention mechanisms, transfer learning, efficiency, optimization, neural network performance, sparse models, inference speed, binary classification.

1. Introduction

Glaucoma is the second most common cause of blindness and disability in the world [1]. It affects more than 90 million people worldwide [1]. It is estimated that in 2013, the number of people aged 40-80 years suffering from glaucoma was 64.3 million, in 2020 - about 76.0 million, and by 2040 this figure is projected to increase to 111.8 million [2].

The main difficulty in diagnosing glaucoma is that its symptoms usually appear only when vision has already deteriorated significantly. Due to the asymptomatic course and slow development of the disease, many people are unaware of the problem, which makes it difficult to detect and predict it early, especially in the initial stages. However, timely diagnosis can slow down the progression of the disease and prevent vision loss.

Since glaucoma causes retinal damage due to damage to the optic nerve head and increased intraocular pressure, an important step in its detection is the segmentation of the optic disc. This procedure is difficult due to the small size of the disc and possible blood supply disorders [3].

Intelitsis'25: The 6th International Workshop on Intelligent Information Technologies & Systems of Information Security, April 04, 2025, Khmelnytskyi, Ukraine
* Corresponding author.
† These authors contributed equally.
vovikusspambox@gmail.com (V. Kysil); drachuk@vnmu.edu.ua (O. Drachuk); histology@vnmu.edu.ua (A. Korol); valentina.gnenna@gmail.com (V. Hnenna); elena.zaitseva@fri.uniza.sk (E. Zaitseva)
0009-0003-9387-6609 (V. Kysil); 0000-0002-0504-4059 (O. Drachuk); 0000-0001-5773-4222 (A. Korol); 0000-0002-0058-8399 (V. Hnenna); 0000-0002-9087-0311 (E. Zaitseva)

The standard diagnosis of glaucoma includes an assessment of the optic nerve head by examining the retina, measuring intraocular pressure, analyzing visual fields, and taking into account other related factors [4]. At the same time, traditional methods are time-consuming, largely dependent on the ophthalmologist's experience, and may be accompanied by human error, which makes it difficult to detect the disease in the early stages.

A promising area for improving the diagnostic process is modern information consulting and diagnostic technologies. They help to improve medical decisions and the quality of treatment of patients with retinal pathologies and are already used in numerous systems for diagnosing various diseases [5-8].

Over the past decade, machine learning and deep learning methods have developed significantly, not only improving the accuracy of glaucoma detection but also speeding up the processing of large amounts of images and data, making the work of doctors easier [9]. Studies confirm that the use of these technologies allows for the automatic diagnosis of glaucoma in the early stages by analyzing fundus and optical coherence tomography images to classify the eye as normal or glaucoma-affected [10, 11].

Recent advancements in deep learning have led to increasingly large and complex models, making deployment on limited-resource environments such as edge devices and low-power CPUs impractical. While specialist models tailored for specific tasks offer high accuracy, they often retain redundant parameters that inflate model size and computational overhead. The challenge lies in optimizing these models for real-world applications while maintaining their effectiveness.

Compression techniques [12-14] such as pruning (removes unnecessary weights, reducing computation and memory footprint), quantization (converts high-precision weights into lower-bit representations, minimizing storage requirements and enhancing inference speed; applied to an already-learned model for increasing performance speed), weight sharing (groups similar weights to reduce parameter redundancy while preserving network expressiveness; also applied to an already-learned model for increasing performance speed and decrease storage space), and knowledge distillation (transfers knowledge from a larger, more complex model to a smaller, more efficient model) can significantly reduce model size and inference time. However, an aggressive compression strategy may result in degraded accuracy, raising the question: How far can a model be compressed while maintaining usability?

2. Related works

The balance between model size and performance has been widely studied. Han et al. [15] introduced structured and unstructured pruning to eliminate redundant weights, demonstrating notable efficiency improvements. Frankle & Carbin [16] proposed the Lottery Ticket Hypothesis, suggesting that sparse subnetworks can match full model performance. Other works [17-19] explore quantization and architectural optimizations to further enhance efficiency.

Additionally, research by Bala and Eschbach [20] highlights the importance of optimized color spaces for improving neural network visibility in medical imaging. Their findings inform our preprocessing techniques for fundus image analysis.

Studies such as Wu et al. [18] emphasize the delicate balance between compression level and accuracy. Excessive pruning or quantization can introduce errors that degrade performance, highlighting the need for a strategic approach to compression. Knowledge distillation has also been explored as a means to retain essential knowledge in highly compressed models while mitigating accuracy loss.

Research has shown that specialized models tend to outperform general-purpose models in their respective domains [21-24]. This motivated selection of an image transformer model for optic nerve extraction, as transformers have demonstrated superior performance in medical imaging tasks. By leveraging their self-attention mechanisms, these models excel in capturing fine-grained spatial dependencies, which is crucial for precise segmentation of the optic nerve in fundus images.

3. Methodology

3.1. Baseline model description

CNN-based encoder-decoder architecture was utilized, structured as follows:

1. Initial feature extraction – VGG-16 backbone with four pooling layers.
2. Intermediate processing – custom layers for high-level feature abstraction.
3. Decoding – three decoder layers with skip connections to preserve spatial information.

This model is fine-tuned for binary classification of optic nerve localization in fundus images. Model is used for image segmentation into single-channel optic mask [25]. The model is tasked to learn high-level features, which, in practice, may further decompose into finer feature representations depending on the model's depth and granularity. Feature rough estimate is 17 discrete low-level features that the CNN must detect, process, and classify to correctly locate the optic nerve. Feature breakdown follows:

Feature 1: Bright round/oval spot (optic disc):
1. Brightness intensity contrast (1).
2. Circular shape detection (1).
3. Local texture gradient (1).
4. Edge detection (1).
5. Color consistency (1).

Feature 2: Conjunction of several blood vessels:
6. Vessel edge detection (1).
7. Vessel thickness estimation (1).
8. Vessel directionality (1).
9. Vessel junction detection (1).
10. Vessel curvature (1).

Feature 3: Rounded dark border around the bright spot:
11. Contrast difference (1).
12. Border shape (1).
13. Edge sharpness (1).
14. Shadowing effects (1).

Feature 4: Not on the black part of the image:
15. Background segmentation (1).
16. Foreground detection (1).
17. Image region classification (1).

Image describing structure of base-64 transformer model is provided in Figure 1.

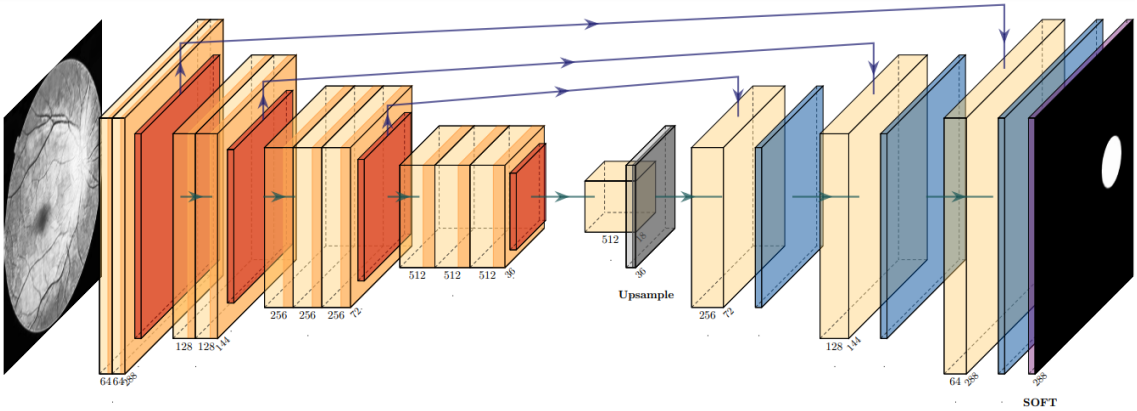

Figure 1: Encoder-decoder architecture for current sample.

3.2. Image preprocessing

To enhance image quality and improve feature extraction, histogram equalization and CLAHE (Contrast Limited Adaptive Histogram Equalization) in the LAB color space are applied. These techniques help normalize contrast and highlight relevant structures in the fundus images. The choice of the LAB color space is inspired by research from [26, 27], which demonstrates that optimized color spaces can improve neural network visibility and feature extraction in medical imaging.

3.3. Compression techniques

Iterative structured pruning is being applied to model, progressively removing low-importance weights while monitoring accuracy loss, resulting model was fine-tuned. In this article pruning was pretty aggressive (50%). Dropout layers (30% rate) were also applied to force model to compress knowledge and decrease amount of knowledge lost on pruning application, lessening retrain time and shortening preparation for further compression.

A smaller student model is trained to mimic the larger teacher model, learning its essential features while significantly reducing parameter count. In this article larger first iteration model was used as teacher model for all further iterations of student model learning.

Post-training quantization was employed to convert 32-bit floating-point weights to 8-bit integers, reducing memory consumption. Spoiler: memory consumption remained unchanged, inference time remained largely unchanged (5% speed increase), space on disk actually increased (why?).

K-Means clustering was used to group similar weights, replacing them with shared values to reduce parameter storage. Spoiler: ineffective for accuracy, inference time unchanged, memory consumption unchanged.

Transfer learning was employed to teach the first large iterations of the transformer faster, leveraging pre-trained models to accelerate convergence and improve performance. Prior research has demonstrated the effectiveness of transfer learning in deep learning applications, particularly in medical imaging tasks, where pre-trained architectures significantly enhance detection accuracy and efficiency [28].

3.4. Transformer model considerations

In addition to the CNN-based approach detailed above, current research evaluates an image transformer model for optic nerve extraction. Transformer architectures are renowned for their self-attention mechanisms, which excel at capturing long-range dependencies and fine-grained contextual relationships-a quality that is particularly beneficial for specialized tasks in medical imaging. In the context of optic nerve segmentation, transformers are capable of discerning subtle variations in image features that are crucial for accurate detection.

While recent literature has introduced a range of advanced compression strategies tailored to transformers (example - low-rank factorization [29, 30] and sparse attention mechanisms [31, 32]), current work does not implement these methods. Instead, the transformer model is utilized in its standard form to serve as a benchmark for specialized model performance. This approach allows to directly compare the efficiency gains achieved through CNN compression techniques - pruning, knowledge distillation, quantization, and weight sharing - with the inherent advantages of transformer-based models operating under similar CPU constraints.

By focusing on this comparison, we seek to determine where the transformer's compressed performance can match accuracy of base transformer model. This evaluation is critical to overall theme of liminal efficiency, where we explore the threshold at which aggressive compression techniques can be applied without incurring unacceptable losses in accuracy, while considering the trade-offs between generalist and specialist model architectures. Liminal efficiency is defined as <5% accuracy loss from starting model without losing binary feature classification metrics and increasing performance.

4. Experiments and testing

For experiments we use datasets ODIR-5K [33] – local ophthalmologist-provided datasets, which consists of 500 random images selected for training and evaluation.

We calculated following metrics: inference time, total training time, number of training epochs for sufficient results (here training epoch is one run over training image set), evaluation prediction count, total true positive predictions, total true negative predictions, total false positive predictions, total false negative predictions, additional false detections, model size, and complex metrics: accuracy, recall, and F1-score.

For testing the results 871 images tested in a CPU environment.

Experiments are conducted on model bases of 64, 32, 16, 8. Iteration model graphic images for reference of differences in iteration for base 32 and base 16 are displayed on Figure 2 and Figure 3 respectively (difference is in thickness of convolution blocks – these are multipliers of base).

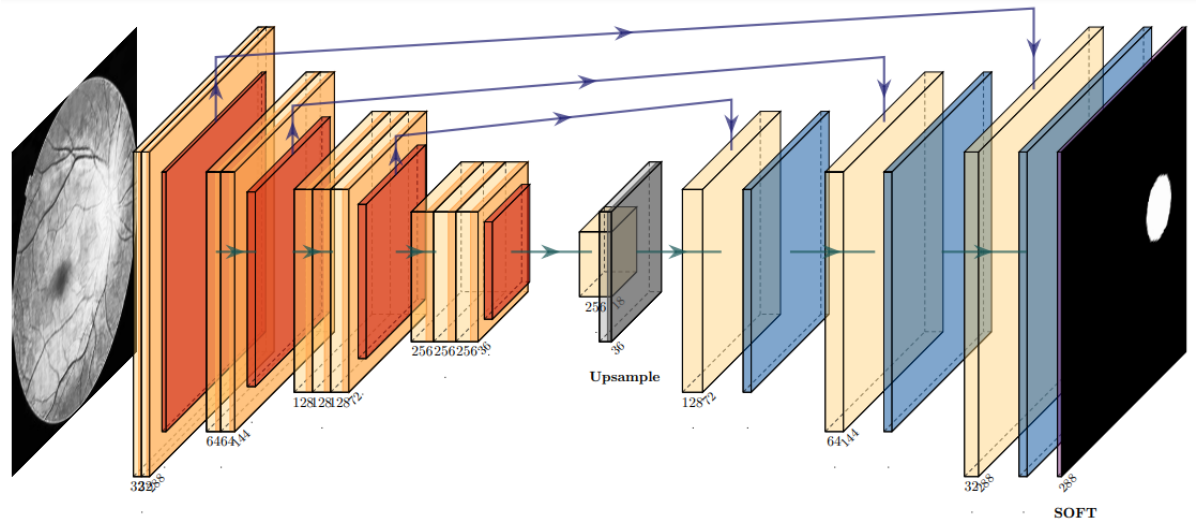

Figure 2: Encoder-decoder architecture for base of 32.

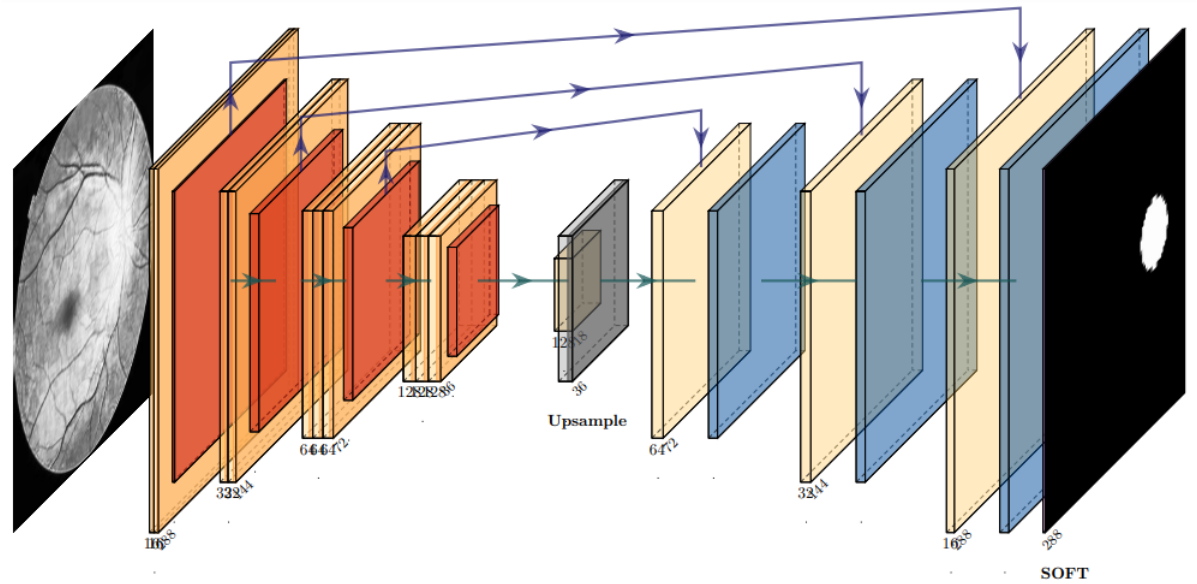

Figure 3: Encoder-decoder architecture for base of 16.

Training was conducted using k-fold cross-validation algorithm with resulting loss being extracted from evaluation fold. Additionally, to preserve edge cases knowledge a set of highest loss cases in maximum size of 1 fold was assembled and used every step to additionally randomly select

a batch to train upon using weighted random where higher loss cases are selected more often. Both knowledge distillation and pruning were trained on similar edge cases.

4.1. Pruning experiments

Pruning experiment results are provided in Table 1. Pruning experiments involve following:

1. Record metrics for each iteration, epoch and training session.
2. At iteration 1 transfer-learning is applied to model encoder blocks 1 through 4, then dropout is applied at each activation except the very last block using dropout-30% layers. After that model is trained to find optic nerve until sufficient results are obtained.
3. At subsequent iterations: previous iteration model is pruned by 50% with dropout turned off, dropout is turned back on after 1 training epoch and model is subsequently fine-tuned until sufficient accuracy is obtained.
4. Repeat for following "bases" of transformer: 64 (initial), 32, 16, 8. We stopped at base 8 because model stopped converging fast enough requiring obnoxious amount of training time (recorded results are results of first converged training, however last iteration did not converge at all).

Table 1
Results of pruning learning

Transformer configuration	x64 dropout	x32 dropout	x16 dropout	x8 dropout
Size	152 M	38 M	9.5 M	2.4 M
Train epochs for sufficient result	24	30	56	122
Train time	3h 34m	4h 15m	5h 57m	13h 1m
Total train time, Σ	3h 34m	7h 49m	13h 46m	1d 2h 47m
Inference time	0.54 s	0.2 s	0.09 s	0.06 s
Accuracy	98.16%	98.28%	95.29%	54.54%
Precision	98.73%	98.96%	98.93%	97.34%
Recall	99.42%	99.30%	96.29%	55.36%
F1-score	99.07%	99.13%	97.59%	70.58%
Additional false detections	16	18	107	346
Edge case detection	yes	yes	yes	no

4.2. Knowledge distillation

Knowledge distillation experiment results are provided in Table 2. Knowledge distillation experiments involve following:

1. Record metrics for each iteration, epoch and training session.
2. At iteration 1 transfer-learning is applied to model encoder blocks 1 through 4. After that model is trained to find optic nerve until sufficient results are obtained.
3. At subsequent iterations: use results from iteration 1 as training target.
4. Repeat for following iteration "bases" of transformer: 64(initial, step 2), 32, 16, 8. I stopped at base 8 because model(again) stopped converging fast enough, requiring obnoxious amount of training time. This time base 8 model was trained for another 40 epochs till sufficient accuracy.

Table 2
Results of knowledge distillation

Transformer configuration	x64	x32 t	x16	x8
Train epochs for sufficient result	10	78	80	160
Train time	1h 3m	10h 21m	7h 45m	13h 26m
Total train time	1h 3m	11h 24m	8h 48m	14h 29m
Inference time	0.6 s	0.2 s	0.09s	0.06s
Accuracy	99.20%	94.60%	90.36%	94.37%
Precision	99.53%	99.63%	100.00%	95.56%
Recall	99.65%	94.88%	90.23%	98.67%
F1-score	99.59%	97.20%	94.87%	97.09%
Additional false detections	1	11	41	10
Edge case detection	yes	yes	no	no

Resulting images that denote edge cases handling are displayed in Figures 4-8.

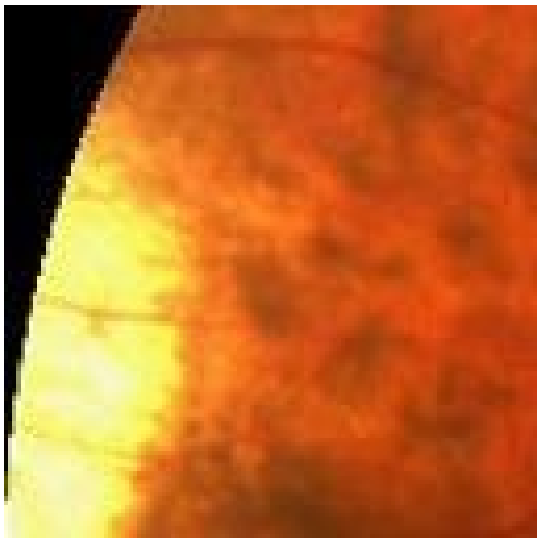

Figure 4: x16 dropout-pruning – positive edge case result – here blood vessels did converge to where optic nerve is, even if it outside the visible part. Knowledge distillation did not grab anything. [Local Image] result file of detection on 4330_left image using x16 dropout model detection.

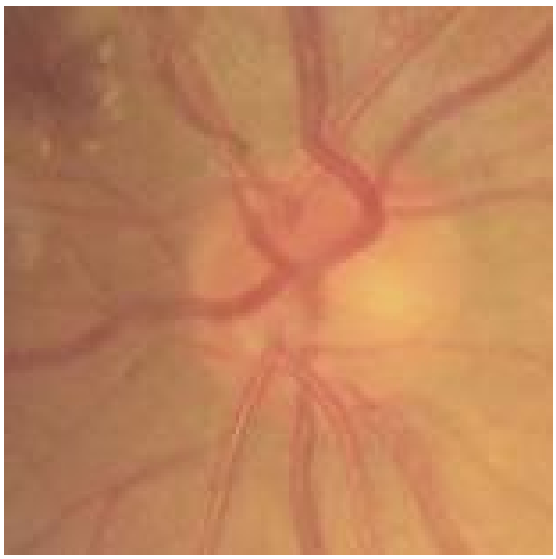

Figure 5: x16 dropout-pruning – positive edge case result - subtle visibility. [Local Image] result file of detection on 4381_left image using x16 dropout model detection.

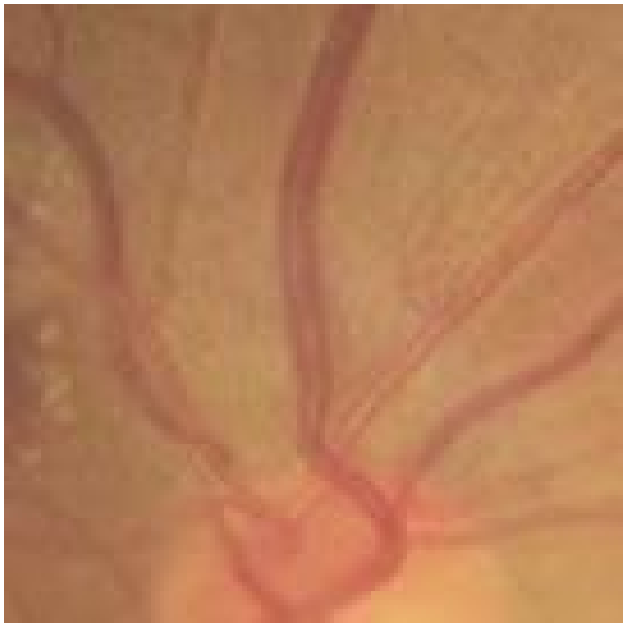

Figure 6: x16 no-dropout knowledge distillation — negative edge case result – subtle visibility, round bright feature was not grabbed. [Local Image] result file of detection on 4381_left image using x16 no-dropout model detection.

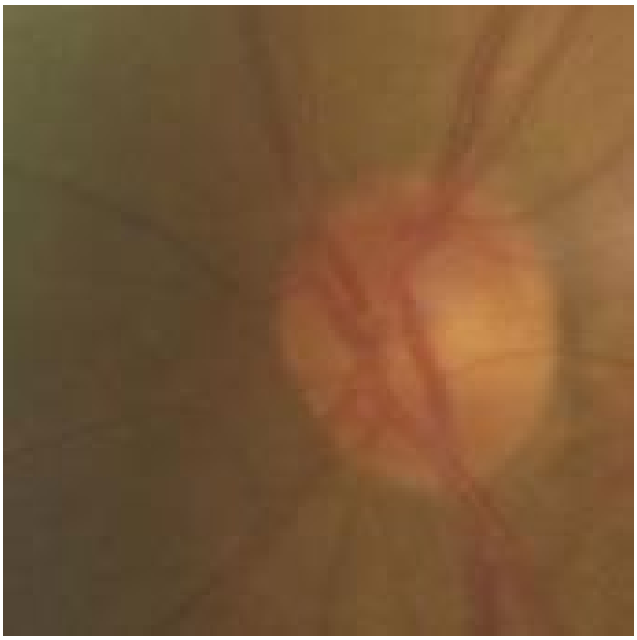

Figure 7: x16 dropout-pruning – positive edge case result — correct subtle visibility feature was grabbed. [Local Image] result file of detection on 4288_left image using x16 dropout model detection.

Figure 8: x16 no-dropout knowledge distillation – negative edge case result – wrong feature. [Local Image] result file of detection on 4288_left image using x16 no-dropout model detection.

4.3. Quantization testing

Quantization testing experiments involve following:

1. Prepare final model from training for certain model base for quantization.
2. Train prepared model to adjust it for quantization and quantize when results are sufficient.
3. Evaluate performance.

Results: gains after quantization are negligible. No useful results.

4.4. Weight sharing

Weight sharing testing involve following: Use final model from training for certain model base and apply weight sharing, fine-tune and reapply weight sharing till accuracy is acceptable.

Results: unusable, breaks detection. Training and weight sharing after each epoch did not improve result. Applying weight sharing after every training did not improve resulting accuracy. Applying weight sharing after several epochs passed did not improve accuracy. Issue with weight sharing was that there were too many interconnected target clusters for good enough selection of optic nerve. Issue example provided in Figure 9.

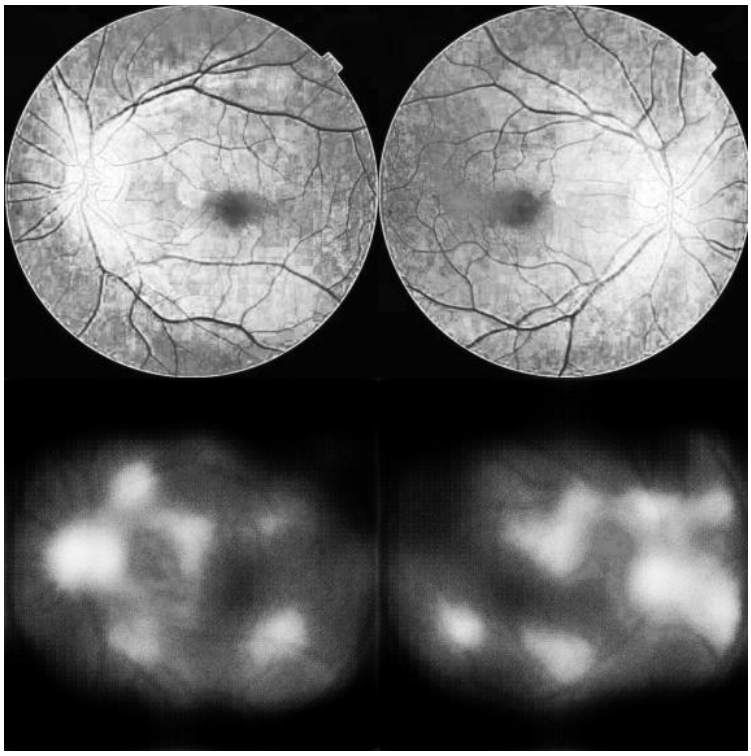

Figure 9: Weight sharing lowest loss results.

5. Conclusions

Analysis of experiments' results shows that the x16 base dropout pruning variant is a liminal point in terms of accuracy/performance ratio. It exhibits negligible accuracy loss compared to previous results while being smaller in size and 5.8 times faster than the x64 base transformer. Moreover, it retains edge case detection capability, indicating superior knowledge retention and fine-tuning potential. This variant remains undertrained and can be fine-tuned further along with un-constraining with setting dropout to 0. Alternatively, data from knowledge distillation experiments suggest that the x8 base transformer has performed well in general cases. Therefore, an x8 no-dropout pruned variant may be obtained from the x16 dropout 30 variant by pruning it to x8 base and fine-tuning it with dropout set to 0 for maximum performance and knowledge retention.

Declaration on Generative AI

During the preparation of this work, the authors used X-GPT-4o in order to: Grammar and spelling check. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] K. Allison, D. Patel, O. Alabi, Epidemiology of Glaucoma: The Past, Present, and Predictions for the Future, Cureus (2020). doi:10.7759/cureus.11686.

[2] Y.-C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, C.-Y. Cheng, Global Prevalence of Glaucoma and Projections of Glaucoma Burden through 2040, Ophthalmology 121.11 (2014) 2081–2090. doi:10.1016/j.ophtha.2014.05.013.

[3] J. B. Jonas, T. Aung, R. R. Bourne, A. M. Bron, R. Ritch, S. Panda-Jonas, Glaucoma, Lancet 390.10108 (2017) 2183–2193. doi:10.1016/s0140-6736(17)31469-1.

[4] X. Huang, M. R. Islam, S. Akter, F. Ahmed, E. Kazami, H. A. Serhan, A. Abd-alrazaq, S. Yousefi, Artificial intelligence in glaucoma: opportunities, challenges, and future directions, Biomed. Eng. OnLine 22.1 (2023). doi:10.1186/s12938-023-01187-8.

[5] T. Hovorushchenko, A. Herts, Ye. Hnatchuk, Concept of Intelligent Decision Support System in the Legal Regulation of the Surrogate Motherhood. CEUR-WS 2488 (2019) 57-68.

[6] T. Hovorushchenko, Ye. Hnatchuk, A. Herts, A. Moskalenko, V. Osyadlyi, Theoretical and Applied Principles of Information Technology for Supporting Medical Decision-Making Taking into Account the Legal Basis, CEUR-WS 3038 (2021) 172-181.

[7] Y. Hnatchuk, T. Hovorushchenko, O. Pavlova, Methodology for the development and application of clinical decisions support information technologies with consideration of civil-legal grounds, Radioelectron. Comput. Syst. № 1 (2023) 33–44. doi:10.32620/reks.2023.1.03.

[8] T. Hovorushchenko, A. Herts, Y. Hnatchuk, O. Sachenko, Supporting the Decision-Making About the Possibility of Donation and Transplantation Based on Civil Law Grounds, in: Advances in Intelligent Systems and Computing, Springer International Publishing, Cham, 2020, p. 357–376. doi:10.1007/978-3-030-54215-3_23.

[9] J. C. Mathew, V. Ilango, V. Asha, Machine Learning Techniques, Detection and Prediction of Glaucoma– A Systematic Review, Int. J. Recent Innov. Trends Comput. Commun. 11.5s (2023) 283–309. doi:10.17762/ijritcc.v11i5s.6655.

[10] V. Kysil, P. Popov, O. Drachuk, V. Hnenna, I. Martyniuk, Concept of Information Technology for Diagnosis and Prognosis of Glaucoma Based on Machine Learning Methods. CEUR-WS. 3675 (2024) 171-181.

[11] T. Hovorushchenko, A. Moskalenko, V. Osyadlyi, Methods of medical data management based on blockchain technologies, J. Reliab. Intell. Environ. 9.1 (2022) 5-16. doi:10.1007/s40860-022-00178-1.

[12] A. Saxena, A. K. Bishwas, A. A. Mishra, R. Armstrong, Comprehensive Study on Performance Evaluation and Optimization of Model Compression: Bridging Traditional Deep Learning and Large Language Models, 2024. URL: https://arxiv.org/pdf/2407.15904.

[13] E. Dupuis, D. Novo, I. O'Connor, A. Bosio, Sensitivity Analysis and Compression Opportunities in DNNs Using Weight Sharing, in: 2020 23rd International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS), IEEE, 2020. doi:10.1109/ddecs50862.2020.9095658.

[14] Y. Na, H. H. Kim, S. B. Yoo, Shared knowledge distillation for robust multi-scale super-resolution networks, Electron. Lett. (2022). doi:10.1049/ell2.12526.

[15] S. Han, J. Pool, J. Tran, W. J. Dally, Learning both weights and connections for efficient neural networks, 2015. URL: https://arxiv.org/abs/1506.02626.

[16] J. Frankle, M. Carbin, The lottery ticket hypothesis: Finding sparse, trainable neural networks, 2019. URL: https://arxiv.org/abs/1803.03635.

[17] Y. Cheng, D. Wang, P. Zhou, T. Zhang, Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges, IEEE Signal Process. Mag. 35.1 (2018) 126–136. doi:10.1109/msp.2017.2765695.

[18] B. Wu, K. Keutzer, X. Dai, P. Zhang, Y. Wang, F. Sun, Y. Wu, Y. Tian, P. Vajda, Y. Jia, FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019. doi:10.1109/cvpr.2019.01099.

[19] A. Boggust, V. Sivaraman, Y. Assogba, D. Ren, D. Moritz, F. Hohman, Compress and Compare: Interactively Evaluating Efficiency and Behavior Across ML Model Compression Experiments, IEEE Trans. Vis. Comput. Graph. (2024) 1–11. doi:10.1109/tvcg.2024.3456371.

[20] P. V. Dantas, W. Sabino da Silva, L. C. Cordeiro, C. B. Carvalho, A comprehensive review of model compression techniques in machine learning, Appl. Intell. (2024). doi:10.1007/s10489-024-05747-w.

[21] T. Schick, H. Schütze, Comparing specialized small and general large language models, 2024. URL: https://arxiv.org/abs/2402.12819.

[22] C. Shi, Y. Su, C. Yang, Y. Yang, D. Cai, Specialist or Generalist? Instruction Tuning for Specific NLP Tasks, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Stroudsburg, PA, USA, 2023. doi:10.18653/v1/2023.emnlp-main.947.

[23] S. Arif, A. H. Azeemi, A. A. Raza, A. Athar, Generalists vs. Specialists: Evaluating Large Language Models for Urdu, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Stroudsburg, PA, USA, 2024, c. 7263–7280. doi:10.18653/v1/2024.findings-emnlp.426.

[24] N. I. Morar, N. Holtham, L. Hackel, K. Davami, M. Sharma, A. DeWald, R. Roy, Effects of high energy laser peening followed by pre-hot corrosion on stress relaxation, microhardness and fatigue life and strength of single crystal nickel CMSX-4® superalloy, Procedia Struct. Integr. 57 (2024) 625–632. doi:10.1016/j.prostr.2024.03.069.

[25] A. Salvador, N. Kolkin, J. Johnson, Recurrent neural networks for semantic instance segmentation, 2017. URL: https://arxiv.org/abs/1712.00617.

[26] A. Sarhan, A. Al-Khaz'Aly, A. Gorner, A. Swift, J. Rokne, R. Alhajj, A. Crichton, Utilizing Transfer Learning and a Customized Loss Function for Optic Disc Segmentation from Retinal Images, in: Computer Vision – ACCV 2020, Springer International Publishing, Cham, 2021, p. 687–703. doi:10.1007/978-3-030-69541-5_41.

[27] A. Tiwari, M. Pant, Optimized Deep-Neural Network for Content-based Medical Image Retrieval in a Brownfield IoMT Network, ACM Trans. Multimedia Comput., Commun., Appl. (2022). doi:10.1145/3546194.

[28] M. Christopher, A. Belghith, C. Bowd, J. A. Proudfoot, M. H. Goldbaum, R. N. Weinreb, C. A. Girkin, J. M. Liebmann, L. M. Zangwill, Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs, Sci. Rep. 8.1 (2018). doi:10.1038/s41598-018-35044-9.

[29] S. Wang, B. Li, M. Khabsa, H. Fang, H. Ma, Linformer: Self-attention with linear complexity, 2020. URL: https://arxiv.org/abs/2006.04768.

[30] K. Choromanski, V. Likhosherstov, D. Dohan, X. He, J. Jiao, et al., Rethinking attention with performers, 2021. URL: https://arxiv.org/abs/2009.14794.

[31] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, 2020. URL: https://arxiv.org/abs/2004.05150.

[32] M. Zaheer, G. Guruganesh, N. Koliousis, A. Dubey, J. Ainslie, et al., Big Bird: Transformers for longer sequences, 2020. URL: https://arxiv.org/abs/2007.14062.

[33] A. Mvd, Ocular Disease Recognition (ODIR-5K) Dataset, 2019. URL: https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k/data.