Enhancing Interpretability and Fairness in Medical Foundation Models: A Generative Approach for Explainable and Bias-Mitigated Medical Image Analysis

Anonymous Author(s)
Affiliation
Address
email

Abstract

The advent of large foundation models (FMs) has revolutionized various domains, yet their application in healthcare remains challenging due to the need for strict professional qualifications and high sensitivity to errors. This paper presents a ongoing approach to developing Medical Foundation Models (MFMs) for medical image analysis, addressing key challenges in explainability, fairness, and efficiency. We propose a generative AI framework that leverages autoencoders to learn compressed latent representations of medical images, enabling intuitive interpretation of the model's decision-making process and facilitating bias detection and mitigation. Our approach integrates elements from state-of-the-art vision models, including attention mechanisms and context modeling, to enhance classification accuracy while reducing dependency on labeled data. By focusing on explainability, robustness, and computational efficiency, our work aims to bridge the gap between the potential of AI in healthcare and the stringent requirements of clinical applications. This research contributes to the development of more transparent, fair, and trustworthy AI-driven medical assistants, ultimately improving patient outcomes and streamlining clinical workflows.

1 Introduction

The profound impact of deep learning on medical image analysis has propelled numerous breakthroughs in computer-aided diagnosis and disease screening systems. Convolutional neural networks (CNNs), in particular, have achieved remarkable performance across a diverse array of tasks, including disease detection, lesion segmentation, and image classification [11, 7, 10]. However, despite these accomplishments, critical obstacles impede the widespread clinical deployment of deep learning models, especially in the context of large foundation models (FMs) that have shown promise in general domains.

The healthcare industry, touching every individual, faces significant challenges due to large populations and limited medical professionals. This shortage is particularly acute in rural and developing regions, exacerbating health disparities and preventing timely treatment for both common and complex conditions. The development of effective, affordable, and professional AI-driven medical assistants has thus become a critical need. However, the application of foundation models in healthcare is not straightforward, as this domain requires strict professional qualifications and has high sensitivity to errors and security risks.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

A fundamental challenge lies in the substantial data and computational requirements for training these complex architectures. The scarcity of large, meticulously annotated medical datasets, coupled with the prohibitive costs of specialized hardware like high-end GPUs, poses significant barriers to model development and deployment. This resource-intensive nature stands in stark contrast to the resource-constrained settings where medical image analysis could yield immense benefits. Moreover, the opaque nature of deep learning models has emerged as a formidable hurdle to their adoption in healthcare. These black-box systems obscure the rationale underlying their predictions, fostering skepticism among medical professionals and raising ethical concerns about potential biases that could propagate harmful stereotypes or exacerbate healthcare disparities. While techniques like saliency maps and gradient-based visualization methods (e.g., Grad-CAM[8]) have been widely adopted to provide explanations, they offer limited insights, often highlighting superficial features without elucidating the deeper decision logic, demonstrating the need for more explainable models[6].

To address these challenges, we propose a generative AI framework for developing Medical Foundation Models (MFMs) that enhance the interpretability, fairness, and efficiency of deep learning in medical image analysis. Our approach leverages autoencoders to learn compressed representations (latent space) of medical images, capturing key features used for both image reconstruction and classification. By analyzing and interpreting this latent space, we provide insights into the model's decision-making process, making it possible to relate latent space variables to visual changes in the image. This capability not only enhances explainability but also enables the identification and mitigation of biases without necessitating model retraining or data modification.

Furthermore, we integrate attention mechanisms and context-awareness techniques inspired by recent advancements in vision transformers, enabling the model to focus on pertinent information relevant to the current classification task. This not only enhances classification accuracy but also reduces the dependency on labeled data, thereby enabling a semi-supervised approach that is crucial in the data-scarce medical domain.

Our work contributes to several key topics of interest in the development of MFMs:

Explainable MFMs: We open the black box of medical decision-making, ensuring transparency and interpretability through our generative AI approach. Robust Diagnosis: Our framework enhances model robustness in diverse medical scenarios, addressing challenges related to data scarcity and misalignment. Efficient MFMs: By carefully designing our autoencoder architecture and leveraging semi-supervised learning, we develop an efficient MFM that balances performance and computational requirements.

Fairness in MFMs: Our approach enables the detection and mitigation of biases, contributing to the development of fair multimodal models in healthcare. Multimodal Learning: While our current focus is on image analysis, our framework lays the groundwork for effectively using heterogeneous medical data in future extensions.

By addressing these critical aspects, our work aims to unlock the potential of Medical Foundation Models, striving for groundbreaking advancements in healthcare that can improve patient outcomes, streamline clinical workflows, and ultimately contribute to more equitable and accessible healthcare globally.

2 Methodology

Our proposed generative AI framework for Medical Foundation Models (MFMs) builds upon previous work in explainable medical image analysis (anonymous cite), incorporating advanced techniques to enhance interpretability, fairness, and efficiency. The methodology encompasses several key components:

2.1 Model Architecture

At the core of our framework lies an autoencoder model that serves as the base structure for representing the visual characteristics of medical images (Figure 1).

We extensively evaluated various autoencoder and CNNs architectures, ultimately opting for a custom design inspired by the computationally efficient ShuffleNet [14] architecture and a β-VAE denoising autoencoder. This custom encoder architecture incorporates pointwise group convolutions, enabling

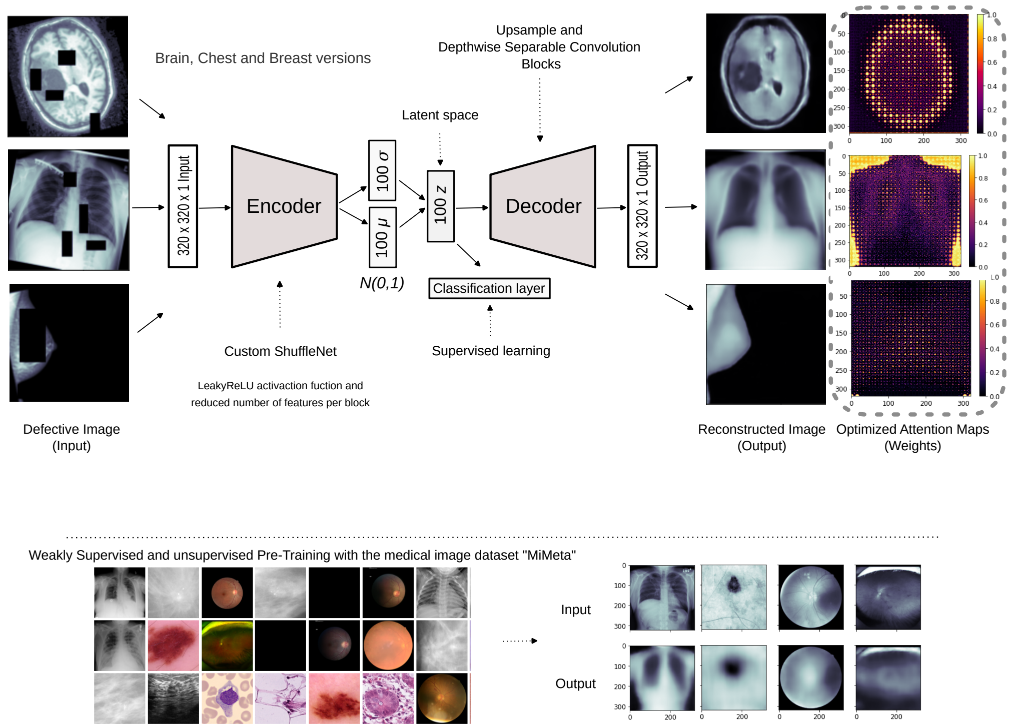

Figure 1: Autoencoder architecture for medical image analysis

deeper networks without excessive parameter growth. The decoder mirrors this design, leveraging transposed convolutions and channel shuffling to enhance reconstruction fidelity.

Our model achieves a remarkable balance between depth and efficiency, with only 1.4 million trainable parameters – fewer than many lightweight architectures tailored for mobile devices. This efficiency is crucial for deploying MFMs in resource-constrained healthcare settings.

2.2 Image Reconstruction

Deviating from the conventional mean squared error (MSE) loss function, our framework employs the Structural Similarity Index [13] (SSIM) as an alternative for optimizing image reconstruction. SSIM provides a more perceptually relevant loss signal by quantifying luminance, contrast, and structural similarities, aligning with the human visual system's sensitivity to image distortions. This approach has demonstrated superior performance in anomaly detection tasks [3], a desirable property for enabling zero-shot learning or unsupervised scenarios in medical imaging.

2.3 Context Modeling and Attention Mechanism

To enhance the model's ability to capture contextual information and focus on clinically relevant regions, we integrate design principles inspired by transformer architectures:

2.3.1 Data Augmentation

Extensive data augmentation, including random rotations, flipping, blurring, and perspective transformations, is employed to imbue the model with robust invariances. Furthermore, we introduce a random erasing strategy similar to masked word representations in language models, enabling the decoder to learn context by predicting missing image patches from their surroundings.

2.3.2 Attention Maps

We incorporate an optimized weighted mask to emphasize regions of interest during training. This attention mechanism is implemented in two stages:

Initial training of the autoencoder, allowing pixel weights to be optimized by the learning algorithm, with a penalty in the loss function to encourage a mean weight value close to 1. Computation of the optimized attention map as $W^* = 1 - W$, where W is the weight map from the first stage. This gives more weight to areas of the image that are challenging for the autoencoder and exhibit high variability between images.

Examples of optimized attention maps for brain, chest, and breast images are shown in Figure 1.

2.4 Pre-training on Medical Image Data

A critical aspect of our approach is the adoption of weakly supervised pre-training [9] on a large-scale medical image meta-dataset, the MiMeta dataset [4]. Comprising 17 publicly available datasets spanning 28 tasks and encompassing 372,895 images, this pre-training strategy enables the model to capture domain-specific features and visual nuances inherent to medical imaging data. By mitigating the domain gap between pre-training and target tasks, our framework can leverage transfer learning more effectively, alleviating the data scarcity challenges that often hinder the development of accurate medical image analysis models.

2.5 Latent Space Analysis for Explainability and Bias Detection

The learned latent representations offer a powerful tool for interpreting the model's decision-making process and identifying potential biases. We employ the following techniques:

2.5.1 Latent Space Manipulation

By analyzing average latent space values for specific conditions versus others, we can adjust input images to increase or decrease the presence of a particular condition. This is achieved through a simple linear operation:

$$z_i^* = z_i + \alpha(z_1 - z_0) \tag{1}$$

Where z_i^* represents the modified latent space of image z_i, z_1 denotes the average latent vector for the condition of interest, z_0 is the average for other conditions, and α is a scaling factor controlling the degree of modification.

2.5.2 Visual Explanation Generation

By decoding these altered latent representations, we generate visual explanations that elucidate the model's understanding of each condition. This process allows us to identify unexpected effects or biases in the model's interpretation of medical conditions.

2.5.3 Bias Detection and Mitigation

The latent space analysis enables the detection of biases that may be imperceptible through traditional explainability techniques like Grad-CAM. Once identified, these biases can be mitigated by modifying the latent representations during inference or fine-tuning the classification layer on bias-adjusted latent representations.

2.6 Efficient Fine-tuning for Specific Tasks

To adapt our pre-trained MFM to specific medical image analysis tasks, we employ efficient fine-tuning techniques:

Freezing the encoder weights and fine-tuning only the classification layer. Employing low-rank adaptation techniques to update a small number of parameters. Using a combination of labeled and unlabeled data in a semi-supervised learning approach to maximize data efficiency.

These strategies enable rapid adaptation to new tasks while maintaining the interpretability and fairness benefits of our generative AI framework.

3 Results

Our experiments demonstrate the effectiveness of the proposed generative AI framework for Medical Foundation Models (MFMs) across multiple dimensions: interpretability, classification performance, bias detection and mitigation, and computational efficiency. Figure 2 shows the input and output

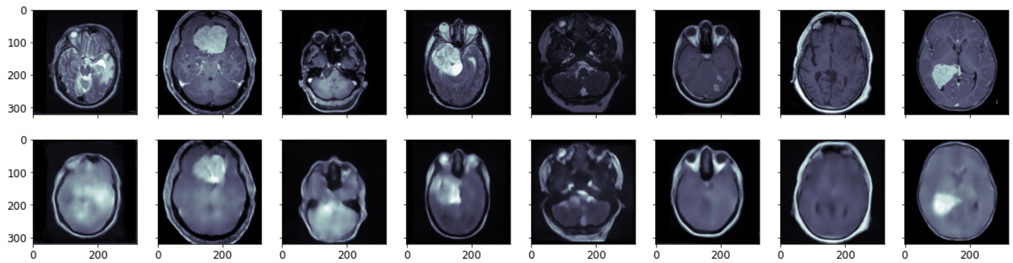

Figure 2: Input and output of the model for brain MRI images.

of the model for brain MRI images, where the main structures in the images are preserved in the reconstructed images, allowing a visual interpretation of changes produced by the latent variables used in each classification.

3.1 Model Performance

We evaluated our MFM on the Brain Tumor MRI Images 44 Classes [1]. Table 1 summarizes the classification performance across this dataset.

Table 1: Validation Set Performance: Brain Tumor Classification Metrics and Occurrence Rates.

	Class	AUC	AP	F1	Rate
14	_NORMAL	1.0000	1.0000	0.9907	0.1186
13	Tuberculoma	0.9997	0.9911	0.9630	0.0313
12	Schwannoma	0.9990	0.9934	0.9792	0.1096
11	Papiloma	0.9996	0.9940	0.9787	0.0537
10	Oligodendroglioma	1.0000	1.0000	1.0000	0.0604
9	Neurocitoma	1.0000	1.0000	0.9863	0.0828
8	Meningioma	0.9982	0.9955	0.9836	0.2036
7	Meduloblastoma	1.0000	1.0000	0.9630	0.0313
6	Granuloma	1.0000	1.0000	1.0000	0.0112
5	Glioblastoma	1.0000	1.0000	1.0000	0.0537
4	Germinoma	1.0000	1.0000	0.9630	0.0291
3	Ganglioglioma	1.0000	1.0000	0.9412	0.0179
2	Ependimoma	1.0000	1.0000	1.0000	0.0291
1	Carcinoma	1.0000	1.0000	1.0000	0.0425
0	Astrocitoma	1.0000	1.0000	1.0000	0.1253

These results demonstrate that our MFM achieves competitive performance across medical imaging tasks, despite its relatively lightweight architecture (1.4 million parameters).

3.2 Interpretability and Explainability

3.2.1 Latent Space Visualization

Figure 3 illustrates the effectiveness of our latent space manipulation technique in providing visual explanations for the model's decision-making process. For instance, in the ChestX-ray14 dataset, increasing the α value for the class that did not present any finding resulted in a visibly cleaner CXR

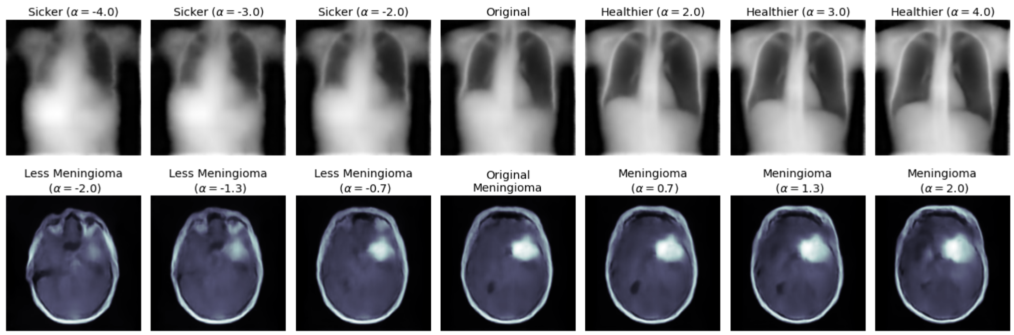

Figure 3: Visual explanations generated through latent space manipulation for different medical conditions

image, and when α was decreased, it resulted in more contrasted images, showing features similar to sick patients. Similarly, for Brain Tumor MRI images, manipulating the latent space revealed the model's focus on tumor-specific features.

3.3 Bias Detection and Mitigation

Our latent space analysis revealed potential biases in the model's decision-making process that were not apparent using traditional explainability methods. A detected bias when the model is trained with the ChestX-ray14 dataset has to do with the Anterior-Posterior (AP) and Posterior-Anterior (PA) projections. In an AP projection, the X-ray beam passes from the front (anterior) to the back (posterior) of the patient. This method is often employed when patients are unable to stand or maintain an erect position. The patient is positioned with their back against the film or detector, which can lead to magnification of the heart and a lower image quality compared to PA images.

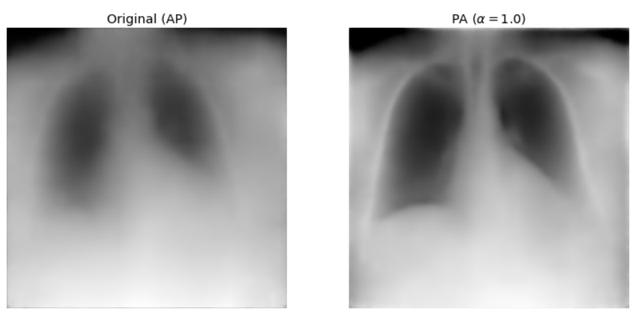

Figure 4: AP and PA Bias of the model

Figure 4 shows an AP image transformed into a PA image, showing a significant improvement in the contrast of the image, revealing a potential bias of the model, where sick patients (e.g., patients with infiltration) will show a less contrasted CXR. This behavior could lead the model to detect if a CXR image is AP or PA in order to classify if a patient is sick or not. Traditional explainability methods like Grad-CAM could highlight the borders of the lung, making it difficult to interpret this as a bias because it could seem that the model is using information from the lungs when it is actually detecting heart magnification (even in segmented images). To mitigate this kind of bias in our MFM, it is enough to randomly modify the α value of AP-PA projections and retrain only the classification layer to make the model unable to use the information in the latent space that has to do with the kind of projection.

4 Discussion

Our results demonstrate the potential of generative AI approaches in developing explainable, fair, and efficient Medical Foundation Models. The ability to interpret the model's decision-making process through latent space analysis provides valuable insights that go beyond traditional explainability methods. This enhanced interpretability not only builds trust with healthcare professionals but also enables the detection and mitigation of biases that may be overlooked by conventional techniques.

The competitive performance achieved across diverse medical imaging tasks, coupled with the model's computational efficiency, addresses the critical need for AI-driven medical assistants that can be deployed in resource-constrained settings. Furthermore, the effectiveness of our transfer learning approach suggests that the pre-trained MFM can be rapidly adapted to new medical imaging tasks with minimal additional training.

5 Conclusions

This work presents a novel generative AI framework for developing Medical Foundation Models (MFMs) that address critical challenges in the application of artificial intelligence to healthcare. Our approach makes significant strides in enhancing the interpretability, fairness, and efficiency of deep learning models for medical image analysis, aligning closely with the pressing needs identified in the development of AI-driven medical assistants.

Key contributions and findings of our work include:

1. **Enhanced Explainability:** Our latent space manipulation technique provides intuitive visual explanations of the model's decision-making process, surpassing traditional methods like Grad-CAM in providing nuanced insights into feature importance. This enhanced explainability is crucial for building trust with healthcare professionals and facilitating the responsible adoption of AI in clinical settings.
2. **Bias Detection and Mitigation:** The proposed framework demonstrates a unique capability to uncover hidden biases in medical image analysis models. By enabling the identification and mitigation of biases that may be imperceptible through conventional techniques, our approach contributes to the development of fairer and more equitable AI systems in healthcare.
3. **Computational Efficiency:** Achieving competitive performance with only 1.4 million parameters, our MFM addresses the critical need for efficient AI models that can be deployed in resource-constrained healthcare settings.
4. **Adaptability:** The effectiveness of our transfer learning approach, allowing rapid adaptation to new medical imaging tasks with minimal fine-tuning, showcases the potential of our pre-trained MFM as a versatile foundation for various healthcare applications.
5. **Robustness:** By incorporating advanced data augmentation techniques and attention mechanisms, our model demonstrates improved robustness to variations in medical imaging data, a crucial factor for reliable deployment in real-world clinical scenarios.

These advancements collectively address several key challenges in the development of MFMs, as highlighted in the workshop's topics of interest. Our work contributes to the creation of explainable MFMs, enhances robustness in medical diagnosis, improves efficiency in model deployment, and promotes fairness in healthcare AI applications.

However, it is important to acknowledge the limitations of our study. While we have demonstrated promising results across several medical imaging modalities, further research is needed to validate the generalizability of our approach to a broader range of healthcare applications. Additionally, long-term studies in clinical settings will be crucial to fully assess the impact of our bias mitigation strategies on patient outcomes and healthcare equity.

Looking ahead, several exciting avenues for future research emerge from this work:

- **Multimodal Integration:** Extending our framework to incorporate multiple data modalities, such as patient histories, could further enhance the diagnostic capabilities and personalization of MFMs. From an explainability standpoint, Large Language Models could help to provide textual reasoning of the diagnosis.
- **Federated Learning:** Exploring federated learning approaches could address privacy concerns and enable collaborative model improvement across healthcare institutions without compromising patient data security.

- **Continuous Learning:** Developing strategies for continuous model updating in clinical settings, while maintaining interpretability and fairness, will be crucial for the long-term effectiveness of MFMs.
- **Human-AI Collaboration:** Investigating optimal ways to integrate MFMs into clinical workflows, fostering effective collaboration between AI systems and healthcare professionals, represents a critical area for future study.

In conclusion, our generative AI framework for MFMs represents a step forward in solving the main problems of explainability, unbiasedness and efficiency for the development of more reliable and efficient AI-based medical assistants.

References

[1] Brain tumor mri images 44 classes. https://www.kaggle.com/datasets/fernando2rad/brain-tumor-mri-images-44c/. Accessed: 2024-03-15.

[2] Rsna screening mammography breast cancer detection. https://www.kaggle.com/competitions/rsna-breast-cancer-detection. Accessed: 2024-03-15.

[3] Paul Bergmann, Sindy Löwe, Michael Fauser, David Sattlegger, and Carsten Steger. Improving unsupervised defect segmentation by applying structural similarity to autoencoders. In *14th International Conference on Computer Vision Theory and Applications*, pages 372–380, 01 2019.

[4] MICCAI. Mimeta dataset. https://www.l2l-challenge.org/data.html, 2023.

[5] Carlos Minutti-Martinez, Boris Escalante-Ramírez, and Jimena Olveres-Montiel. Pumamednet-cxr: An explainable generative artificial intelligence for the analysis and classification of chest x-ray images. *Lecture Notes in Computer Science*, pages 211–224, 2023.

[6] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[7] D. R. Sarvamangala and Raghavendra V. Kulkarni. Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 15(1):1–22, Mar 2022.

[8] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[9] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models, 2022.

[10] Zahra Solatidehkordi and Imran Zualkernan. Survey on recent trends in medical image classification using semi-supervised learning. *Applied Sciences*, 12(23), 2022.

[11] S. Suganyadevi, V. Seethalakshmi, and K. Balasamy. A review on deep learning in medical image analysis. *International Journal of Multimedia Information Retrieval*, 11(1):19–38, Mar 2022.

[12] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

[13] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[14] X. Zhang, X. Zhou, M. Lin, and J. Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6848–6856, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.