DEVELOPMENT OF LOW-OVERHEAD SOFT ERROR MITIGATION TECHNIQUE FOR SAFETY CRITICAL NEURAL NETWORKS APPLICATIONS

KHALID ADAM ISMAIL HAMMAD

Doctor of Philosophy

UNIVERSITI MALAYSIA PAHANG

Universiti Malaysia PAHANG
Engineering • Technology • Creativity

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Doctor of Philosophy.

(Supervisor's Signature)

Full Name : DR. IZZELDIN IBRAHIM MOHAMED ABDELAZIZ

Position : SENIOR LECTURER

Date : 18 MAY 2021

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Universiti Malaysia Pahang or any other institutions.

____

(Student's Signature)

Full Name : KHALID ADAM ISMAIL HAMMAD

ID Number : PEL17005

Date : 18 MAY 2021

DEVELOPMENT OF LOW-OVERHEAD SOFT ERROR MITIGATION TECHNIQUE FOR SAFETY CRITICAL NEURAL NETWORKS APPLICATIONS

KHALID ADAM ISMAIL HAMMAD

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Doctor of Philosophy

College of Engineering

UNIVERSITI MALAYSIA PAHANG

MAY 2021

ACKNOWLEDGEMENTS

I am grateful and I would like to express my deepest gratitude to God Almighty and whoever supported me to complete this thesis, including my supervisor, attached university, friends, and family. First and foremost, I am thankful to my supervisor Dr. Izzeldin Ibrahim Mohamed Abdelaziz Abdelaziz for his germinal ideas, invaluable guidance, continuous encouragement and unwavering support in making this research possible. He has always impressed me with his outstanding professional conduct, his strong conviction for science, and his belief that PhD. program is only a start of a life-long learning experience. I appreciate his consistent support from the first day I applied to graduate program to these concluding moments.

I am greatly indebted to my brother, Mr. Younis Mohammed Younis for his assistance, encouragement and the sacrifices this made during this research. I really appreciate for his standing by me at all times, may Allah reward you abundantly.

Last but not least, I acknowledge my sincere indebtedness and gratitude to my family for their love, dream and sacrifice throughout my life. I am also grateful to my parent for their sacrifice, patience, and understanding that were inevitable to make this work possible. I cannot find the appropriate words that could properly describe my appreciation for their devotion, support and faith in my ability to attain my goals.

Finally, I thanks Allah for giving me good health, I am grateful to Allah for making me alive to complete this study, I look forward to him to continue to direct me in whatever step I take in life. Praise be to Allah the Lord of the world!!!

ABSTRAK

Deep Neural Networks (DNNs) telah banyak digunakan dalam aplikasi di sektor kesihatan. Aplikasi kesihatan berasaskan DNN adalah sistem keselamatan-kritikal yang memerlukan pelaksanaan kebolehpercayaan yang tinggi kerana risiko kematian atau kecederaan manusia yang tinggi sekiranya berlaku kerosakan. Beberapa pemecut DNN digunakan untuk melaksanakan model DNN ini, dan GPU saat ini adalah pemecut DNN yang paling menonjol dan didominasi. Walau bagaimanapun, GPU terdedah kepada kesilapan lembut yang secara dramatik mempengaruhi tingkah laku GPU; Kesalahan tersebut boleh merosakkan nilai data atau operasi logik, yang mengakibatkan Silent Data Corruption (SDC). Penyebaran SDC berlaku dari tahap fizikal ke peringkat aplikasi (SDC yang berlaku pada komponen perkakasan GPU) mengakibatkan salah klasifikasi objek dalam model DNN, yang membawa kepada akibat yang buruk. Pentadbiran Makanan dan Ubat-ubatan (FDA) melaporkan bahawa 1078 kejadian buruk (10.1%) adalah kesilapan yang berlaku secara tidak sengaja (iaitu, kesalahan kecil) yang melibatkan 52 kecederaan dan dua kematian. Beberapa teknik tradisional telah diusulkan untuk melindungi peranti elektronik dari kesalahan lembut dengan replikasi model DNN. Walau bagaimanapun, teknik ini menyebabkan overhed luas, prestasi, dan tenaga, menjadikannya sukar untuk dilaksanakan dalam sistem penjagaan kesihatan yang mempunyai batas waktu yang ketat. Untuk mengatasi masalah ini, kajian ini mengembangkan Teknik Mitigasi Selektif berdasarkan Redundansi Triple Modular standard (S-MTTM-R) untuk menentukan bahagian yang rentan model, membezakan kesalahan Malfunction dan Light-Malfunction. Analisis kerentanan komprehensif dilakukan dengan menggunakan penyuntik kesalahan SASSIFI pada model CNN AlexNet dan DenseNet201: lapisan, kernel, dan arahan untuk menunjukkan ketahanan kedua-dua model dan mengenal pasti bahagian yang paling rentan dan mengeraskannya dengan menyuntikkannya apabila dilaksanakan pada GPU NVIDIA. Hasil eksperimen menunjukkan bahawa S-MTTM-R mencapai peningkatan yang ketara dalam penyamaran ralat. No-Malfunction telah dinaiktaraf untuk AlexNet dari 54.90%, 67.85%, dan 59.36% menjadi 62.80%, 82.10%, dan 80.76% dalam tiga mod RF, IOA, dan IOV. Untuk DenseNet, No-Malfunction telah dipertingkatkan dari 43.70%, 67.70%, dan 54.68% menjadi 59.90%, 84.75%, dan 83.07% dalam tiga mod RF, IOA, dan IOV. S-MTTM-R memainkan peranan penting dengan menurunkan peratusan ralat dalam kes salah klasifikasi (Malfunctioni) dari 3.70% menjadi 0.38% dan 5.23% menjadi 0.23% masing-masing untuk AlexNet dan DenseNet. Hasil analisis prestasi menunjukkan bahawa S-MTTM-R mencapai overhead yang lebih rendah berbanding dengan teknik perlindungan yang terkenal: Algorithm-Based Fault Tolerance (ABFT), Double Modular Redundancy (DMR), and Triple Modular Redundancy (TMR). Berdasarkan hasil ini, kajian ini menunjukkan bukti kuat bahawa S-MTTM-R yang dikembangkan berhasil mengurangkan kesalahan lembut bagi model DNN pada GPU dengan overhead rendah dalam tenaga, prestasi, dan kawasan yang menunjukkan peningkatan yang luar biasa dengan kebolehpercayaan model dalam domain rawatan kesihatan.

ABSTRACT

Deep Neural Networks (DNNs) have been widely applied in healthcare applications. DNN-based healthcare applications are safety-critical systems that require high-reliability implementation due to a high risk of human death or injury in case of malfunction. Several DNN accelerators are used to execute these DNN models, and GPUs are currently the most prominent and the dominated DNN accelerators. However, GPUs are prone to soft errors that dramatically impact the GPU behaviors; such error may corrupt data values or logic operations, which result in Silent Data Corruption (SDC). The SDC propagates from the physical level to the application level (SDC that occurs in hardware GPUs' components) results in misclassification of objects in DNN models, leading to disastrous consequences. Food and Drug Administration (FDA) reported that 1078 of the adverse events (10.1%) were unintended errors (i.e., soft errors) encountered, including 52 injuries and two deaths. Several traditional techniques have been proposed to protect electronic devices from soft errors by replicating the DNN models. However, these techniques cause significant overheads of area, performance, and energy, making them challenging to implement in healthcare systems that have strict deadlines. To address this issue, this study developed a Selective Mitigation Technique based on the standard Triple Modular Redundancy (S-MTTM-R) to determine the model's vulnerable parts, distinguishing Malfunction and Light-Malfunction errors. A comprehensive vulnerability analysis was performed using a SASSIFI fault injector at the CNN AlexNet and DenseNet201 models: layers, kernels, and instructions to show both models' resilience and identify the most vulnerable portions and harden them by injecting them while implemented on NVIDIA's GPUs. The experimental results showed that S-MTTM-R achieved a significant improvement in error masking. No-Malfunction have been improved from 54.90%, 67.85%, and 59.36% to 62.80%, 82.10%, and 80.76% in the three modes RF, IOA, and IOV, respectively for AlexNet. For DenseNet, No-Malfunction have been improved from 43.70%, 67.70%, and 54.68% to 59.90%, 84.75%, and 83.07% in the three modes RF, IOA, and IOV, respectively. Importantly, S-MTTM-R decreased the percentage of errors that case misclassification (Malfunction) from 3.70% to 0.38% and 5.23% to 0.23%, for AlexNet and DenseNet, respectively. The performance analysis results showed that the S-MTTM-R achieved lower overhead compared to the well-known protection techniques: Algorithm-Based Fault Tolerance (ABFT), Double Modular Redundancy (DMR), and Triple Modular Redundancy (TMR). In light of these results, the study revealed strong evidence that the developed S-MTTM-R was successfully mitigated the soft errors for the DNNs model on GPUs with low-overheads in energy, performance, and area indicated a remarkable improvement in the healthcare domains' model reliability.

TABLE OF CONTENT

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *http://arxiv.org/abs/*, 1603.04467.

Aboagye, E. O., James, G. C., & Kumar, R. (2018). Evaluating the Performance of Deep Neural Networks for Health Decision Making. *Procedia Computer Science*, (131), 866–872. https://doi.org/10.1016/j.procs.2018.04.288

Alemzadeh, H., Iyer, R. K., Kalbarczyk, Z., & Raman, J. (2013). Analysis of safety-critical computer failures in medical devices. *IEEE Security and Privacy*, 11(4), 14–26. https://doi.org/10.1109/MSP.2013.49

Alemzadeh, H., Raman, J., Leveson, N., Kalbarczyk, Z., & Iyer, R. K. (2016a). Adverse events in robotic surgery: A retrospective study of 14 years of FDA data. *PLoS ONE*, 11(4), 1–20. https://doi.org/10.1371/journal.pone.0151470

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Van Esesn, B. C., Awwal, A. A. S., & Asari, V. K. (2018). The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches. *http://arxiv.org/abs*/, 1803.01164.

Amin, M. G., & Erol, B. (2018). Understanding deep neural networks performance for radar-based human motion recognition. In *Proceedings of the 2018 IEEE Radar Conference (RadarConf18)*, (1461–1465). https://www.10.1109/RADAR.2018.8378780

Anarado, I., Anam, M. A., Verdicchio, F., & Andreopoulos, Y. (2016). Mitigating Silent Data Corruptions in integer matrix products: toward reliable multimedia computing on unreliable hardware. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(11), 2476–2489. https://www.10.1109/TCSVT.2016.2589622

Arifeen, T., Hassan, A. S., & Lee, J. A. (2019). A fault tolerant voter for approximate triple modular redundancy. *Electronics (Switzerland*), 8(3). https://doi.org/10.3390/electronics8030332

Azizimazreah, A., Gu, Y., Gu, X., & Chen, L. (2018). Tolerating soft errors in deep learning accelerators with reliable on-chip memory designs. In *Proceedings of 2018 IEEE International Conference on Networking, Architecture and Storage, NAS 2018,* (1–10). https://doi.org/10.1109/NAS.2018.8515692

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *Proceedings of International Conference on Engineering and Technology (ICET)* (pp. 1-6). https://10.1109/ICEngTechnol.2017.8308186

Autran, J. L., & Munteanu, D. (2017). Soft Errors: from particles to circuits. *CRC Press*.

Arifeen, T., Hassan, A. S., & Lee, J. A. (2020). Approximate triple modular redundancy: A survey. *IEEE Access*, 8, 139851-139867. https://10.1109/ACCESS.2020.3012673

Baumann, R. C. (2005). Radiation-induced soft errors in advanced semiconductor technologies. *IEEE Transactions on Device and Materials Reliability*, 5(3), 305–316. https://www.10.1109/TDMR.2005.853449

Bettola, S., & Piuri, V. (1998). High performance fault-tolerant digital neural networks. *IEEE Transactions on Computers*, 47(3), 357–363. https://doi.org/10.1109/12.660173

Botella, G., García, C., & Meyer-Bäse, U. (2013). Hardware implementation of machine vision systems: image and video processing hardware implementation of machine vision systems. *Eurasip Journal on Advances in Signal Processing*, (1), 2–5. https://doi.org/10.1186/1687-6180-2013-152

Braun, C., Halder, S., & Wunderlich, H. J. (2014). A-abft: autonomous algorithm-based fault tolerance for matrix multiplications on graphics processing units. In *Proceedings of 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, (443–454). https://doi.org/10.1109/DSN.2014.48

Breier, J., Hou, X., Jap, D., Ma, L., Bhasin, S., & Liu, Y. (2018). Practical Fault Attack on Deep Neural Networks. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security* (2204-2206). https://doi.org/10.1145/3243734.3278519

Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. *In Neurocomputing* (227–236). https://doi.org/10.1007/978-3-642-76153-9_28

Bryson, E. Strohmaier, J. Dongarra, H. Simon, and H. Meuer. (Nov. 2018). The Top 500 List, [online]. *https:// www.top500.org/lists/2018/11*.

Cai, Z., Fan, Q., Feris, R. S., & Vasconcelos, N. (2016). A unified multi-scale deep convolutional neural network for fast object detection. In *Proceedings of European Conference on Computer Vision*, (354–370). https://doi.org/10.1007/978-3-319-46493-0_22

Chen, Y.-H., Krishna, T., Emer, J. S., & Sze, V. (2016). Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE Journal of Solid-State Circuits*, 52(1), 127–138. https://doi.org/10.1109/JSSC.2016.2616357

Clemente, J. A., Mansour, W., Ayoubi, R., Serrano, F., Mecha, H., Ziade, H., El Falou, W., & Velazco, R. (2016). Hardware implementation of a fault-tolerant Hopfield Neural Network on FPGAs. *Neurocomputing*, (171), 1606–1609.

Codrescu, L., Anderson, W., Venkumanhanti, S., Zeng, M., Plondke, E., Koob, C., ... & Maule, R. (2014). Hexagon DSP: An architecture optimized for mobile multimedia and communications. *IEEE Micro*, 34(2), 34-43. https://www.10.1109/MM.2014.12

Collobert, R., Kavukcuoglu, K., & Farabet, C. (2011). Torch7: A matlab-like environment for machine learning. In *BigLearn*, *NIPS workshop*, 1–6. http://infoscience.epfl.ch/record/192376/files/Collobert_NIPSWORKSHOP_2011.pdf

Cook, J. S., & Gupta, N. (2015). History of Supercomputing and Supercomputer Centers. In *Research and Applications in Global Supercomputing* (pp. 33–55). https://www.10.4018/978-1-4666-7461-5.ch002

Cook, S. (2012). CUDA programming: a developer's guide to parallel computing with GPUs. *Newnes*. http://dx.doi.org/10.1016/B978-0-12-415933-4.00001-6

De Oliveira, Daniel Alfonso Goncalves, Pilla, L. L., Hanzich, M., Fratin, V., Fernandes, F., Lunardi, C., Cela, J. M., Navaux, P. O. A., Carro, L., & Rech, P. (2017). Radiation-induced error criticality in modern HPC parallel accelerators. In *Proceedings of IEEE International Symposium on High Performance Computer Architecture (HPCA)*, (577–588). https://www.10.1109/HPCA.2017.41

De Oliveira, Daniel Alfonso Goncalves, Pilla, L. L., Santini, T., & Rech, P. (2016). Evaluation and mitigation of radiation-induced soft errors in graphics processing units. *IEEE Transactions on Computers*, 65(3), 791–804. https://doi.org/10.1109/TC.2015.2444855

DeBardeleben, N., Blanchard, S., Monroe, L., Romero, P., Grunau, D., Idler, C., & Wright, C. (2013). GPU behavior on a large HPC cluster. In *Proceedings of European Conference on Parallel Processing*, (680–689). https://doi.org/10.1007/978-3-642-54420-0_66

Despres, P., & Jia, X. (2017). A review of GPU-based medical image reconstruction. *Physica Medica*, (42), 76–92. https://doi.org/10.1016/j.ejmp.2017.07.024

Dodd, P. E., & Massengill, L. W. (2003). Basic mechanisms and modeling of single-event upset in digital microelectronics. *IEEE Transactions on Nuclear Science*, 50 III(3), 583–602. https://www.10.1109/TNS.2003.813129

dos Santos, F. F., Draghetti, L., Weigel, L., Carro, L., Navaux, P., & Rech, P. (2017). Evaluation and mitigation of soft-errors in neural network-based object detection in three GPU architectures. In *Proceedings of 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, (169–176). https://doi.org/10.1109/DSN-W.2017.47

dos Santos, F. F., Pimenta, P. F., Lunardi, C., Draghetti, L., Carro, L., Kaeli, D., & Rech, P. (2018). Analyzing and Increasing the Reliability of Convolutional Neural Networks on GPUs. *IEEE Transactions on Reliability*, 68(2), 663–677. https://doi.org/10.1109/TR.2018.2878387

Dunnmon, J. A., Yi, D., Langlotz, C. P., Ré, C., Rubin, D. L., & Lungren, M. P. (2019). Assessment of convolutional neural networks for automated classification of chest radiographs. *Radiology*, 290(2), 537-544. https://doi.org/10.1148/radiol.2018181422

Duzellier, S. (2005). Radiation effects on electronic devices in space. *Aerospace Science and Technology*, 9(1), 93–99. https://doi.org/10.1016/j.ast.2004.08.006

Dodds, N. A. (2012). Single event latchup: hardening strategies, triggering mechanisms, and testing considerations (*Doctoral dissertation*). https://etd.library.vanderbilt.edu/available/etd-11032012-225718/unrestricted/dodds_dissertation_FINAL.pdf

Defour, D., & Petit, E. (2013, July). GPUburn: A system to test and mitigate GPU hardware failures. In *Proceedings of IEEE International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS)*, (pp. 263-270). https://doi.org/10.1007/s11042-017-4699-5

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

Fang, B., Pattabiraman, K., Ripeanu, M., & Gurumurthi, S. (2016). A Systematic Methodology for Evaluating the Error Resilience of GPGPU Applications. *IEEE Transactions on Parallel and Distributed Systems*, 27(12), 3397–3411. https://doi.org/10.1109/TPDS.2016.2517633

Fang, B., Pattabiraman, K., Ripeanu, M., & Gurumurthi, S. (2014). GPU-Qin: A methodology for evaluating the error resilience of GPGPU applications. In *Proceedings of IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, (221–230). https://www.10.1109/ISPASS.2014.6844486

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, 161, 1–13. https://doi.org/10.1016/j.cmpb.2018.04.005

Fonseca, A., & Cabral, B. (2017). Prototyping a GPGPU Neural Network for Deep-Learning Big Data Analysis. *Big Data Research*, (8), 50–56. https://doi.org/10.1016/j.bdr.2017.01.005

Fu, J., Zheng, H., & Mei, T. (2017). Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. *In Proceedings of* *the IEEE Conference on Computer Vision and Pattern Recognition*, (4438–4446). https://www.10.1109/CVPR.2017.476.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1(2), 119–130. https://www.doi.org/10.1016/0893-6080(88)90014-7Get

Gainaru, A., Cappello, F., Snir, M., & Kramer, W. (2012). Fault prediction under the microscope: A closer look into HPC systems. In *Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis*, (1-11). https://doi.org/10.1109/SC.2012.57

George, S., Kim, S., Shah, S., Hasler, J., Collins, M., Adil, F., Wunderlich, R., Nease, S., & Ramakrishnan, S. (2016). A Programmable and Configurable Mixed-Mode FPAA SoC. *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems*, 24(6), 2253–2261. https://doi.org/10.1109/TVLSI.2015.2504119

Glaskowsky, P. N. (2009). NVIDIA's Fermi: the first complete GPU computing architecture. *White Paper,* 18. https://www.nvidia.com/content/PDF/fermi_white_papers/P.Glaskowsky_NVIDIA%27s_Fermi-The_First_Complete_GPU_Architecture.pdf

Gomez, A. N., Ren, M., Urtasun, R., & Grosse, R. B. (2017). The reversible residual network: Backpropagation without storing activations. *arXiv preprint arXiv:,* 1707.04585.

Gomez, L. B., Cappello, F., Carro, L., DeBardeleben, N., Fang, B., Gurumurthi, S., Pattabiraman, K., Rech, P., & Reorda, M. S. (2014). GPGPUs: how to combine high computational power with high reliability. In *Proceedings of the Conference on Design, Automation & Test in Europe*, (341). https://www.10.7873/DATE.2014.354

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). *Cambridge: MIT press.* https://doi.org/10.4258/hir.2016.22.4.351

Gottapu, R. D., & Dagli, C. H. (2018). DenseNet for anatomical brain segmentation. *Procedia Computer Science*, (140), 179–185. https://doi.org/10.1016/j.procs.2018.10.327

Grewal, M., Srivastava, M. M., Kumar, P., & Varadarajan, S. (2018). RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans. In *Proceedings of International Symposium on Biomedical Imaging*, (281–284). https://doi.org/10.1109/ISBI.2018.8363574

GTX, N. G. (2014). 750 Ti White Paper,". NVIDIA Corporation," Online at: Http://International. Download. Nvidia. Com/Geforce-Com/International/Pdfs/GeForce-GTX-750-Ti-Whitepaper. Pdf.

Gu, J., Liu, H., Zhou, Y., & Wang, X. (2017). DeepProf: Performance Analysis for Deep Learning Applications via Mining GPU Execution Patterns. *http://arxiv.org*/abs/, 1707.03750

Guo, Y., Wu, H., Chai, W., Ma, J., & Zhou, G. (2017). Integrity checking based soft error recovery method for DSP. In *Proceedings of Prognostics and System Health Management Conference, PHM-Chengdu*, (1–4). https://doi.org/10.1109/PHM.2016.7819772

Gupta, S., Agrawal, A., Gopalakrishnan, K., & Narayanan, P. (2015). Deep learning with limited numerical precision. In *Proceedings of International Conference on Machine Learning, ICML 2015*, (1737–1746). https://www.arXiv:1502.02551

Hari, S. K. S., Tsai, T., Stephenson, M., Keckler, S. W., & Emer, J. (2015). SASSIFI : Evaluating Resilience of GPU Applications. In *Proceedings of IEEE Workshop on Silicon Errors in Logic – System Effects*, (90–95). https://www.10.1109/ISPASS.2017.7975296.

Hari, S. K. S., Tsai, T., Stephenson, M., Keckler, S. W., & Emer, J. (2017). SASSIFI: An architecture-level fault injection tool for GPU application resilience evaluation. In *Proceedings of ISPASS 2017 - IEEE International Symposium on Performance Analysis of Systems and Software*, 1, (249–258). https://doi.org/10.1109/ISPASS.2017.7975296

Haykin, S. (1998). Neural Networks: A Comprehensive Foundation Upper. *Saddle River. NJ, USA*,

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

Hsueh, M.-C., Tsai, T. K., & Iyer, R. K. (1997). Fault injection techniques and tools. *Computer*, 30(4), 75–82. https://www.10.1109/2.585157

Huang, G., Liu, Z., Pleiss, G., Van Der Maaten, L., & Weinberger, K. (2019). Convolutional Networks with Dense Connectivity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/tpami.2019.2918284

Huang, Z., Zhu, X., Ding, M., & Zhang, X. (2020). Medical Image Classification Using a Light-Weighted Hybrid Neural Network Based on PCANet and DenseNet. *IEEE Access*, (8), 24697–24712. https://doi.org/10.1109/ACCESS.2020.2971225

Huynh, T. V. (2017). Deep neural network accelerator based on FPGA. In *Proceedings of 2017 4th NAFOSTED Conference on Information and Computer Science*, (254–257). https://www.10.1109/NAFOSTED.2017.8108073

Ibrahim, Y., Wang, H., Bai, M., Liu, Z., Wang, J., Yang, Z., & Chen, Z. (2020). Soft Error Resilience of Deep Residual Networks for Object Recognition. *IEEE Access*, (8), 19490–19503. https://doi.org/10.1109/ACCESS.2020.2968129

Ignatov, A., Timofte, R., Chou, W., Wang, K., Wu, M., Hartley, T., & Van Gool, L. (2018). Ai benchmark: Running deep neural networks on android smartphones. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops* (0-0). http://ai-benchmark.com

Islam, M., Atputharuban, D. A., Ramesh, R., & Ren, H. (2019). Real-Time Instrument Segmentation in Robotic. *IEEE Robotics and Automation Letters*, 4(2), 2188–2195. https://doi.org/10.1109/LRA.2019.2900854

Jafari-Nodoushan, M., Miremadi, S. G., & Ejlali, A. (2008). Control-flow checking using branch instructions. In *Proceedings of the 5th International Conference on Embedded and Ubiquitous Computing, EUC 2008*, (1), 66–72. https://doi.org/10.1109/EUC.2008.44

Jagannathan, S., Mody, M., & Mathew, M. (2016). Optimizing convolutional neural network on DSP. In *Proceedings of 2016 IEEE International Conference on Consumer Electronics, ICCE 2016*, (371–372). https://doi.org/10.1109/ICCE.2016.7430652

Janney, S. S., & Chakravarty, S. (2019). Deep learning in medical and surgical instruments. *In Bioelectronics and Medical Devices*. *Elsevier Ltd*. 00040-6. https://doi.org/10.1016/b978-0-08-102420-1.00040-6

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 2014 ACM Conference on Multimedia*, (675–678). https://doi.org/10.1145/2647868.2654889

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Jouppi, N. (2019). Google supercharges machine learning tasks with TPU custom chip. 18 May 2016

Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., Bates, S., Bhatia, S., Boden, N., & Borchers, A. (2017). In-datacenter performance analysis of a tensor processing unit. In *Proceedings of 2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*, (1–12). https://doi.org/10.1145/3079856.3080246

Kalaiselvi, T., Sriramakrishnan, P., & Somasundaram, K. (2017). Survey of using GPU CUDA programming model in medical image analysis. *Informatics in Medicine Unlocked*, (9), 133–144. https://doi.org/10.1016/j.imu.2017.08.001

Kastensmidt, F., & Rech, P. (2016). Radiation effects and fault tolerance techniques for FPGAs and GPUs. *In FPGAs and Parallel Architectures for Aerospace Applications* (pp. 3–17). https://doi.org/10.1007/978-3-319-14352-1_1

Kazanzides, P. (2009, September). Safety design for medical robots. In *Proceedings of 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 7208-7211). https://doi.org/10.1109/IEMBS.2009.5335275

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2019). A Survey of the Recent Architectures of Deep Convolutional Neural Networks. 1–62. *http://arxiv.org/abs*/, 1901.06032

Ko, J. H., Mudassar, B., Na, T., & Mukhopadhyay, S. (2017). Design of an Energy-Efficient Accelerator for Training of Convolutional Neural Networks using Frequency-Domain Computation. In *Proceedings of Design Automation Conference*, (Part 12828). https://doi.org/10.1145/3061639.3062228

Ko, Y., Jeyapaul, R., Kim, Y., Lee, K., & Shrivastava, A. (2017). Protecting Caches from Soft Errors: A Microarchitect's Perspective. *ACM Transactions on Embedded Computing Systems (TECS)*, 16(4), 93. https://doi.org/10.1145/3063180

Koga, R., Penzin, S. H., Crawford, K. B., & Crain, W. R. (1998). Single event functional interrupt (SEFI) sensitivity in microcircuits. In *Proceedings of the European Conference on Radiation and Its Effects on Components and Systems* (311–318). https://doi.org/10.1109/radecs.1997.698915

Koo, Y., You, C., & Kim, S. (2018). OpenCL-Darknet: An OpenCL Implementation for Object Detection. In *Proceedings of 2018 IEEE International Conference on Big Data and Smart Computing*, (631–634). https://doi.org/10.1109/BigComp.2018.00112

Kovalev, V., Kalinovsky, A., & Kovalev, S. (2016). Deep learning with theano, torch, caffe, tensorflow, and deeplearning4j: Which one is the best in speed and accuracy?. In *Proceedings of International Conference on Pattern Recognition and Information Processing*, (99–103). http://imlab.grid.by/

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. https://doi.org/10.1145/3065386

Kuznetsov, E., & Stegailov, V. (2019). Porting CUDA-Based Molecular Dynamics Algorithms to AMD ROCm Platform Using HIP Framework: Performance Analysis. *Russian Supercomputing Days*, 121–130. https://doi.org/10.1007/978-3-030-36592-9_11

Kwon, H., Samajdar, A., & Krishna, T. (2018). Maeri: Enabling flexible dataflow mapping over DNN accelerators via reconfigurable interconnects. *ACM SIGPLAN Notices*, 53(2), 461-475. https://doi.org/10.1145/3296957.3173176

Kwon, S. J. (2011). Artificial neural networks. *Artificial Neural Networks*, 1–426.

Kundu, P., Chopra, S., & Lad, B. K. (2019). Multiple failure behaviors identification and remaining useful life prediction of ball bearings. *Journal of Intelligent Manufacturing*, 30(4), 1795-1807. https://doi.org/10.1007/s10845-017-1357-8

Lai, J., Chen, Y., Han, B., Ji, L., Shi, Y., Huang, Z., ... & Feng, Q. (2019). A DenseNet-based diagnosis algorithm for automated diagnosis using clinical ECG data. *Nan Fang yi ke da xue xue bao= Journal of Southern Medical University*, 39(1), 69-75. https://www.10.12122/j.issn.1673-4254.2019.01.11

Lane, N. D., Bhattacharya, S., Georgiev, P., Forlivesi, C., Jiao, L., Qendro, L., & Kawsar, F. (2016, April). Deepx: A software accelerator for low-power deep learning inference on mobile devices. In *Proceedings of 2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)* (pp. 1-12). https://doi.org/10.1109/IPSN.2016.7460664

Lee, M., Hwang, K., & Sung, W. (2014, May). Fault tolerance analysis of digital feed-forward deep neural networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* (5031-5035). https://www.10.1109/ICASSP.2014.6854560

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361(10), 1995. https://www.researchgate.net/profile/Yann_Lecun/publication/2453996_Convolutional_Networks_for_Images_Speech_and_Time-Series/links/0deec519dfa2325502000000.pdf

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.

Li, G., Hari, S. K. S., Sullivan, M., Tsai, T., Pattabiraman, K., Emer, J., & Keckler, S. W. (2017a). Understanding error propagation in Deep Learning Neural Network (DNN) accelerators and applications. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2017*. https://doi.org/10.1145/3126908.3126964

Li, G., Pattabiraman, K., Cher, C. Y., & Bose, P. (2016). Understanding Error Propagation in GPGPU Applications. In *Proceedings of International Conference for High Performance Computing, Networking, Storage and Analysis, SC, November*, (240–251). https://doi.org/10.1109/SC.2016.20

Li, Z., Wang, Y., Zhi, T., & Chen, T. (2017). A survey of neural network accelerators. *Frontiers of Computer Science*, 11(5), 746–761. https://doi.org/10.1007/s11704-016-6159-1

Liang, Z., Zhang, G., Huang, J. X., & Hu, Q. V. (2014). Deep learning for healthcare decision making with EMRs. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* (556–559). https://www.10.1109/BIBM.2014.6999219

Libano, F., Wilson, B., Anderson, J., Wirthlin, M. J., Cazzaniga, C., Frost, C., & Rech, P. (2019). Selective hardening for neural networks in FPGAs. *IEEE Transactions on Nuclear Science*, 66(1), 216–222. https://doi.org/10.1109/TNS.2018.2884460

Lindholm, E., Nickolls, J., Oberman, S., & Montrym, J. (2008). Nvidia Tesla: a unified graphics and computing architecture to enable flexible, programmable graphics and high-performance computing. *IEEE Micro*, 39–55. https://www.10.1109/MM.2008.31

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, (234), 11–26. https://doi.org/10.1016/j.neucom.2016.12.038

Liu, Y., Wei, L., Luo, B., & Xu, Q. (2017). Fault injection attack on deep neural network. In *Proceedings of IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, ICCAD, 2017-Novem*, (131–138). https://doi.org/10.1109/ICCAD.2017.8203770

Lopes, I. C., Benevenuti, F., Kastensmidt, F. L., Susin, A. A., & Rech, P. (2018). Reliability analysis on case-study traffic sign convolutional neural network on APSoC. In *Proceedings of IEEE 19th Latin-American Test Symposium, LATS 2018*, (1–6). https://doi.org/10.1109/ICCAD.2017.8203770

Lopes, I. C., Kastensmidt, F. L., & Susin, A. A. (2017). SEU susceptibility analysis of a feed forward neural network implemented in a SRAM-based FPGA. LATS 2017 - In *Proceedings of 18th IEEE Latin-American Test Symposium*, (1–6). https://www.10.1109/LATW.2017.7906770

Lopes, I. D. C. (2017). Convolutional neural network reliability on an APSoC platform a traffic-sign recognition case study. *(Master dissertation).* http://hdl.handle.net/10183/171094

Lu, L., & Liang, Y. (2018). SpWA: An efficient sparse winograd convolutional neural networks accelerator on FPGAs. In *Proceedings of Design Automation Conference*, (Part F1377). https://doi.org/10.1145/3195970.3196120

Lu, S., Lu, Z., & Zhang, Y. (2019). Pathological brain detection based on AlexNet and transfer learning. *Journal of Computational Science,* (30), 41–47. https://doi.org/10.1016/j.jocs.2018.11.008

Lunardi, C., Previlon, F., Kaeli, D., & Rech, P. (2018). On the Efficacy of ECC and the Benefits of FinFET Transistor Layout for GPU Reliability. *IEEE Transactions on Nuclear Science*, 65(8), 1843–1850. https://www.10.1109/TNS.2018.2823786

Lyons, R. E., & Vanderkulk, W. (1962). The use of triple-modular redundancy to improve computer reliability. *IBM journal of research and development*, 6(2), 200-209. https://www.10.1147/rd.62.0200

Mahatme, N. N., Jagannathan, S., Loveless, T. D., Massengill, L. W., Bhuva, B. L., Wen, S.-J., & Wong, R. (2011). Comparison of combinational and sequential error rates for a deep submicron process. *IEEE Transactions on Nuclear Science*, 58(6), 2719–2725. https://www.10.1109/TNS.2011.2171993

Marques, J., Andrade, J., & Falcao, G. (2017, October). Unreliable memory operation on a convolutional neural network processor. In *Proceedings of 2017 IEEE International Workshop on Signal Processing Systems (SiPS)* (pp. 1-6). https://doi.org/10.1109/SiPS.2017.8110024

Mayer, D. C., Koga, R., & Womack, J. M. (2007). The impact of radiation-induced failure mechanisms in electronic components on system reliability. *IEEE Transactions on Nuclear Science*, 54(6), 2120–2124. https://doi.org/10.1109/TNS.2007.910294

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of 2016 Fourth International Conference on 3D Vision (3DV),* (565–571). https://www.10.1109/3DV.2016.79

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2017). Deep learning for healthcare: Review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6), 1236–1246. https://doi.org/10.1093/bib/bbx044

Mittal, S. (2020). A survey of FPGA-based accelerators for convolutional neural networks. *Neural computing and applications*, *32*(4), 1109-1139. https://doi.org/10.1007/s00521-018-3761-1

Mittal, S. (2019). A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform. *Journal of Systems Architecture*, 97(January), 428–442. https://doi.org/10.1016/j.sysarc.2019.01.011

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., ... & Wu, H. (2017). Mixed precision training. *arXiv preprint arXiv:*1710.03740.

Narasimham, B., Bhuva, B. L., Schrimpf, R. D., Massengill, L. W., Gadlage, M. J., Amusan, O. A., Holman, W. T., Witulski, A. F., Robinson, W. H., Black, J. D., Benedetto, J. M., & Eaton, P. H. (2007). Characterization of digital single event transient pulse-widths in 130-nm and 90-nm CMOS technologies. *IEEE Transactions on Nuclear Science*, 54(6), 2506–2511. https://doi.org/10.1109/TNS.2007.910125

Nazemi, M., Pasandi, G., & Pedram, M. (2018). NullaNet: Training deep neural networks for reduced-memory-access inference. *ArXiv Preprint ArXiv*:1807.08716.

Nidhin, T. S., Bhattacharyya, A., Behera, R. P., Jayanthi, T., & Velusamy, K. (2017). Understanding radiation effects in SRAM-based field programmable gate arrays for implementing instrumentation and control systems of nuclear power plants. *Nuclear Engineering and Technology*, 49(8), 1589–1599. https://doi.org/10.1016/j.net.2017.09.002

Nielsen, M. A. (2015). Neural networks and deep learning. *San Francisco, CA: Determination press*.

Noh, J., Correas, V., Lee, S., Jeon, J., Nofal, I., Cerba, J., ... & Kwon, S. (2015). Study of neutron soft error rate (SER) sensitivity: investigation of upset mechanisms by comparative simulation of FinFET and planar MOSFET SRAMs. *IEEE Transactions on Nuclear Science*, 62(4), 1642-1649. https://www.10.1109/TNS.2015.2450997

Nuño-Maganda, M., & Torres-Huitzil, C. (2011). A temporal coding hardware implementation for spiking neural networks. *ACM SIGARCH Computer Architecture News*, 38(4), 2. https://doi.org/10.1145/1926367.1926369

Oh, K. S., & Jung, K. (2004). GPU implementation of neural networks. *Pattern Recognition*, 37(6), 1311-1314. https://doi.org/10.1016/j.patcog.2004.01.013

Oliveira, D., Rech, P., & Navaux, P. O. (2019). Hardening Strategies for HPC Applications. *Anais Estendidos do Simpósio em Sistemas Computacionais de Alto Desempenho*, 112–113. https://doi.org/10.5753/wscad_estendido.2019.8708

Oliveira, Daniel A.G., Rech, P., Quinn, H. M., Fairbanks, T. D., Monroe, L., Michalak, S. E., Anderson-Cook, C., Navaux, P. O. A., & Carro, L. (2014). Modern GPUs radiation sensitivity evaluation and mitigation through duplication with comparison. *IEEE Transactions on Nuclear Science*, 61(6), 3115–3122. https://doi.org/10.1109/TNS.2014.2362014

Petscharnig, S., & Schöffmann, K. (2018). Learning laparoscopic video shot classification for gynecological surgery. *Multimedia Tools and Applications*, *77*(7), 8061-8079. https://doi.org/10.1007/s11042-017-4699-5

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics*, 69, 218–229. https://doi.org/10.1016/j.jbi.2017.04.001

Pigou, L., Dieleman, S., Kindermans, P. J., & Schrauwen, B. (2015). Sign language recognition using convolutional neural networks. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, (8925), 572–578. https://doi.org/10.1007/978-3-319-16178-5_40

Piuri, V. (2001). Analysis of Fault Tolerance in Artificial Neural Networks. *Journal of Parallel and Distributed Computing*, 61(1), 18–48. https://doi.org/10.1006/jpdc.2000.1663

Pourbabaee, B., Roshtkhari, M. J., & Khorasani, K. (2018). Deep Convolutional Neural Networks and Learning ECG Features for Screening Paroxysmal Atrial Fibrillation Patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 48(12), 2095–2104. https://doi.org/10.1109/TSMC.2017.2705582

Previlon, F. G. (2019). Characterization and Remediation for Soft Error Reliability on GPU. Northeastern University. *(Doctoral dissertation)*. https://repository.library.northeastern.edu/files/neu:m044pf65j/fulltext.pdf

Pown, M., & Lakshmi, B. (2020). Investigation of Radiation Hardened TFET SRAM Cell for Mitigation of Single Event Upset. *IEEE Journal of the Electron Devices Society*, 8, 1397-1403. https://www.10.1109/JEDS.2020.3002265

Rajih, E., Tholomier, C., Cormier, B., Samouëlian, V., Warkus, T., Liberman, M., Widmer, H., Lattouf, J. B., Alenizi, A. M., Meskawi, M., Valdivieso, R., Hueber, P. A., Karakewicz, P. I., El-Hakim, A., & Zorn, K. C. (2017). Error reporting from the da Vinci surgical system in robotic surgery: A Canadian multispecialty experience at a single academic centre. *Canadian Urological Association Journal*, 11(5), E197–E202. https://doi.org/10.5489/cuaj.4116

Rampasek, L., & Goldenberg, A. (2016). TensorFlow: biology's gateway to deep learning?. *Cell systems*, 2(1), 12-14. https://doi.org/10.1016/j.cels.2016.01.009

Rech, P., Pilla, L. L., Navaux, P. O. A., & Carro, L. (2014). Impact of GPUs parallelism management on Safety-Critical and HPC applications reliability. In *Proceedings of 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, (455–466). https://doi.org/10.1109/DSN.2014.49

Rech, P., Pilla, L. L., Silvestri, F., Frost, C., Navaux, P. O. A., Reorda, M. S., & Carro, L. (2013). Neutron sensitivity and hardening strategies for Fast Fourier Transform on GPUs. In *Proceedings of the European Conference on Radiation and Its Effects on Components and Systems*, (1–5). https://doi.org/10.1109/RADECS.2013.6937457

Rech, P, Aguiar, C., Ferreira, R., Silvestri, M., Griffoni, A., Frost, C., & Carro, L. (2012). Neutron-induced soft errors in graphic processing units. In *Proceedings of IEEE Radiation Effects Data Workshop*, (1–6). https://doi.org/10.1109/REDW.2012.6353714

Rech, Paolo, Carro, L., Wang, N., Tsai, T., Kumar, S., Hari, S., & Keckler, S. W. (2014). Measuring the Radiation Reliability of SRAM Structures in GPUs Designed for HPC. In *Proceedings of IEEE 10th Workshop on Silicon Errors in Logic - System Effects (SELSE)*. https://www.cs.utexas.edu/users/skeckler/pubs/SELSE_2014_Reliability.pdf

Rech, Paolo, Fairbanks, T. D., Quinn, H. M., & Carro, L. (2013). Threads distribution effects on graphics processing units neutron sensitivity. *IEEE Transactions on Nuclear Science*, 60(6), 4220–4225. https://doi.org/10.1109/TNS.2013.2286970

Redmon, "Darknet: Open Source Neural Networks in C," [online]. Available: http://pjreddie.com/darknet/, 2016.

Restrepo-Calle, F., Martínez-Álvarez, A., Cuenca-Asensi, S., & Jimeno-Morenilla, A. (2013). Selective swift-r. *Journal of Electronic Testing*, 29(6), 825–838. https://doi.org/10.1007/s10836-013-5416-6

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Santos, F. F. D., & Rech, P. (2017, November). Analyzing the criticality of transient faults-induced SDCS on GPU applications. In *Proceedings of the 8th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems* (pp. 1-7). https://doi.org/10.1145/3148226.3148228

Schirmeier, H. B. (2016). Efficient fault-injection-based assessment of software-implemented hardware fault tolerance, *(Doctoral dissertation)*. https://eldorado.tu-dortmund.de/bitstream/2003/35175/1/Dissertation.pdf

Seifert, N. (2010). Radiation-induced soft errors: A chip-level modeling perspective. *Foundations and Trends® in Electronic Design Automation*, 4(2–3), 99–221. https://doi.org/10.1561/1000000018

Sheikh, A. T., & El-Maleh, A. H. (2018). Double Modular Redundancy (DMR) Based Fault Tolerance Technique for Combinational Circuits. *Journal of Circuits, Systems and Computers*, 27(6), 1–17. https://doi.org/10.1142/S0218126618500974

Shi, Q., Omar, H., & Khan, O. (2017). Exploiting the Tradeoff between Program Accuracy and Soft-error Resiliency Overhead for Machine Learning Workloads. *http://arxiv.org/abs/*,1707.02589.

Shortliffe, E. H. (1977, October). Mycin: A knowledge-based computer program applied to infectious diseases. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, (p. 66). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2464549/

Shvets, A. A., Rakhlin, A., Kalinin, A. A., & Iglovikov, V. I. (2018). Automatic instrument segmentation in robot-assisted surgery using deep learning. In *Proceedings of 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, (624–628). https://www.10.1109/ICMLA.2018.00100

Siingh, D., & Singh, R. P. (2010). The role of cosmic rays in the Earth's atmospheric processes. *Pramana*, 74(1), 153–168. https://doi.org/10.1007/s12043-010-0017-8

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, (1–14). https://arxiv.org/abs/1409.1556

Stephenson, M., Sastry Hari, S. K., Lee, Y., Ebrahimi, E., Johnson, D. R., Nellans, D., O'Connor, M., & Keckler, S. W. (2015). Flexible software profiling of GPU architectures. In *Proceedings of International Symposium on Computer Architecture, 13-17-June*, (185–197). https://doi.org/10.1145/2749469.2750375

Supercompuers, T. (2018). TOP500 Supercompuers. 2018 Nov. https://www.top500.org/lists/2018/11/

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going Deeper with Convolutions Christian. *Journal of Chemical Technology and Biotechnology*, 91(8), 2322–2330. https://doi.org/10.1002/jctb.4820

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117. https://doi.org/10.1016/j.neunet.2014.09.003

Schwing, A. G., & Urtasun, R. (2015). Fully connected deep structured networks. *arXiv preprint arXiv:*,1503.02351.

T, D., & T, L. (2019). Error correction for soft errors. *Journal of Electrical & Electronic Systems*, 07(04). https://doi.org/10.4172/2332-0796.1000276

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (1701–1708). https://doi.org/10.1109/CVPR.2014.220.

Tang, J., Yuan, F., Shen, X., Wang, Z., Rao, M., He, Y., ... & Wu, H. (2019). Bridging biological and artificial neural networks with emerging neuromorphic devices: fundamentals, progress, and challenges. *Advanced Materials*, 31(49), 1902761. https://doi.org/10.1002/adma.201902761

Torres-Huitzil, C., & Girau, B. (2017). Fault and error tolerance in neural networks: A review. *IEEE Access*, 5, 17322–17341. https://10.1109/ACCESS.2017.2742698

Twomey, N. J. (2013). Digital signal processing and artificial intelligence for the automated classification of food allergy *(Doctoral dissertation, University College Cork).* https://cora.ucc.ie/handle/10468/1236

Tselonis, S., & Gizopoulos, D. (2016). GUFI: A framework for GPUs reliability assessment. ISPASS 2016 - In *Proceedings of International Symposium on Performance Analysis of Systems and Software*, (90–100). https://doi.org/10.1109/ISPASS.2016.7482077

Teifel, J. (2008). Self-voting dual-modular-redundancy circuits for single-event-transient mitigation. *IEEE Transactions on Nuclear Science*, 55(6), 3435-3439. https://10.1109/TNS.2008.2005583

Vankeirsbilck, J., & Hallez, H. (2015). Integration of Soft Errors in Functional Safety: a conceptual study. *Annual Journal of Electronics*, *9(September)*, 108–111. https://lirias.kuleuven.be/handle/123456789/508481%0Ahttps://www.researchgate.net/profile/Jens_Vankeirsbilck/publication/281967937_Integration_of_Soft_Errors_in_Functional_Safety_a_conceptual_study/links/55ffe39c08ae07629e51e1fe.pdf

Vedaldi, A., & Zisserman, A. (2016). Vgg convolutional neural networks practical. *Dep. Eng. Sci. Univ. Oxford*, 66. https://www.academia.edu/download/49291782/CNN_practical.pdf

Velazco, R., Cheynet, P., Tissot, A., Haussy, J., Lambert, J., & Ecoffet, R. (1999). Evidences of SEU tolerance for digital implementations of Artificial Neural Networks: one year MPTB flight results. In *Proceedings of 1999 Fifth European Conference on Radiation and Its Effects on Components and Systems. RADECS 99 (Cat. No. 99TH8471)*, (565–568). https://10.1109/RADECS.1999.858648

Vidya, C. S., & Vijaya Kumar, B. P. (2016). Reliability Analysis in Healthcare Imaging Applications. *Indian Journal of Science and Technology*, 9(34). https://doi.org/10.17485/ijst/2016/v9i34/100988

Wang, S., Raju, A., & Huang, J. (2017, April). Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos. In *Proceedings of 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, (pp. 620-623). https://doi.org/10.1109/ISBI.2017.7950597

Wang, Z., & Fey, A. M. (2018). Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 13(12), 1959–1970. https://doi.org/10.1007/s11548-018-1860-1

Weigel, L., Fernandes, F., Navaux, P., & Rech, P. (2017). Kernel vulnerability factor and efficient hardening for histogram of oriented gradients. In *Proceedings of IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*, (1–6). https://10.1109/DFT.2017.8244439

Watkins, A., & Tragoudas, S. (2020). Radiation Hardened Latch Designs for Double and Triple Node Upsets. *IEEE Transactions on Emerging Topics in Computing*, *8*(3), 616–626. https://10.1109/TETC.2017.2776285

Xing, K. (2018). Training for'Unstable'CNN Accelerator: A Case Study on FPGA. *arXiv preprint arXiv:*1812.01689. https://arxiv.org/abs/1812.01689

Yang, Y., Luo, H., Xu, H., & Wu, F. (2016). Towards Real-Time Traffic Sign Detection and Classification. *IEEE Transactions on Intelligent Transportation Systems*, 17(7), 2022–2031. https://doi.org/10.1109/TITS.2015.2482461

Yu, D., & Deng, L. (2010). Deep learning and its applications to signal and information processing. *IEEE Signal Processing Magazine*, 28(1), 145–154. https://10.1109/MSP.2010.939038

Zhang, Jeff, Gu, T., Basu, K., & Garg, S. (2018). Analyzing and mitigating the impact of permanent faults on a systolic array based neural network accelerator. In *Proceedings of the IEEE VLSI Test Symposium, 2018-April(Ml),* (1–6). https://doi.org/10.1109/VTS.2018.8368656

Zhang, Jianpeng, Xie, Y., Wu, Q., & Xia, Y. (2019). Medical image classification using synergic deep learning. *Medical Image Analysis*, 54, 10–19. https://doi.org/10.1016/j.media.2019.02.010

Zhao, R., Luk, W., Niu, X., Shi, H., & Wang, H. (2017). Hardware Acceleration for Machine Learning. In *Proceedings of IEEE Computer Society Annual Symposium on VLSI, ISVLSI, 2017-July*, (645–650). https://doi.org/10.1109/ISVLSI.2017.127

Zhou, Y., & Jiang, J. (2016). An FPGA-based accelerator implementation for deep convolutional neural networks. In *Proceedings of 2015 4th International Conference on Computer Science and Network Technology*, (829–832). https://doi.org/10.1109/ICCSNT.2015.7490869

Ziade, H., Ayoubi, R., & Velazco, R. (2004). A survey on fault injection techniques. *Int. Arab J. Inf. Technol.*, 1(2), 171–186. http://www.citemaster.net/get/f18d7caa-7207-11e3-b274-00163e009cc7/04-Hissam.pdf

Zhou, X., Tang, Y., Jia, Y., Hu, D., Wu, Y., Xia, T., ... & Pang, H. (2019). Single-event effects in SiC double-trench MOSFETs. *IEEE Transactions on Nuclear Science*, 66(11), 2312-2318. https://10.1109/TNS.2019.2944944

You, Z., & Wei, S. (2018). White Paper on AI Chip Technologies. In *Beijing Inno. Cen. Fut. Chips (ICFC)*.