**LITERATURE REVIEW**

Họ tên: Trần Văn Duy MSSV: 18520675

Lê Phước Nhật Nam 18521122

1. **Tổng quan**

Ngày nay với sư phát triển của trí tuệ nhân tạo đã đem lại ngày căng nhiều lợi ích cho cuộc sống bằng các ứng dụng thực tiễn. Các ứng dụng của trí tuệ nhân tạo trải dài trên nhiều linh vực như xử lí ảnh [8], giao tiếp tín hiệu không dây [9], xử lí ngôn ngữ tự nhiên [10], … Với việc có rất nhiều ứng dụng và khả năng ứng dụng vào thực tế cao như vậy nên quá trình phát triển của trí tuệ nhân tạo ngày căng được thúc đẩy nhanh hơn.

Song song với việc phát triển nhanh đó, lại phát sinh thêm những vẫn đề liên quan. Cụ thể, khi thực hiện một ứng dụng của trí tuệ nhân tạo thì chúng ta luôn cần một hệ thống đủ lớn để xử lí những phép toán phức tạp và số lượng các phép toán lên tới hàng triệu phép tính. Và cũng cần đủ lượng bộ nhớ để lưu trữ hàng triệu thông số với nguồn năng lượng đủ để duy trì hệ thống hoạt động.

Một điều nữa, không phải lúc nào chúng ta cũng có thể kết nối tới một server đủ lớn để duy trì cho hệ thống trí tuệ nhân tạo của chúng ta hoạt động ổn định và hiệu quả. Chính từ những điều đó, những thiết bị phần cứng có thể đáp ứng được những yêu cầu của một hệ thống trí tuệ nhân tạo thay cho server ra đời và kéo theo rất nhiều bài công trình nghiên cứu về điều này như [1], [2], [3], [4], [5], ... Nhưng vẫn còn một điều bất cập đó là về bộ nhớ và khả năng tính toán trong các thiết bị phần cứng vẫn là một điều khiến cho việc triển khai các hệ thống trí tuệ nhân tạo lên một thiết bị phần cứng gặp nhiều khó khăn. Khó khăn đó là do giới hạn về phần cứng của nhiều thiết bị không đủ để có thể mang trên mình một hệ thống trí tuệ nhân tạo.

Và từ chính những thách thức đó, qua nhiều công trình nghiên cứu của các nhà khoa học. Chúng ta đã tìm được cách khắc phục về giới hạn phần cứng bằng những phương pháp đặc biệt như pruning, quantization, huffman-coding trong [1], [6], [7], … đã đem lại những kết quả vô cùng tốt về việc hệ thống trí tuệ nhân tạo được nén lại để loại bỏ một phần giới hạn về phần cứng nhưng vẫn đảm bảo độ chính xác khi hoạt động ngoài thực tế. Với một mô hình mô hình trí tuệ nhân tạo đã được nén thì chúng ta có thể tiết kiệm được năng lượng tiêu thụ, tăng tốc quá trình xử lí, giảm độ phức tạp khi tính toán, lượng bộ nhớ cần thiết giảm. Đây là một bước tiến lớn trong việc nghiên cứ để tạo ra những thiết bị chuyên dụng cho trí tuệ nhân tạo có khả năng linh hoạt để hoạt động trên nhiều thiết bị khác nhau.

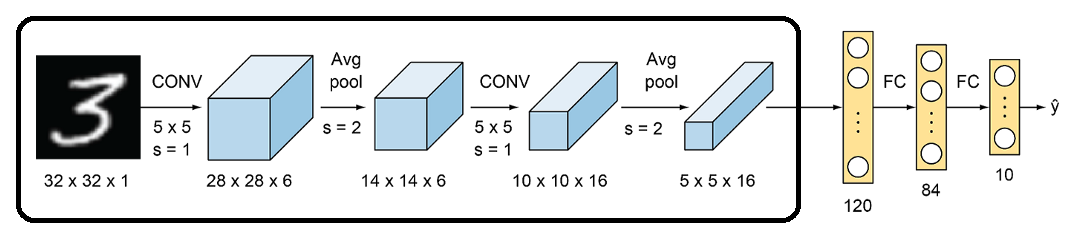
Trong khuôn khổ của project môn học, chỉ có kĩ thuật quantization sẽ được đề cập và nghiên cứu để triển khai.

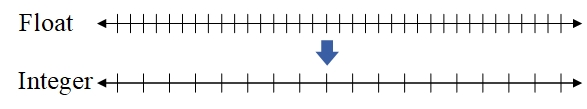
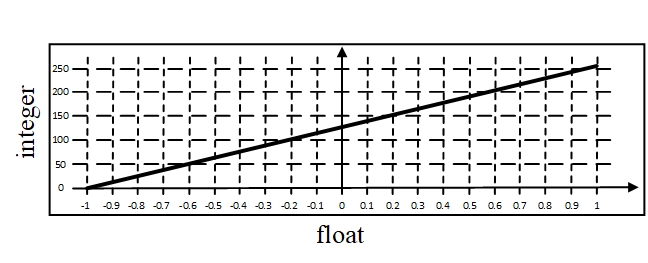
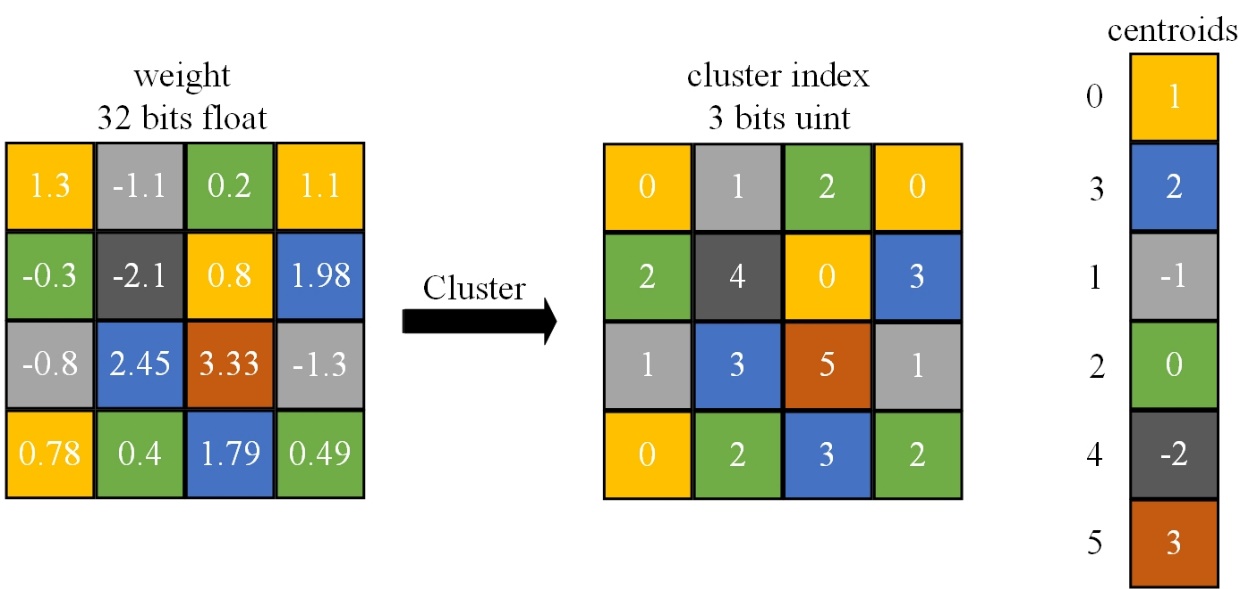
1. **Các công trình liên quan**

Với sự xuất hiện của kĩ thuật quantization, rất nhiều công trình nghiên cứu khoa học đã được ra đời đem lại nhiều kết quả tốt. Kết quả [1] đã kết hợp nhiều kĩ thuật giúp nén mô hình trí tuệ nhân tạo lại trong đó có quantizaion đã đạt được kết quả tốt. Bài báo [2] đã giới thiệu mô hình quantization network và đã áp dụng vô các lĩnh vực như image classification, object detection bằng mô hình Alexnet và Resnet. Với kết quả được đưa ra trong [3] đã chỉ ra được việc có sự khác nhau trong việc sử dụng các mô hình quantization với độ chính xác khác nhau. Với mô hình quantization với bit-width thấp mà vẫn đạt được độ chính xác thì trong [4] đã được đề cập. Và trong [5] đã đề cập đến việc thiết lập thông số cho quantization network bằng Half-wave Gaussian.

1. **Giải pháp được đề xuất và mục tiêu**

Ở đề tài này, nhóm đề xuất xây dựng một mô hình bao gồm các lớp convolution và pooling (không bao gồm phần fully connected) trong mô hình Lenet-5. Mô hình của nhóm sẽ được áp dụng kĩ thuật quantization để hiện thực lại kĩ thuật này trên một môt hình đơn giản trong khuôn khổ Project môn học. Kiến trúc được xây dựng sẽ dựa trên tài nguyên phần cứng được cung cấp bởi board Virtex-7 VC707. Kiến trúc cụ thể sẽ được xây dựng được minh họa trong hình 1 và minh họa cho kĩ thuật quantization sẽ được trinh bày trong hình 2.



Hình 1. Kiến trúc mô hình Lenet-5 đã loại bỏ fully connected

Hình 2. Minh họa cho kĩ thuật quatization

Mục tiêu của đề tài này là có thể thực hiện được việc áp dụng kĩ thuật quantization vào mô hình Lenet-5 mà vẫn giữ được độ chính xác khi hoạt động. Các thông số cuối cùng khi hệ thống hoàn thiện được kì vọng sẽ có thể hoạt động với tần số hoạt động của hệ thống sẽ lớn hơn 100MHz, năng lượng sử dụng sẽ nhỏ hơn 1W và chỉ sử dụng bộ nhớ on-chip.

1. **Phân công nhóm**

|  |  |
| --- | --- |
| Tên thành viên | Trách nhiệm |
| Trần Văn Duy | * Lên ý tưởng project, kế hoạch làm việc. * Kiểm thử độ khả thì của hệ thống bằng python. * Thiết kế cấu trúc phần cứng. * Kiểm thử phần cứng sau khi thiết kế. |
| Lê Phước Nhật Nam | * Lên ý tưởng project, kế hoạch làm việc. * Thiết kế cấu trúc phần cứng. * Kiểm thử phần cứng sau khi thiết kế. |

1. **Tài liệu tham khảo**

[1] Han, Song, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." *arXiv preprint arXiv:1510.00149* (2015).

[2] Yang, Jiwei, et al. "Quantization networks." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

[3] Nayak, Prateeth, David Zhang, and Sek Chai. "Bit efficient quantization for deep neural networks." *arXiv preprint arXiv:1910.04877* (2019).

[4] Wang, Peisong, et al. "Two-step quantization for low-bit neural networks." *Proceedings of the IEEE Conference on computer vision and pattern recognition*. 2018.

[5] Cai, Zhaowei, et al. "Deep learning with low precision by half-wave gaussian quantization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[6] C. Pal, S. Pankaj, W. Akram, A. Acharyya and D. Biswas, "Modified Huffman based compression methodology for Deep Neural Network Implementation on Resource Constrained Mobile Platforms," 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018, pp. 1-5, doi: 10.1109/ISCAS.2018.8351234.

[7] Z. Chiliang, H. Tao, G. Yingda and Y. Zuochang, "Accelerating Convolutional Neural Networks with Dynamic Channel Pruning," 2019 Data Compression Conference (DCC), 2019, pp. 563-563, doi: 10.1109/DCC.2019.00075.

[8] Y. -n. Dong and G. -s. Liang, "Research and Discussion on Image Recognition and Classification Algorithm Based on Deep Learning," 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2019, pp. 274-278, doi: 10.1109/MLBDBI48998.2019.00061.

[9] H. Ye, G. Y. Li and B. Juang, "Power of Deep Learning for Channel Estimation and Signal Detection in OFDM Systems," in IEEE Wireless Communications Letters, vol. 7, no. 1, pp. 114-117, Feb. 2018, doi: 10.1109/LWC.2017.2757490.

[10] D. W. Otter, J. R. Medina and J. K. Kalita, "A Survey of the Usages of Deep Learning for Natural Language Processing," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 2, pp. 604-624, Feb. 2021, doi: 10.1109/TNNLS.2020.2979670.