**Configurations**

We will go through all of the configurations that we apply in this project

**Batch Normalization:** Internal covariate shift slows down and makes it harder to train models. To normalize layers input, we apply Batch Normalization [1]. It acts as a regularizer and helps models to achieve the same accuracy with fewer training steps, hence, saving time as well [2].

**Dropout**: When the data is limited, our model can be overfitted. Dropout technique is introduced to address this problem [3]. However, dropout will increase your training time so in other words, applying dropout also means creating a trade-off between accuracy and training time [4].

**AdamW:** In a conference paper at ICLR 2019 by Loshchilov & Hutter, Adam with L2 regularization had been proved to have worse generalization performance than AdamW [5]. The paper also mentioned that AdamW had yielded better training loss and test error than Adam [6].

**GELUs:** GELUs stands for Gaussian Error Linear Units. It is a high-performing neural network activation function [7]. As the Neural Network goes deeper and deeper while training, sigmoid activation has been proved to be less effective than RELUs (Rectified Linear Units).

**He initialization:** He initialization is an initialization introduced in 2015 [8]. This method will allow us to train deep rectified models from scratch and look into deeper architectures [9]

**References**

[1] [2] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”, Mar. 2015. doi: 10.48550/arXiv.1502.03167.

[3] [4] N. Srivastava, G. Hinton, A. Krizhevsky, L. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting”, vol 15, no. 1, pp. 1929-1958, 2014. [Online]. Available: <https://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf?utm_content=buffer79b43&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer>

[5] [6] I. Loshchilova and F. Hutter, “Decoupled weight decay regularization”, 2017, doi: 10.48550/arXiv.1711.05101

[7] D. Hendrycks and K. Gimpel, “Gaussian Error Linear Units (GELUs)”, 2016, doi: 10.48550/arXiv.1606.08415

[8] [9] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”, 2015, doi: 10.48550/arXiv.1502.01852