Literature Review:

Variational Inference with Normalized Flows

Joshua C. Cubero
SEAS8510-DA3
George Washington University

Background & Context

Research around probability distributions and Bayesian methods has been important for several reasons such as improved decision making when known outcomes are uncertain and improved predictive modeling (Taboga, 2021). Bayesian models enable structured frameworks for updating new assumptions when new data is received, leading to more informed decision making. Additionally, Bayesian models improve predictive power as the dataset increases. There exists a core problem when conducting Bayesian research that arises from the approximation of computing difficult probability densities (Blei, et. al, 2017). Here, the probability density's size and complexity make Bayesian approximations difficult to achieve and thus practitioners must employ an approximation method that enables them to approximate a smaller probability density that maps to the much larger, more complex probability density. In this literature review, we study research on *Variational Inference with Normalizing Flows.*

Variational inference (VI) is a machine learning technique that employs optimization to approximate probability densities (Blei, et. al, 2017). According to Rezende and Mohamed, researchers have had some success employing VI, but which posterior approximation to choose has limited VI's effectiveness. VI approximates a posterior distribution from a known distribution, but no solution had been able to truly reflect the base probability distribution (Rezende & Mohamed, 2015). Research has shown that truer posterior approximations result in marked performance improvements. Rezende and Mohamed present a new method for approximate posterior distribution by normalizing flows, which transforms the probability distribution via a set of invertible mappings. Their work culminates in demonstrating the VI with normalized flows outperforms competing posterior approximation techniques.

Proposed Solution

Rezende and Mohamed achieve VI with normalized flows via the amalgamation three techniques: Amortized Variational Inference, Normalizing Flows, and Inference with Normalizing Flows. Amortized variational inference involved the use of stochastic back propagation, inference networks, and deep latent Gaussian models. Next, normalized flows involve transforming a known probability distribution through a series of invertible mappings. Normalized flows move the base density model through the series of invertible mappings to arrive at the truest probability distribution. Normalized flows employ finite flows and infinitesimal flows. Lastly, the researchers performed inference with normalized flows which enabled scalable inference, which typically have a time complexity $O(LD^3)$. Rezende and Mohamed recognized this inefficiency and developed an algorithm that enabled invertible transformations in linear time. The final algorithm boasted $O(LN^2) + O(KD)$, making VI with Normalized Flows quadratic time complexity, worst case (Rezende & Mohamed, 2015).

Results

The research team compared the normalized flow model to the true distribution and the distribution approximation resulting from a Non-linear Independent Components Estimation (NICE) model on a set of unnormalized 2D probability densities. The team conducted three approximations with K = {2, 8, 32} where the posterior approximation with normalized flows demonstrated greatest convergence at K = 32. The NICE model had less performance with only marginal convergence compared the approximation with normalized flows. Next, Rezende and Mohamed compare Deep Latent Gaussian Model (DLGM) + NICE to DLGM + Normalized Flows using the MNIST dataset. The researchers achieved reduced KL-divergence between approximate posterior and true posterior distribution when employing DLGM + NF. Lastly, the team conducted experiments using the CIFAR=10 images dataset and were once again able to demonstrate that lengthened K, associated with normalized flows, produced improved posterior approximations (Rezende & Mohamed, 2015).

Conclusion

Rezende and Mohamed developed a simple method for learning non-Gaussian posterior densities by learning and mapping fewer complex densities to more complex densities. By combining an amortized VI with Monte Carlo gradient estimation, the team achieved greater improvements when compared to competing approximations. Lastly, Rezende and Mohamed acknowledged these results are only supported by intuition, reasoning, and empirical results. Thus, *Variational Inference with Normalizing Flows* should be further researched, formalized with rigorous research and mathematical proofs, and supported by detailed theoretical analysis.

References

Taboga, M. (2021). *Bayesian inference | Introduction with explained examples*. Www.statlect.com. https://www.statlect.com/fundamentals-of-statistics/Bayesian-inference

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

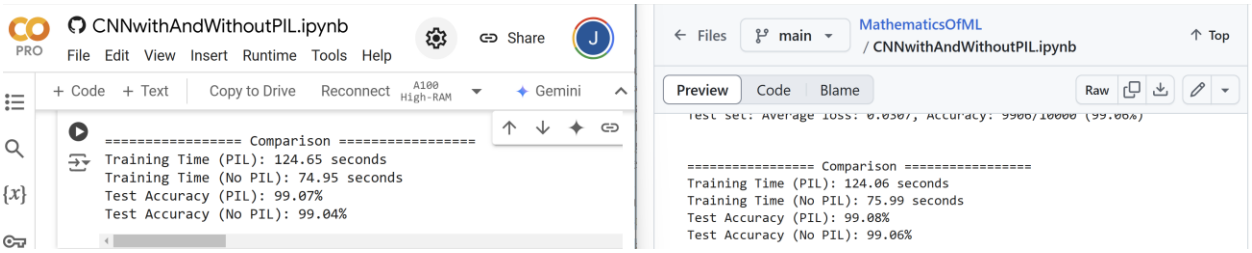
Rezende, D., & Mohamed, S. (2015). Variational Inference with Normalizing Flows. *International Conference on Machine Learning*, 1530–1538.

ConvNets, K-Means Clustering, and Vanishing Gradients: Comparing GitHub to Local Execution

Joshua C. Cubero
SEAS8510-DA3
George Washington University

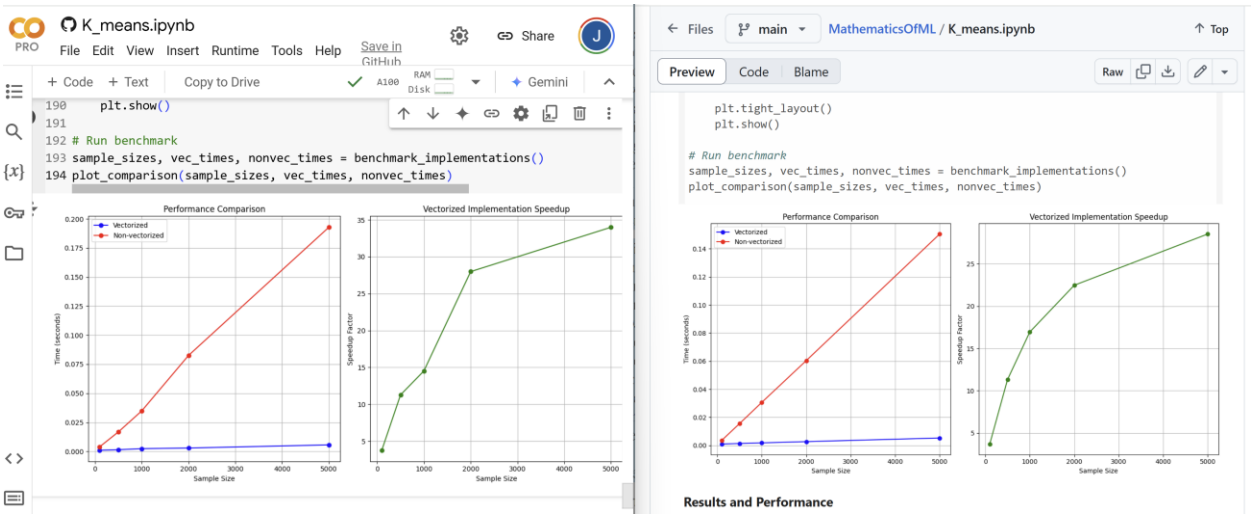
CNN With and Without PIL

The first notebook analyzed was the CNN With and Without PIL, in which PIL or Python Image Library. The null hypothesis here might state that employing the CNN method on the MNIST dataset, where two models are evaluated, one using PIL and the other not using PIL, the model not using PIL would outperform the model using PIL. Executing the code from the GitHub repository returned nearly identical scores for training time and test accuracy.



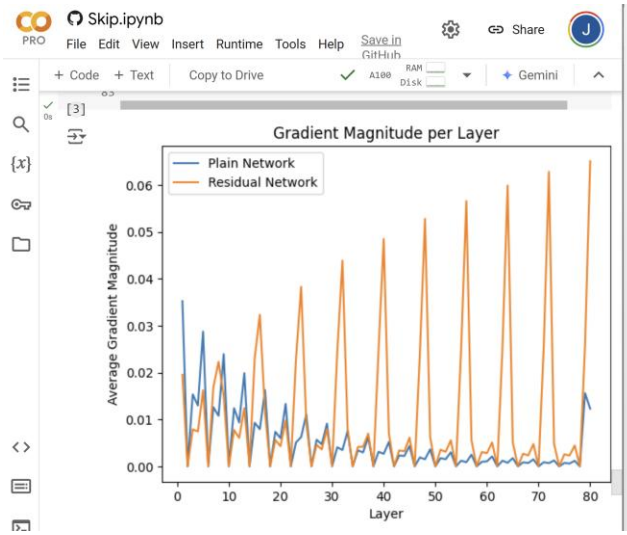
K-Means: Vectorized vs Non-Vectorized

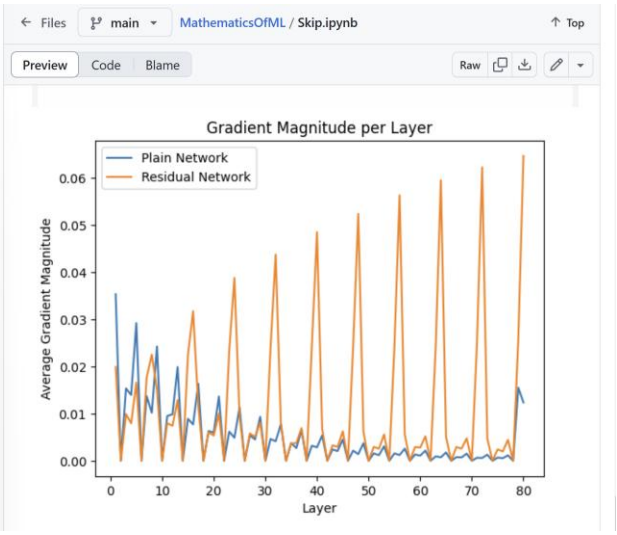
The K-Means vectorized vs non-vectorized experiment focused on the difference between employing an $O(n^4)$ non-vectorized solution against an $O(n^2)$ vectorized solution. Again, the null hypothesis here would suggest that the vectorized solution would outperform the non-vectorized solution. Executing the notebook in the local Colab instance yielded identical results as those in the GitHub repository.



Vanishing Gradient Problem

The vanishing gradient problem typically plagues deep neural networks with many nodes requiring back propagation. Vanishing gradient problem manifests when loss gradients are so small that the loss curve flattens and fails to optimize. The vanishing gradient problem in this exploration is solved by employing a skip connection that adds the output to input. The local instance of the notebook returned identical results as the GitHub repository.


Skip.ipynb
File Edit View Insert Runtime Tools Help
Share
+ Code + Text
Copy to Drive
Gemini
[3]
Gradient Magnitude per Layer
Plain Network
Residual Network
Average Gradient Magnitude
Layer


Files
main
MathematicsOfML / Skip.ipynb
Top
Preview
Code
Blame
Raw
Gradient Magnitude per Layer
Plain Network
Residual Network
Average Gradient Magnitude
Layer