REFERENCES

Kannan, Anjuli., Kurach, Karol., Ravi, Sujith., Kaufmann, Tobias., Tomkins, Andrew., Miklos, Baliant., Corrado, Greg., Lukács, László., Ganea, Marina., Young, Peter., Ramavajjala, Vivek. 2016. *Smart Reply: Automated Response Suggestion for Email*. abs/1901.05639.

Cover, Thomas M., Thomas, Joy A. 2006. *Elements Of Information Theory*. John Wiley Sons, Inc.

Sun, Shiliang., Cao, Zehui., Zhu, Han., Zhao, Jin., 2019. *A Survey of Optimization Methods from a Machine Learning Perspective*. abs/1906.06821

Staudemeyer, C. R., Morris, E. C. 2019. *Understanding LSTM - a tutorial into Long Short-Term Memory Recurrent Neural Networks*. abs/1909.09586

Goodfellow, Ian., Bengio, Yoshua., Courville, Aaron. 2017. *Deep Learning*. The MIT Press.

Lan, Zhenzhong., Chen, Mingda., Goodman, Sebastian., Gimpel, Kevin., Sharma, Piyush., Soricut, Radu. 2019. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. abs/1909.11942

Polyak, B.T. 1964. *Some methods of speeding up the convergence of iteration methods*. Ussr computational mathematics and mathematical physics, volume 4(5).

Golub, Gene H., Loan, Charles F. Van. 1983. *Matrix Computations*. Johns Hopkins University Press, Baltimore, Maryland, 476 pp.

Huang, Liang., Zhao, Kai., Ma, Mingbo. 2017. *When to Finish? Optimal Beam Search for Neural Text Generation (modulo beam size)*. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2134–2139, Copenhagen, Denmark. Association for Computational Linguistics.

Horn, Roger A., Johnson, Charles R. 1990. *Matrix Analysis*. Cambridge University Press

Grimmett, Geoffrey., Stirzaker, David. 2001. *Probability and Random Processes*. Oxford University Press Inc., New York

Chee, Jerry., Li, Ping, 2020. *Understanding and Detecting Convergence for Stochastic Gradient Descent with Momentum*. abs/2008.12224