References

1. Zoom In: An Introduction to Circuits
   Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M. and Carter, S., 2020. Distill. DOI: 10.23915/distill.00024.001

2. Softmax Linear Units
   Elhage, N., Hume, T., Olsson, C., Nanda, N., Henighan, T., Johnston, S., ElShowk, S., Joseph, N., DasSarma, N., Mann, B., Hernandez, D., Askell, A., Ndousse, K., Jones, A., Drain, D., Chen, A., Bai, Y., Ganguli, D., Lovitt, L., Hatfield-Dodds, Z., Kernion, J., Conerly, T., Kravec, S., Fort, S., Kadavath, S., Jacobson, J., Tran-Johnson, E., Kaplan, J., Clark, J., Brown, T., McCandlish, S., Amodei, D. and Olah, C., 2022. Transformer Circuits Thread.

3. Compressed sensing
   Donoho, D.L., 2006. IEEE Transactions on information theory, Vol 52(4), pp. 1289--1306. IEEE.

4. Local vs. Distributed Coding
   Thorpe, S.J., 1989. Intellectica, Vol 8, pp. 3--40.

5. Representation learning: A review and new perspectives
   Bengio, Y., Courville, A. and Vincent, P., 2013. IEEE transactions on pattern analysis and machine intelligence, Vol 35(8), pp. 1798--1828. IEEE.

6. Curve Detectors [link]
   Cammarata, N., Goh, G., Carter, S., Schubert, L., Petrov, M. and Olah, C., 2020. Distill.

7. Superposition of many models into one
   Cheung, B., Terekhov, A., Chen, Y., Agrawal, P. and Olshausen, B., 2019. Advances in neural information processing systems, Vol 32.

8. Linguistic regularities in continuous space word representations [PDF]
   Mikolov, T., Yih, W. and Zweig, G., 2013. Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies, pp. 746--751.

9. Linguistic regularities in sparse and explicit word representations
   Levy, O. and Goldberg, Y., 2014. Proceedings of the eighteenth conference on computational natural language learning, pp. 171--180.

10. Unsupervised representation learning with deep convolutional generative adversarial networks
    Radford, A., Metz, L. and Chintala, S., 2015. arXiv preprint arXiv:1511.06434.

11. Visualizing and understanding recurrent networks [PDF]
    Karpathy, A., Johnson, J. and Fei-Fei, L., 2015. arXiv preprint arXiv:1506.02078.

12. Learning to generate reviews and discovering sentiment [PDF]
    Radford, A., Jozefowicz, R. and Sutskever, I., 2017. arXiv preprint arXiv:1704.01444.

13. Object detectors emerge in deep scene cnns [PDF]
    Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A., 2014. arXiv preprint arXiv:1412.6856.

14. Network Dissection: Quantifying Interpretability of Deep Visual Representations [PDF]
    Bau, D., Zhou, B., Khosla, A., Oliva, A. and Torralba, A., 2017. Computer Vision and Pattern Recognition.

15. Understanding the role of individual units in a deep neural network
    Bau, D., Zhu, J., Strobelt, H., Lapedriza, A., Zhou, B. and Torralba, A., 2020. Proceedings of the National Academy of Sciences, Vol 117(48), pp. 30071--30078. National Acad Sciences.

16. On the importance of single directions for generalization [PDF]
    Morcos, A.S., Barrett, D.G., Rabinowitz, N.C. and Botvinick, M., 2018. arXiv preprint arXiv:1803.06959.

17. On Interpretability and Feature Representations: An Analysis of the Sentiment Neuron
    Donnelly, J. and Roegiest, A., 2019. European Conference on Information Retrieval, pp. 795--802.

18. High-Low Frequency Detectors
    Schubert, L., Voss, C., Cammarata, N., Goh, G. and Olah, C., 2021. Distill. DOI: 10.23915/distill.00024.005

19. Multimodal Neurons in Artificial Neural Networks
    Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A. and Olah, C., 2021. Distill. DOI: 10.23915/distill.00030

20. Convergent learning: Do different neural networks learn the same representations?
    Li, Y., Yosinski, J., Clune, J., Lipson, H., Hopcroft, J.E. and others,, 2015. FE@ NIPS, pp. 196--212.

21. Feature Visualization [link]
    Olah, C., Mordvintsev, A. and Schubert, L., 2017. Distill. DOI: 10.23915/distill.00007

22. Adversarial examples are not bugs, they are features
    Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B. and Madry, A., 2019. Advances in neural information processing systems, Vol 32.