Interpreting Neural Networks through Mahalanobis Distance

Alan Oursland

alan.oursland@gmail.com

October 2024

Abstract

This paper introduces a theoretical framework that connects neural network linear layers with the Mahalanobis distance, offering a new perspective on neural network interpretability. While previous studies have explored activation functions primarily for performance optimization, our work interprets these functions through statistical distance measures, a less explored area in neural network research. By establishing this connection, we provide a foundation for developing more interpretable neural network models, which is crucial for applications requiring transparency. Although this work is theoretical and does not include empirical data, the proposed distance-based interpretation has the potential to enhance model robustness, improve generalization, and provide more intuitive explanations of neural network decisions.

1 Introduction

Neural networks have revolutionized machine learning, achieving remarkable success across diverse applications. Central to their efficacy is the use of activation functions, which introduce non-linearity and enable the modeling of complex relationships within data. While Rectified Linear Units (ReLU) have gained prominence due to their simplicity and effectiveness [Nair and Hinton, 2010], the exploration of alternative activation functions remains an open and valuable area of research [Ramachandran et al., 2018].

Neural network units are often viewed as linear separators that define decision boundaries between classes [Minsky and Papert, 1969] with larger activation values suggesting stronger contributions of features to those decisions. Our work challenges this perspective, exploring how individual neurons can be understood through the lens of statistical distance measures. Clustering techniques aim to minimize the distance between data points and feature prototypes, with smaller values indicating stronger membership to the feature or cluster [MacQueen, 1967a]. Our work explores the intersection between these perspectives, leveraging the distance-minimization approach of clustering techniques to lay the groundwork for novel neural network designs based on statistical distance measures.

This paper establishes a novel connection between neural network architectures and the Mahalanobis distance, a statistical measure that accounts for the covariance structure of data [Mahalanobis, 1936]. We present a robust mathematical framework that bridges neural networks to this statistical distance measure and lay the groundwork for future research into neural network interpretability and design. This distance-based interpretation has the potential to enhance model robustness, improve generalization, and offer more intuitive explanations of neural network decisions. Our key contributions are:

1. We establish a mathematical connection between neural network linear layers and the Mahalanobis distance, demonstrating how Absolute Value (Abs) activations facilitate distance-based interpretations.

2. We analyze the solution space that neural networks are likely to learn when approximating Mahalanobis distance, exploring the effects of non-uniqueness in whitening transformations and the role of Abs-activated linear nodes.

3. We discuss the broader implications of this framework for neural network design and interpretability, laying the groundwork for more interpretable models.

2 Background and Related Work

2.1 Activation Functions

Activation functions introduce non-linearity in neural networks, enabling them to model complex data relationships. The field has evolved from early sigmoid and hyperbolic tangent functions [Rosenblatt, 1958] to the widely adopted Rectified Linear Unit (ReLU) [Nair and Hinton, 2010], which mitigates the vanishing gradient problem in deep networks [Glorot and Bengio, 2010, Krizhevsky et al., 2012].

ReLU variants intended to address its shortcomings include Leaky ReLU [Maas et al., 2013]; Parametric ReLU (PReLU) [He et al., 2015]; and Exponential Linear Unit (ELU) [Clevert et al., 2016]. Additionally, newer activation functions like Swish [Ramachandran et al., 2018] and GELU [Hendrycks and Gimpel, 2016] have been proposed to further enhance network performance and training dynamics.

Tanh and Sigmoid activations are still used in many architectures such as recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks [Hochreiter and Schmidhuber, 1997].

The variety of activation functions used in modern networks reflects the diverse needs of different architectures. The exploration of activation functions remains an active area of research, with ongoing investigations into their impact on neural network performance, generalization, and interpretability [Ramachandran et al., 2018]. Despite extensive research, the interpretation of activation functions in terms of statistical measures remains an open area of investigation.

2.2 Overview of Distance Metrics in Clustering

Distance metrics are fundamental in clustering algorithms, determining how similarity between data points is measured. Various clustering methods employ different distance measures:

- K-Means typically uses Euclidean distance (ℓ_2 norm), assuming spherical clusters and equal feature importance [MacQueen, 1967b].
- Gaussian Mixture Models (GMMs) employ Mahalanobis distance, accounting for data covariance and modeling elliptical clusters [Reynolds, 2009].
- Radial Basis Function (RBF) networks use Gaussian-like activations based on Euclidean distance, creating spherical clusters around learned centers [Broomhead and Lowe, 1988].
- Hierarchical and Agglomerative Clustering can use various metrics (Euclidean, Manhattan, correlation-based), affecting dendrogram shape [Murtagh, 1983].
- DBSCAN, while often using Euclidean distance, can employ any metric for density-based clustering [Ester et al., 1996].
- Spectral Clustering incorporates similarity measures like Gaussian kernel functions [Von Luxburg, 2007].

The Mahalanobis distance stands out for its ability to account for feature correlations and scale differences, making it particularly useful in multivariate analysis [Mahalanobis, 1936, De Maesschalck et al., 2000]. It provides a scale-invariant measure that adjusts for the covariance structure of the data, offering advantages in high-dimensional spaces.

Understanding these distance metrics and their properties is crucial for selecting appropriate clustering algorithms and interpreting their results. As we explore the connection between neural networks and distance-based interpretations, these insights from clustering algorithms provide valuable context and inspiration.

2.3 Neural Network Interpretability and Statistical Models

The interpretability of neural networks remains a critical challenge, often referred to as the "black-box" nature of these models [Lipton, 2016]. In applications requiring transparency, such as healthcare and finance, understanding the decision-making processes of neural networks is paramount [Rudin, 2019]. Various approaches have been developed to enhance interpretability, including feature visualization, saliency maps, and prototype-based methods [Erhan et al., 2009, Simonyan and Zisserman, 2013, Kim et al., 2018a].

Recent advancements in explainable AI (XAI) have introduced tools like SHAP (SHapley Additive exPlanations) [Lundberg and Lee, 2017] and LIME (Local Interpretable Model-Agnostic Explanations) [Ribeiro et al., 2016], which provide both local and global insights into model predictions. SHAP offers a consistent approach to feature attribution grounded in cooperative game theory, providing robust explanations even for complex models like deep neural networks [Lundberg and Lee, 2017, Markov, 2020]. LIME, by contrast, offers localized interpretability by approximating the model's behavior with simpler, interpretable models, making it particularly useful for individual predictions [Ribeiro et al., 2016, Ali et al., 2024].

Connections between neural networks and statistical models, such as Bayesian neural networks [Neal, 1996, Blundell et al., 2015], continue to provide a probabilistic framework for understanding model uncertainty. Moreover, concept-based interpretability methods like TCAV (Testing with Concept Activation Vectors) [Kim et al., 2018b] have emerged, allowing for a more granular analysis of what neural networks learn, which can be crucial in high-stakes domains like healthcare [Hanif et al., 2024].

While significant strides have been made in explaining model behavior, there remains a gap in establishing direct mathematical connections between neural network components, specifically activation functions, and statistical distance measures like the Mahalanobis distance. Addressing this gap can provide deeper insights into feature learning and decision-making processes, enhancing both interpretability and robustness of neural network models.

3 Mathematical Framework

Gaussians fall out of second-order Taylor series approximations [Bishop, 2006, Section 4.4], making them effective for modeling data, even when the data is not explicitly Gaussian. Gaussian mixtures can serve as piecewise linear approximations of complex distributions and surfaces. They are a good choice for modeling point clouds such as the ones neural networks are trained on.

In this section, we develop the mathematical foundation that connects neural networks to the Mahalanobis distance, thereby providing a framework for interpreting neural network operations through the lens of statistical distance metrics. We begin by revisiting key concepts related to Gaussian distributions and the Mahalanobis distance, followed by a detailed exploration of how neural network components, particularly linear layers and activation functions, can approximate these distance metrics. This framework not only enhances our understanding of neural network behavior but also lays the groundwork for leveraging statistical principles to improve network

interpretability and training dynamics.

3.1 Mahalanobis Distance for a Multivariate Gaussian Distribution

A multivariate Gaussian (Normal) distribution is a fundamental concept in statistics, describing a d-dimensional random vector $\mathbf{x} \in \mathbb{R}^d$ with a mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ [Bishop, 2006]. We denote this distribution as $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The Mahalanobis distance quantifies the distance between a point $\mathbf{x}$ and the mean $\boldsymbol{\mu}$ of a distribution, while considering the covariance structure of the data [Mahalanobis, 1936, De Maesschalck et al., 2000]. It is defined as:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}. \tag{1}$$

This metric adjusts for variance across dimensions by effectively whitening the data, resulting in a spherical distance measure.

3.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms data into a new coordinate system, emphasizing directions (principal components) that capture the most variance [Jolliffe, 2002]. When performing PCA on the covariance matrix $\boldsymbol{\Sigma}$, it is decomposed using eigenvalue decomposition:

$$\boldsymbol{\Sigma} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^\top, \tag{2}$$

where:

- $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d]$ is a matrix whose columns are the orthogonal unit eigenvectors of $\boldsymbol{\Sigma}$.
- $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ is a diagonal matrix of the corresponding eigenvalues λ_i, representing the variance along each principal component.

Substituting $\mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^\top$ for $\boldsymbol{\Sigma}$ in the Mahalanobis distance equation (1), we obtain:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{V} \boldsymbol{\Lambda}^{-1} \mathbf{V}^\top (\mathbf{x} - \boldsymbol{\mu})}. \tag{3}$$

To further simplify, we can express the Mahalanobis distance in terms of the principal components:

$$\begin{aligned} D_M(\mathbf{x}) &= \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{V} \boldsymbol{\Lambda}^{-1} \mathbf{V}^\top (\mathbf{x} - \boldsymbol{\mu})} \\ &= \sqrt{(\mathbf{V}^\top (\mathbf{x} - \boldsymbol{\mu}))^\top \boldsymbol{\Lambda}^{-1} (\mathbf{V}^\top (\mathbf{x} - \boldsymbol{\mu}))} \\ &= \sqrt{\sum_{i=1}^{d} \lambda_i^{-1} \left(\mathbf{v}_i^\top (\mathbf{x} - \boldsymbol{\mu}) \right)^2} \\ &= \left\| \lambda_i^{-1/2} \mathbf{v}_i^\top (\mathbf{x} - \boldsymbol{\mu}) \right\|_2 . \end{aligned} \tag{4}$$

where $\| \cdot \|_2$ denotes the Euclidean (ℓ_2) norm. This shows that the Mahalanobis distance can also be expressed as the ℓ_2 norm of the number of standard deviations of $\mathbf{x}$ along each principal component.

3.3 Connecting Neural Networks to Mahalanobis Distance

We consider the Mahalanobis distance along a single principal component.

$$D_{M,i}(\mathbf{x}) = \left| \lambda_i^{-1/2} \mathbf{v}_i^\top (\mathbf{x} - \boldsymbol{\mu}) \right|, \tag{5}$$

This equation projects the centered data $(\mathbf{x} - \boldsymbol{\mu})$ onto the direction of variance defined by the principal component eigenvector and scales it by the inverse square root of the eigenvalue.

Let

$$\mathbf{W} = \lambda_i^{-1/2} \mathbf{v}_i^\top, \tag{6}$$

$$\mathbf{b} = -\lambda_i^{-1/2} \mathbf{v}_i^\top \boldsymbol{\mu}. \tag{7}$$

We can simplify Equation (5) to

$$D_{M,i}(\mathbf{x}) = |\mathbf{W}\mathbf{x} - \mathbf{b}|, \tag{8}$$

This is identical to the equation for a linear layer where $\boldsymbol{W}$ represents the weight matrix, $\boldsymbol{b}$ the bias vector, and the Abs function serves as the activation function. Each linear node with an Abs activation can be interpreted as modeling a one-dimensional Gaussian along a principal component direction, with the decision boundary passing through the mean of the modeled cluster. The layer as a whole represents a subset of principal components from a Gaussian Mixture Model (GMM) that approximates the input distribution. Since each component captures significant features individually, we do not need to aggregate them via an ℓ_2 norm (full Mahalanobis distance computation). Instead, the subsequent layer clusters these principal component features, effectively forming a new GMM that models the outputs of the first layer.

3.4 Non-Uniqueness of Whitening

The principal components of a Gaussian distribution, as used in the Mahalanobis distance, form an orthonormal set of axes. Projecting Gaussian data onto these axes transforms the distribution from an oriented ellipsoid into a spherical Gaussian, effectively converting the data from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This process is known as *whitening* [Bishop, 2006, Section 12.1.3].

However, the transformation to whitened data is not unique. Specifically, any rotation applied in the whitened space results in another valid whitening transformation. Mathematically, if $\mathbf{x}_w$ is the whitened data, then for any orthogonal rotation matrix $\mathbf{R} \in \mathrm{SO}(d)$, the rotated data $\mathbf{x}'_w = \mathbf{R}\mathbf{x}_w$ is also whitened.

This non-uniqueness implies that multiple sets of axes, possibly non-orthogonal in the original space, can serve as a whitening basis. When we transform the rotated basis back to the original space, we obtain a new set of basis vectors $\mathbf{W}$ that still whiten the data but may not correspond to the original principal components and may not even be orthogonal.

In the context of neural networks, this means that although linear nodes can represent directions that effectively whiten the data, they are unlikely to precisely learn the actual principal components when estimating Mahalanobis distances. Instead, they may learn any basis that achieves whitening. Nevertheless, the learned hyperplanes (decision boundaries) should still pass through the data mean $\boldsymbol{\mu}$, allowing for prototype interpretation.

To encourage the network to learn the actual principal components, one could apply an orthogonality constraint or regularization on the weight matrices. This regularization promotes learning orthogonal directions, aligning the learned basis with the true principal components of the data clusters and providing statistically independent features.

4 Implications and Discussion

We discuss implications, potential impact and future work of this reframing of linear layers in neural networks. While this paper provides a robust theoretical foundation for interpreting neural networks through Mahalanobis distance and Abs activation functions, it does not include empirical results. Future work will involve validating these theoretical insights with empirical data to further assess their applicability and performance in real-world scenarios.

4.1 Expected Value Interpretation

The expected value, or mean, is a central concept in statistics, representing the average tendency of a distribution. In neural networks, finding the expected value for each neuron would reveals the features it recognizes. Interpreting linear nodes as approximations of Gaussian principal components provides a path towards recoving the neuron mean value. The estimated mean serves as a prototype for the feature that the neuron has learned to recognize [Li et al., 2018], representing the 'ideal' input for that neuron. This interpretation enhances the transparency of the feature extraction process, potentially leading to more interpretable models and improved architectures.

4.2 Equivalence between Abs and ReLU Activations

While our analysis utilizes linear layers with Abs activation functions to model deviations along principal component directions, ReLU activations can provide comparable information within the same framework.

For the *Abs activation*, each linear node computes:

$$y_{\text{Abs}} = \left| \mathbf{w}^\top \mathbf{x} + b \right|, \tag{9}$$

where the weights $\mathbf{w}$ and bias b are set such that $\mathbf{w}^\top \boldsymbol{\mu} + b = 0$. This centers the decision boundary at the cluster mean $\boldsymbol{\mu}$, and within a confidence interval δ, the pre-activation output ranges from $-\delta$ to $+\delta$.

For the *ReLU activation*, we adjust the bias to shift the decision boundary just outside the cluster:

$$y_{\text{ReLU}} = \max\left(0, -\mathbf{w}^\top \mathbf{x} - b + \delta\right). \tag{10}$$

Here, the pre-activation output ranges from 0 to 2δ within the cluster. Although ReLU zeros out negative inputs, by negating the pre-activation and adjusting the bias, it effectively captures the magnitude of deviations similar to the Abs activation.

The hyperplanes defined by $\mathbf{w}$ maintain the same orientation in both cases, providing equivalent views of the cluster. Subsequent layers can adapt to either activation's output range, making Abs and ReLU functionally comparable in capturing essential features.

This suggests that techniques developed for networks with Abs activations may be adaptable to ReLU activations, bridging theoretical insights with practical neural architectures commonly utilizing ReLU.

4.3 Activations as Distance Metrics

Traditional neural networks typically employ an "intensity metric model," where larger activation values indicate stronger feature presence. In contrast, a "distance metric model" interprets smaller activation values as indicating closer proximity to a learned feature or prototype. The following observations suggest directions for future work:

- Most error functions (e.g., Cross Entropy Loss, Hinge Loss) are designed for intensity metrics. Output layers using Abs activation may require modification of their output values.
- While some architectures, like Radial Basis Function networks [Broomhead and Lowe, 1988], utilize distance metrics, they are not widely adopted in modern deep learning.
- Distance metrics conflict with the goal of sparse output layers. In a distance metric model, zero is the strongest signal, making it illogical for most outputs to have the strongest signal.
- The Gaussian connection suggests transforming distance metrics through exponential ($y = e^{-x^2}$) or Laplace ($y = e^{-|x|}$) functions to convert them into intensity metrics. However, these may suffer from vanishing gradients. A approximation of these functions could combine Abs and ReLU: $y = \text{ReLU}(-\text{Abs}(x) + \text{confidence\_bound})$.
- Distance and intensity metrics can be interconverted through negation. Subsequent layer weights can apply their own negation, obscuring the metric type learned by internal nodes.
- There may exist regularization techniques that encourage distance metric learning [Weinberger and Saul, 2009].

4.4 Model Initialization and Pretraining

Interpreting neurons as learning distances from cluster means suggests novel approaches to model initialization and pretraining. This perspective offers an alternative to standard random initialization techniques [Kamilov et al., 2017] by incorporating data-driven insights into the model's starting configuration.

Rather than initializing with random weights, an approach could involve clustering the input data (e.g., using k-means) and calculating the covariance of each cluster. Applying Principal Component Analysis (PCA) to these covariance matrices can provide a basis for directly initializing network parameters. This strategy leverages the structure of the data to guide the network's early learning stages. This process, and approximations of this process, may offer several advantages:

- Faster convergence by starting with parameters informed by the data distribution
- Enhanced interpretability, as network weights are aligned with meaningful features from the outset
- Improved generalization by incorporating information about cluster structures

4.5 Model Translation and Componentization

The interpretation of neurons as principal components of Gaussians suggests a potential mapping between neural networks and hierarchical Gaussian Mixture Models (GMMs) [Jacobs et al., 1991]. By performing PCA on the clusters in a GMM, we can extract principal components, converting them directly into neurons. Conversion from neurons to Gaussian representations may also be possible. The process of directly translating betweem neural networks and GMMs offers several potential advantages:

- **Enhanced Interpretability**: Neural networks can be better understood through their GMM equivalents, providing insights into the data distribution and feature representations.
- **Application of Statistical Techniques**: Established statistical methods used in GMM analysis can be applied to neural networks, potentially improving training and evaluation.

- **Hybrid Models**: Combining neural networks and GMMs can leverage the strengths of both, enhancing performance in tasks like clustering and classification.
- **Model Decomposition**: Large networks might be decomposable into smaller, context-specific subnetworks, facilitating easier analysis and maintenance.
- **Efficient Storage and Computation**: Subnetworks can be stored offline and dynamically loaded based on data context, improving memory efficiency and reducing computational overhead.
- **Scalability in Large-Scale Applications**: This approach can lead to faster inference and more efficient resource utilization in applications dealing with massive datasets.

4.6 Direct use of Mahalanobis equation

Equation 5 explicitly incorporates the variance eigenvalue λ, the unit eigenvector $\mathbf{v}$, and the mean $\boldsymbol{\mu}$. Batch Normalization already makes use of λ and $\boldsymbol{\mu}$ [Ioffe and Szegedy, 2015], while the nGPT model employs unit weight vectors, which are analogous to $\mathbf{v}$ [Loshchilov et al., 2024]. The success of these techniques suggest there might be further opportunities to decompose the standard linear layer equation $y = Wx + b$ towards the Mahalanobis equation in a way that leads to improvements in training speed and representation quality.

5 Conclusion

This paper establishes a novel connection between neural network architectures and the Mahalanobis distance, providing a fresh perspective on neural network interpretability. By demonstrating how linear layers with Abs activations can approximate Mahalanobis distances, we bridge the gap between statistical distance measures and neural network operations. This framework offers several key insights:

- It provides a probabilistic interpretation of neural network nodes as learning principal components of Gaussian distributions.
- It suggests new approaches for model initialization, pretraining, and componentization.
- It establishes a potential homomorphism between neural networks and hierarchical Gaussian Mixture Models.

These findings lay the groundwork for future research into more interpretable and robust neural network architectures. By leveraging statistical principles in neural network design, we open new avenues for enhancing model transparency, improving generalization, and developing more efficient training techniques. As the field of AI continues to evolve, such interpretable frameworks will be crucial in building trustworthy and explainable AI systems.

References

ABM Shawkat Ali, Ambreen Hanif, and Amin Beheshti. Explainable ai frameworks: Navigating the present challenges and unveiling innovative applications. *Algorithms*, 17(6):227, 2024. doi: 10.3390/a17060227.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1613–1622. PMLR, 2015.

David S Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. *Royal Signals and Radar Establishment Malvern (United Kingdom) RSRE Memorandum*, 4148, 1988.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *4th International Conference on Learning Representations (ICLR)*, 2016. URL `https://arxiv.org/abs/1511.07289`.

R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000. doi: 10.1016/S0169-7439(99)00047-7.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. In *Technical Report, University of Montreal*, 2009.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

Ambreen Hanif, Amin Beheshti, and ABM Shawkat Ali. Interpreting artificial intelligence models: A systematic review on the application of lime and shap in alzheimer's disease detection. *Brain Informatics*, 11(1):1–19, 2024. doi: 10.1186/s40708-023-00195-9.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. URL `https://arxiv.org/abs/1606.08415`.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 448–456, 2015. URL `https://arxiv.org/abs/1502.03167`.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2 edition, 2002.

Ulugbek S. Kamilov, Hassan Mansour, and Brendt Wohlberg. A survey of computational imaging methods for inverse problems. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2668–2677. PMLR, 2018a.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018b.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zachary C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016. URL `https://arxiv.org/abs/1606.03490`.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. ngpt: Normalized transformer with representation learning on the hypersphere, 2024. URL `https://arxiv.org/abs/2410.01131`.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967a.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297. University of California Press, 1967b.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

Mark Markov. Lime vs shap: A comparative analysis of interpretability tools. *MarkovML Blog*, 2020. URL `https://www.markovml.com/blog/lime-vs-shap-comparative-analysis`.

Marvin L. Minsky and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, USA, 1969. ISBN 9780262631832.

Fionn Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer-Verlag, 1996.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Hk4_qw5K-`. arXiv preprint arXiv:1710.05941.

Douglas A. Reynolds. Gaussian mixture models. In *Encyclopedia of Biometrics*, pages 659–663. Springer, 2009.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Karen Simonyan and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. URL `https://arxiv.org/abs/1312.6034`.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In *International Conference on Machine Learning (ICML)*, pages 1317–1324, 2009. URL `https://proceedings.mlr.press/v5/weinberger09a.html`.