Ensembles of Random Projections for Nonlinear Dimensionality Reduction

Amir-Hossein Karimi — Computer Science Dept., University of Waterloo, ON, Canada
Mohammad Javad Shafiee — Systems Design Eng. Dept., University of Waterloo, ON, Canada
Ali Ghodsi — Stats & Acturial Sciences, University of Waterloo, ON, Canada
Alexander Wong — Systems Design Eng. Dept., University of Waterloo, ON, Canada

Abstract

Dimensionality reduction methods are widely used in information processing systems to better understand the underlying structures of datasets, and to improve the efficiency of algorithms for big data applications. Methods such as linear random projections have proven to be simple and highly efficient in this regard, however, there is limited theoretical and experimental analysis for nonlinear random projections. In this study, we review the theoretical framework for random projections and nonlinear rectified random projections, and introduce *ensemble of nonlinear maximum random projections*. We empirically evaluate the embedding performance on 3 commonly used natural datasets and compare with linear random projections and traditional techniques such as PCA, highlighting the superior generalization performance and stable embedding of the proposed method.

1 Introduction

Many information processing systems and intelligent decision-making systems operate on measured real-world data that often have a large number of components and high dimensionality. To adequately and efficiently handle this sort of data, these system may first obtain lower-dimensional representations of the data samples. As a result, dimensionality reduction enables, among others, data compression, data visualization, machine learning, and handling of large volumes of high-dimensional data enabling researchers across a variety of fields to overcome the *curse of dimensionality* that comes with having more information.

Methods for dimensionality reduction are plentiful and have been successfully applied to applications such as head pose estimation [1], visualization of biomedical data [2], face [3] and speech recognition [4], and gene expression analysis [5] among others. Different techniques are used across various data setups taking into account assumptions about the complexity and degrees of freedom of the input data, and performances and running-time complexities vary based on the desired level of accuracy, and the assumptions made about the underlying manifold.

Common methods for dimensionality reduction include Principle Component Analysis (PCA) [6] that finds the optimal embedding with maximum variance, Multi-dimensional Scaling (MDS) [7] that optimizes an eigenvalue problem to find an embedding that preserves pair-wise Euclidean distances, and Isomap [8] which takes the distribution of neighboring points into account in finding an embedding that preserves pair-wise geodesic distances. Along the lines of Isomap, Locally Linear Embedding (LLE) [9] preserves local properties of the data manifold by attempting to preserve the reconstruction weights of each sample obtained from writing / reconstructing each original sample as a linear combination of its nearest neighbors in the original high dimensional space.

Despite the success of these methods, special care must be considered when choosing an appropriate dimensionality reduction method. Specifically, the dependence on data often leads researchers to experiment with multiple dimensionality reduction methods before moving on to the rest of their algorithms. The challenge with running multiple experiments to settle on an appropriate method is further exacerbated when dealing with high dimensional data or a large number of datapoints. In PCA, for example, computing the covariance matrix for a large number of features becomes exponentially more expensive as the dimensionality increases, and in MDS or Kernel PCA, constructing pairwise distances to feed into the optimization problem grows exponentially with the number of samples and causes an efficiency bottleneck. As information processing systems tackle larger-scale applications, big-data scenarios are becoming the norm and there seems to be a more urgent need to explore more efficient methods for dimensionality reduction that are universal in their applicability to different datasets.

In this regard, *Random Projections* (RP) [10] are simple, efficient, and data-independent methods for dimensionality reduction. The Johnson-Lindenstrauss (JL) theorem [11] asserts that, using a linear projection that is independent of the samples themselves, one can find an embedding in $\mathcal{O}(\log n/\varepsilon^2)$ dimensions where n is the number of samples and ε is the error tolerance. Assuming the embedding satisfies a minimum projected space dimension k, this embedding will preserve pair-wise Euclidean distances with high probability. As we shall review in the following section, this lower-bound depends only on the number of samples n and the error margin ε, but not on the original data dimensionality d rendering random projections as an exceptionally powerful dimensionality reduction tool for very high dimensional data. The simplicity and universal applicability of random projections are further brought to light when one considers how to construct linear random projections: all entries of a $k \times d$ projection matrix can be populated uniformly and independently from a standard Normal distribution [12, 13], or can be independently drawn from $\{-1, 0, +1\}$ [14] resulting in sparse, and consequently more efficient, random projections.

Recently a study demonstrated that the theory for linear random projections can be extended to nonlinear random projections by applying the ReLU activation function elementwise on the embedding [15]. The authors demonstrate that this form of nonlinear random projection performs a class-aware embedding where the embedding places objects of the same class closer to one another after the projection compared to objects of different classes.

Nonlinear dimensionality reduction methods such as this hold promise to offer an advantage over their linear counterparts for real-world data, as real-world data is likely to lie on or near a highly nonlinear manifold. This is the question we explore in this study. We extend the line of work above in this experimental study by employing an ensemble of random projections and using the maximum activations as the embedding coefficient. In the work that follows, we empirically demonstrate how this form of random projection leads to stable low-dimensional embeddings that perform better than linear random projections, nonlinear rectified random projections [15], and PCA.

2 Methodology

Inspired to extend linear random projection to nonlinear random projections to tackle complicated real-world datasets, we first review the theory on linear RPs in the context of dimensionality reduction. Dimensionality reduction attempts to find an embedding $\boldsymbol{Y} \subset \mathbb{R}^k$ of the original set $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\} \subset \mathbb{R}^d$. In particular, dimensionality reduction based on random projections rely on the Johnson-Lindenstrauss (JL) theorem [11] to assert the existence of an embedding that preserves all pair-wise Euclidean (l_2) distance, with high probability. More specifically,

Theorem 2.1 *For any set $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\} \subset \mathbb{R}^d$, any integer n (number of samples), and any $0 < \varepsilon < 1$ (error tolerance), let k be a positive integer satisfying*

$$k \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \log n \tag{1}$$

then, there exists a map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $\boldsymbol{x_i}, \boldsymbol{x_j} \in \boldsymbol{X}$, with probability greater than $1 - \delta$ we have

$$(1-\varepsilon)||\boldsymbol{x_i} - \boldsymbol{x_j}||^2 \leq ||f(\boldsymbol{x_i}) - f(\boldsymbol{x_j})||^2 \leq (1+\varepsilon)||\boldsymbol{x_i} - \boldsymbol{x_j}||^2 \tag{2}$$

One such embedding function $f(\cdot)$ is simply a projection matrix $\boldsymbol{R} \in \mathbb{R}^{d \times k}$ where each coefficient $r_{ij} \sim \frac{1}{\sqrt{k}}\mathcal{N}(0,1) = \mathcal{N}(0, \frac{1}{k})$. Therefore, the above can be written equivalently as

$$Pr\left[\left|||\boldsymbol{y_i} - \boldsymbol{y_j}||^2 - ||\boldsymbol{x_i} - \boldsymbol{x_j}||^2\right| \geq \varepsilon\right] \leq \delta \tag{3}$$

where $\boldsymbol{y_i} = \boldsymbol{R}^T \boldsymbol{x_i}$ and $\boldsymbol{y_j} = \boldsymbol{R}^T \boldsymbol{x_j}$. For a proof of the above theorem, as well as other forms of embeddings, refer to [12, 13, 14]. Giryes *et al.* [15] extended this line of work to nonlinear RPs by applying an activation function on the embedded samples, $\boldsymbol{y_i}\ \forall\, i \in [n]$. In particular, a ReLU operator ($\rho(w) = w \cdot \mathbb{1}\{w \geq 0\}$) was applied elementwise to each coefficient of the embedded samples. This resulted in the introduction of an additional term in (3) which depends on the

Table 1: 1-NN performance of 3 common imaging datasets at different projected dimensions k. All samples have original dimensionality d equal to $32 \times 32 \times 3 = 3072$. Cells in bold demostrate the superior performance of Nonlinear Ensemble RP across various datasets and different projected dimensions $k \geq \mathcal{O}(\log n/\varepsilon^2)$. All reported results are the average of 10 runs.

Datasets			Projection Type	Projected Dimensions k						
Name	# Classes	n (training)		64	128	256	512	1024	2048	3072
CIFAR-10	10	2,500	No RP							27.26± 00.00
			PCA	**30.27± 00.00**	**29.28± 00.00**	**27.90± 00.00**	27.40± 00.00	27.30± 00.00	27.25± 00.00	27.26± 00.00
			Linear RP	25.12± 00.32	26.35± 00.37	26.83± 00.43	26.93± 00.27	26.93± 00.26	27.04± 00.13	27.06± 00.18
			Nonlinear Rectified RP	23.05± 00.51	25.32± 00.44	26.33± 00.50	26.79± 00.22	27.08± 00.30	26.98± 00.21	27.12± 00.12
			Nonlinear Ens. Max RP	19.82± 00.44	23.12± 00.47	25.96± 00.47	**27.93± 00.28**	**29.15± 00.25**	**29.82± 00.38**	**30.02± 00.29**
STL-10	10	2,500	No RP							27.56± 00.00
			PCA	**30.52± 00.00**	**30.00± 00.00**	**28.80± 00.00**	28.16± 00.00	27.70± 00.00	27.57± 00.00	27.56± 00.00
			Linear RP	26.58± 00.67	27.36± 00.33	27.64± 00.33	27.69± 00.27	27.68± 00.22	27.69± 00.19	27.69± 00.18
			Nonlinear Rectified RP	25.12± 00.55	26.84± 00.58	27.56± 00.22	27.61± 00.41	27.88± 00.21	27.98± 00.14	27.98± 00.12
			Nonlinear Ens. Max RP	22.90± 00.68	25.60± 00.73	27.51± 00.54	**28.74± 00.45**	**29.46± 00.39**	**29.90± 00.28**	**30.03± 00.31**
ImageNet (Tiny)	2 (avg of 10 pairs)	1,000	No RP							51.50± 00.00
			PCA	51.50± 00.00	51.50± 00.00	51.50± 00.00	51.50± 00.00	**51.50± 00.00**	51.50± 00.00	51.50± 00.00
			Linear RP	51.60± 03.42	50.75± 02.21	51.64± 02.06	**51.64± 01.72**	51.35± 01.23	**51.58± 01.14**	**51.58± 00.88**
			Nonlinear Rectified RP	50.98± 03.56	50.87± 02.48	51.22± 02.13	50.83± 01.71	50.97± 01.27	51.28± 01.00	51.18± 01.10
			Nonlinear Ens. Max RP	**52.53± 03.22**	**51.61± 02.69**	**51.96± 02.17**	50.76± 01.81	51.13± 01.79	51.36± 01.14	51.27± 01.13

angular distance between samples in the original space (i.e., the $\boldsymbol{x_i}$s)

$$Pr\left[\left|||\boldsymbol{y_i}-\boldsymbol{y_j}||^2-\left(\frac{1}{2}||\boldsymbol{x_i}-\boldsymbol{x_j}||^2+||\boldsymbol{x_i}||\,||\boldsymbol{x_j}||\,\Psi(\boldsymbol{x_i},\boldsymbol{x_j})\right)\right|\geq\varepsilon\right]\leq\delta \quad (4)$$

where $\boldsymbol{y_i}=\rho(\boldsymbol{R}^T\boldsymbol{x_i})$, $\boldsymbol{y_j}=\rho(\boldsymbol{R}^T\boldsymbol{x_j})$, $\Psi(\boldsymbol{x_i},\boldsymbol{x_j})=\frac{1}{\pi}\big(\sin(\theta)-\theta\cos(\theta)\big)$, and $\theta=\angle(\boldsymbol{x_i},\boldsymbol{x_j})$, the angular distance between $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$. The authors show that $\Psi(\boldsymbol{x_i},\boldsymbol{x_j})$ is approximately equal to $0.5(1-\cos(\theta))$, helping us understand that the probability bound (4) suggests that nonlinear rectified RPs activation function will perform class-aware embedding of the data that is sensitive to angles between points: *such embeddings tend to decrease the Euclidean distances between points with a small angle between them ("same class") more than the distances between points with large angles between them ("different classes").*

In addition to ReLU as an activation function for nonlinear RPs, the authors of [15] claim a similar analysis can be derived for the spatial pooling operation commonly used in convolutional neural networks (CNNs). In this work, we explore the effect of choosing the max activation as the embedded feature on the quality of the embeddings for the application of dimensionality reduction. In contrast to spatial pooling used in CNNs, that supports embedding-robustness via spatial invariance, our strategy selects the maximum activation of m *randomly* selected features as the embedded coefficient. This form of nonlinearity is supported by an ensemble of random projection matrices $\{\boldsymbol{R}^{(1)},\cdots,\boldsymbol{R}^{(m)}\}\subset\mathbb{R}^{d\times k}$ that embed each input sample $\boldsymbol{x_i}$ into $\boldsymbol{y_i}$ via

$$\boldsymbol{y}_{ij}=\max\left\{(\boldsymbol{R}_j^{(1)})^T\boldsymbol{x_i},\cdots,(\boldsymbol{R}_j^{(m)})^T\boldsymbol{x_i}\right\}\ \forall\ j\in[k] \quad (5)$$

where $\boldsymbol{y}_{ij}$ is the j^{th} coefficient of embedded point $\boldsymbol{y_i}$, and $\boldsymbol{R}_j^{(l)}$ is the j^{th} column of the l^{th} projection matrix $\boldsymbol{R}^{(l)}$. We refer to this method as *Ensemble of Nonlinear Maximum Random Projections*.

3 Experiments

In this section, we describe our experimental setup, parameters, and metrics used to compare the performance of the proposed method against other dimensionality reduction methods. In the experiments below, n is the number of samples, d is the original dimensionality, and k is the projected dimensionality of the embedded space. Table 1 summarizes the results averaged over 10 runs.

3.1 Evaluation Metric

We assess the quality of the embedding by evaluating how the local structure is retained in the projected space. This is accomplished by measuring the generalization error of 1-nearest neighbor (1-NN) classifier trained on the low-dimensional representation of the data (as is done, e.g., in [16, 17]). Ideally, dimensionality reduction reduces the number of data features while maintaining a certain level of generalization performance. As we shall see, in many cases dimensionality reduction methods lead to improved generalization errors in the lower dimensions, a characteristic much desired.

3.2 Datasets

For our experiments, we selected three datasets that represent tasks from a variety of domains: (i) the CIFAR-10 dataset [18], (ii) the STL-10 dataset [19], (iii) the ImageNet Tiny dataset. These datasets were selected because they satisfy the theoretical conditions for stable embedding using random projections[1].

The first two datasets comprise of 10 classes of natural scene images with image size $32\times32\times3$. This value constitutes the original dimensionality of the samples. The CIFAR-10 and STL-10 datasets contain $50,000$ and $5,000$ training samples, respectively. For computational reasons, we randomly selected 250 samples from each of the 10 classes as the training set, and used the entire test set to compute 1-NN performance.

The final dataset comprised of 200 classes of natural images, with 500 training samples and 50 test samples per class. Each image was the result of a crop of an original image in the ILSVRC dataset [20], where the crop was done using accompanying bounding box information. Even this *tiny* version was prohibitively large to compute 1-NN for, therefore we opted to first resize all cropped images to $32\times32\times3$, and instead of computing generalization error on the collective of 200 classes, we evaluated performance on 10 randomly selected pairs[2] and computed average performance in each projected dimension. Running 1-NN on the entire 200-class dataset resulted in generalization performance of around $\%2$ accuracy for data projected using the proposed method; although the proposed method outperformed other methods, the performances were hardly distinguishable, thus we opted the pairwise strategy.

4 Discussion

We first address the surprising result that lower-dimensional representations of the aforementioned datasets often lead to improved generalization performance while having fewer dimensions. A possible explanation for this increase in performance after dimensionality reduction is due to the nature of the datasets, where pixels / features in natural images contain local information that is repeated across neighboring pixels / features. Furthermore, dimensionality reduction is known to make the representations more robust to noise and outliers, potentially leading to improved generalization performance. We note that this trend does not continue to improve performance as we continue to reduce dimensionality further. This is expected because ultimately the dimension is reduced to 1 feature where almost all of the features are discarded and hence we cannot expect superior performance. Results for really low dimensions are not included due to brevity.

We also note that PCA almost always outperforms linear RP, potentially because PCA actively considers the data in its optimization process. Many studies have compared PCA and linear RP. Dasgupta's seminal work [10] studies the tradeoffs between these two methods, demonstrating that although PCA often performs better than linear RP, linear RP enjoys superior time-complexity ($\mathcal{O}(dn)$ for linear RP vs $\mathcal{O}(d^3)$ for PCA). Furthermore, while linear RP can always stably embed the input set into $k=\mathcal{O}(\log n/\varepsilon^2)$ dimensions, PCA can at *worst case* embed into $k=\Omega(n)$ dimensions.

Furthermore recall that PCA focuses on solving a *global* minimization (i.e., min $\sum_{i=1}^{n}||\boldsymbol{x_i}-\boldsymbol{y_i}||^2$ where $\boldsymbol{x_i}$ is the original sample and $\boldsymbol{y_i}$ is the projected sample). This form of optimization does not guarantee that local pairwise distances are preserved. In contrast

[1] To accommodate the theoretical conditions for highly accurate random projections, the embedded space $\mathbb{R}^k$ must satisfy a minimum dimensionality of $\mathcal{O}(\log n/\varepsilon^2)$. This minimum dimensionality is >100 for the number of samples in our experiments.

[2] Sample pairs: (school bus, remote control), (brown bear, german shepherd), (brown bear, lion), (lion, monarch butterfly), (monarch butterfly, steel-arch bridge), $\cdots$

to PCA, random projections have provably converging bounds on the distortion of local pairwise distances between all pairs of samples in the embedding space with respect to the original space.

It is also interesting to see the results for nonlinear maximum RP outperform both linear RP and nonlinear rectified RP at higher dimensions. This difference in performance is observed while [15] suggest similar performance for ReLU and spatial pooling activation functions. We hypothesize that the low performance of nonlinear rectified RP in high dimensions may be because roughly half of the features are being set to zero, while the low performance of nonlinear maximum RP in low dimensions may stem from the relatively tighter distribution of features in the embedded space (i.e., (5)) and the quality of the l_2 norm in high dimensions. This comparison and reasons for varying performances merit further study.

Furthermore, we observe that the performance of linear RP and nonlinear rectified RP is consistent as the projected dimension k decreases from 3072 to 128 and then drops as k drops from 128 to 64. This suggests that we have crossed the theoretical lower bound for highly stable embedding. In contrast to this, the performance of nonlinear maximum RP follows a gradual decrease in performance as the projected dimension k drops from 3072 to 64.

Finally, a surprising observation in our experiments for ensemble of nonlinear maximum RP was its similar performance to that of ensemble of nonlinear minimum RP, as long as we were consistent in applying the max / min functions on each output dimension. This observation will likely be of importance in future work when deriving theoretical probability bounds for nonlinear maximum RP.

5 Conclusion

In this study, we introduced a new method for nonlinear maximum RPs for stable and class-aware embedding of n data samples from $\mathbb{R}^d$ into $\mathbb{R}^k$. Inspired by theoretical work on linear RPs and nonlinear rectified RPs, and following their stipulation surrounding theory for the proposed method, we perform an experimental study showing the stable and superior embedding of nonlinear maximum RPs compared to prior RP methods on 3 different real-world datasets. Furthermore, we compare the performance of the proposed method with PCA, a commonly used dimensionality reduction technique, and show that the proposed method performs comparatively (in the theoretically allowed range for k) while being computationally much more efficient.

In future work, we would like to derive a theory for the probability bounds of the nonlinear embedding of samples into lower dimensions using the max activation function. Specifically, we would like to assert the claim of [15] that this bound is similar for spatial pooling and ReLU activation functions, and to explore the differences and the interplay between these two nonlinearities, and variants thereof. Additionally, our preliminary experiments on multi-layered nonlinear RPs (not included here for brevity) hint at the compounded power of such projections in further boosting performance. Theory in this vein is promising for theoretically backing empirical results observed by [21] and [22] where CNNs with random weights competitively performed on classification tasks.

6 Acknowledgements

This research has been supported by Canada Research Chairs programs, Natural Sciences and Engineering Research Council of Canada (NSERC), and the Ministry of Research and Innovation of Ontario. The authors also thank Nvidia for the GPU hardware used in this study through the Nvidia Hardware Grant Program.

References

[1] B. Raytchev, I. Yoda, and K. Sakaue, "Head pose estimation by nonlinear manifold learning," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 4. IEEE, 2004, pp. 462–466.

[2] I. S. Lim, P. de Heras Ciechomski, S. Sarni, and D. Thalmann, "Planar arrangement of high-dimensional biomedical data sets by isomap coordinates," in *Computer-Based Medical Systems, 2003. Proceedings. 16th IEEE Symposium*. IEEE, 2003, pp. 50–55.

[3] K. Kim, "Face recognition using principle component analysis," in *International Conference on Computer Vision and Pattern Recognition*, 1996, pp. 586–591.

[4] A. Lima, H. Zen, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "On the use of kernel pca for feature extraction in speech recognition," *IEICE TRANSACTIONS on Information and Systems*, vol. 87, no. 12, pp. 2802–2811, 2004.

[5] R. Xu, S. Damelin, and D. C. Wunsch, "Applications of diffusion maps in gene expression data-based cancer diagnosis analysis," in *Engineering in medicine and biology society, 2007. EMBS 2007. 29th annual international conference of the IEEE*. IEEE, 2007, pp. 4613–4616.

[6] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[7] W. S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[8] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[9] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[10] S. Dasgupta, "Experiments with random projection," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 143–151.

[11] W. B. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," *Contemporary mathematics*, vol. 26, no. 189-206, p. 1, 1984.

[12] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of johnson and lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[13] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM, 1998, pp. 604–613.

[14] D. Achlioptas, "Database-friendly random projections: Johnson-lindenstrauss with binary coins," *Journal of computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.

[15] R. Giryes, G. Sapiro, and A. M. Bronstein, "Deep neural networks with random gaussian weights: A universal classification strategy," 2015.

[16] G. Sanguinetti, "Dimensionality reduction of clustered data sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 535–540, 2008.

[17] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, pp. 66–71, 2009.

[18] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[19] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *International conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, 2015.

[21] K. Jarrett, K. Kavukcuoglu, Y. LeCun *et al.*, "What is the best multi-stage architecture for object recognition?" in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 2146–2153.

[22] A. Karimi, M. Shafiee, A. Ghodsi, and A. Wong, "Synthesizing deep neural network architectures using biological synaptic strength distributions," *arXiv preprint arXiv:1707.00081*, 2017.