SIGNAL REPRESENTATIONS IN MODERN SIGNAL PROCESSING

Rebecca Willett

Electrical and Computer Engineering
University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT

The last decade of John Cozzens's tenure at the NSF witnessed the advent of theory and methods at the heart of modern data science. These advances include (but are not limited to) compressed sensing, sparse coding, inference methods robust to outliers and missing data, and convex optimization tools that facilitate a host of novel inference methods. This paper describes how these methods evolved from classical basis representations of signals to alternative, flexible representations of signal structure. These new representations facilitate more accurate and robust inference in many contexts, and research at the intersection of signal processing, machine learning, and optimization make it possible to learn new representations from complex sensor data. This paper explores several key representations that have emerged in the past decade and their impact on the signal processing community.

Index Terms— Signal representations, subspaces, sparsity, neural networks, union of subspaces

1. INTRODUCTION

Signal representations are at the heart of modern signal processing. An accurate and parsimonious model of the information-bearing component of a signal facilitates more accurate estimation of signals from noisy observations, reconstruction of signals in ill-posed inverse problems, recovery of missing data, higher-power detection and classification methods, and higher-rate compression methods.

Classically, the signal processing community represented signals in terms of their Fourier coefficients. While this was a useful model for radio and radar applications, it has significant limitations in many application domains because signals are not everlasting, but rather are transient and can exhibit sudden changes and singularities. Any signal with a sudden change would have a large number of significant Fourier coefficients. This early observation spurred interest in alternative signal representations.

In general, signal processors seek flexible, parsimonious representations of signals – that is, representations with a relatively small number of degrees of freedom capable of representing broad families of signals. To see the importance of both flexibility and parsimony in the context of signal estimation, recall that the mean squared error of any estimate is the sum of its squared bias and its variance. Representations with few degrees of freedom lead to estimators with less sensitivity to noise and hence lower variance, while flexible representations exhibit low bias.

Classical subspace models and other parametric representations are parsimonious but often too inflexible for modern applications. The past decade of signal processing research has focused on representations that leverage sparse, low-rank, and learned signal models that more effectively navigate this tradeoff between parsimony and flexibility. This paper reviews a subset of the major signal representation categories of the past decade.

This work was supported in part by NSF Award CCF-1418976.

2. KNOWN SUBSPACES

Subspace representations arise in a number of settings [1]. Assume we have a true signal of interest, $x^* = \left[x_1^*, \ldots, x_p^*\right]^\top \in \mathbb{R}^p$. Classical subspace representations represent the signal as $x^* = Uv^*$ where $U \in \mathbb{R}^{p \times r}$ is a basis for an r-dimensional subspace and $v^* \in \mathbb{R}^r$ is a vector of r subspace basis coefficients. Typically $r \ll p$; hence, if we know U, estimating or detecting x^* is a function of only r unknowns rather than p.

For example, fitting a polynomial to observed data can be cast as estimation using a subspace model. In particular we could let the subspace basis correspond to a polynomial basis; then finding the best polynomial approximation to a signal amounts to estimating the polynomial basis coefficients (corresponding to the polynomial coefficients). For a second example, Figure 1 shows a visual representation of this idea applied to small patches of an image, where each patch on the left can be represented as a weighted sum of the three representative patches on the right (which span a patch subspace).

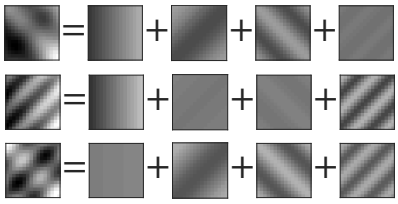


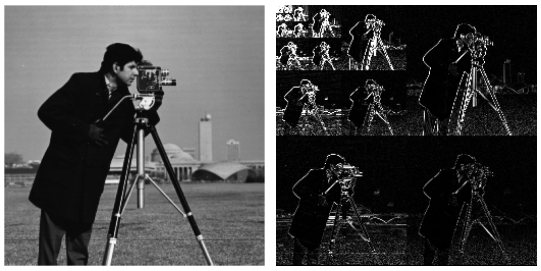
Fig. 1: Image patches represented by a patch subspace. Each patch on the left is a sum of the four patches (which span the patch subspace) on the right. Each column on the right corresponds to the same patch multiplied by a different weight, and different weights result in different patches on the left.

3. SPARSIFYING BASES

The known subspace model described above is too inflexible for many problems. For instance, most photographs do not lie in a single subspace. The advent of wavelet-based representations of signals (*cf.* [2]) allowed for much more flexible models. In particular, researchers noted that many signals (such as photographs) could be represented as a weighted sum of a small number of wavelet basis functions. On the surface, this may appear equivalent to a subspace representation, but the key difference here is that the wavelet subspace could be detected automatically from the data instead of being

fixed ahead of time. In particular, one could compute the wavelet coefficients of a signal and examine the magnitude of all the wavelet coefficients to determine which subset of wavelet basis functions generated a subspace containing most of a signal's energy.

More specifically, we can represent a signal as $x^* = A\theta^*$, where $A \in \mathbb{R}^{p \times p}$ is a p-dimensional wavelet basis and θ^* is the vector of wavelet coefficients. Assume only $s \ll p$ elements of θ^* are non-zeros. If we knew that the indicies of the s non-zero elements of θ^* formed a set $S^* \subseteq \{1, \ldots, m\}$, we could let $A_{S^*} \in \mathbb{R}^{p \times s}$ be a matrix corresponding to the s columns of A with column indices in S^*, and similarly let $\theta^*_S \in \mathbb{R}^s$ be subset of elements of θ^* corresponding to the indices in S^*. Then $x^* = A_S \theta^*_S$ and our model is equivalent to the subspace model described above. However, in general we do not know the set S, so methods leveraging the sparsity of θ^* essentially learn both the subspace containing θ^* and its subspace coefficients.



(a) (b)

Fig. 2: Sparse wavelet representation of an image. (a) Original cameraman image. (b) Log-magnitudes of wavelet basis coefficients of cameraman image. Note that many of the coefficients are zero-valued or close to zero-valued.

Put another way, many signals are *sparse* in a wavelet basis – that is, most of the wavelet coefficients are close to zero-valued. Note the image in Figure 2, showing the cameraman image and the magnitudes of its wavelet coefficients. Notions of sparse representations of signals extend well beyond the wavelet basis. In fact, significant research has been devoted to developing specific bases which allowed for sparse representations of broad ranges of signals for different applications (*cf.,* [3, 4, 5]). The central theme of sparse basis representations is that we only need a weighted sum of a small number of basis functions from a larger collection to accurately represent a signal. In the context of estimating a signal from noisy observations and trying to navigate the bias-variance trade-off described in the introduction, sparse basis representations achieve about the same variance as that associated with subspace models, but significantly lower bias for broad families of signals.

4. VARIATIONAL REPRESENTATIONS

Variational representations of signals classically model a signal as a partial differential equation and use the calculus of variations to compute properties of the signal [6]. Particularly well-known examples include the Mumford-Shah model [7] and total variation [8]. Total-variation in particular has experienced a resurgence of attention in the past decade with the development of new optimization algorithms which facilitate finding a signal which is both a good fit to data and which has low total variation (*cf.,* [9]).

The total variation seminorm [6, 9] measures how much a signal or image varies across pixels, so that a highly variable or noisy signal has a large TV seminorm,[1] while a smooth or piecewise smooth signal would have a relatively small TV seminorm.

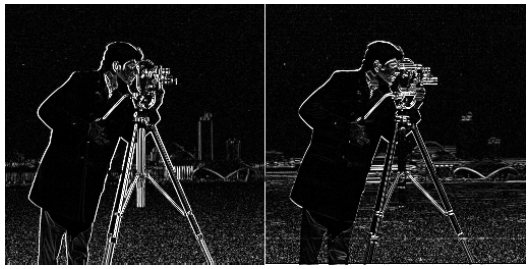


Fig. 3: Magnitude of horizontal and vertical first-order differences in cameraman image. The sum of these magnitudes is the total variation of the image.

The (anisotropic) total variation seminorm for an image is defined as

$$\|x\|_{\mathrm{TV}} \triangleq \sum_{k=1}^{\sqrt{n}-1} \sum_{l=1}^{\sqrt{n}} |x_{k,l} - x_{k+1,l}| + \sum_{k=1}^{\sqrt{n}} \sum_{l=1}^{\sqrt{n}-1} |x_{k,l} - x_{k,l+1}|,$$

where we slightly abuse of notation by using 2D pixel indices instead of vector indices by assuming that $x \in \mathbb{R}^n$ is a square $\sqrt{n} \times \sqrt{n}$ image. This highlights the fact that the TV seminorm is simply a measure of the magnitude of all vertical and horizontal first-order differences, which are illustrated in Figure 3. This property makes TV especially well-suited for image denoising and inverse problems. For instance, if y is a noisy realization of x^*, we might denoise y by solving the optimization problem

$$\hat{x} = \arg\min_x \|y - x\|_2^2 + \lambda \|x\|_{\mathrm{TV}},$$

where $\lambda > 0$ is a tuning parameter [9].

5. MANIFOLDS AND UNIONS OF SUBSPACES

The subspace model described above can be generalized not only using sparse models, but also by using manifold and union of subspaces models. The manifold representation is useful in a number of contexts (*cf.,* [10, 11]). Unlike subspaces, manifolds exhibit curvature; however, if the curvature is not too strong then the manifold can be approximated by a union of shifted subspaces or hyperplanes [12]. In this context, the best subspace representation of a signal depends on the signal itself, somewhat analogously to the signal-dependent subspace corresponding to sparse signal representations.

Manifold and union of subspace models are widespread in image processing (*cf.,* [13, 14]). Figure 4 illustrates how the set of 8×8 patches in an image have the potential to be represented with fewer than 64 degrees of freedom. For instance, methods like non-local means essentially extract all the patches from an image and models those patches as lying near a low-dimensional manifold, and estimate noise-free patches by projecting noisy patches onto locally linear approximations of that manifold [13]. A cartoon illustration of patches lying in a union of subspaces is shown in Figure 5.

To better understand the flexibility and parsimony of a union of subspaces representation, consider K rank-r subspaces lying in a p-dimensional ambient space. Specifically, given a collection of

[1] A seminorm is a norm that can have zero value for non-zero vectors.

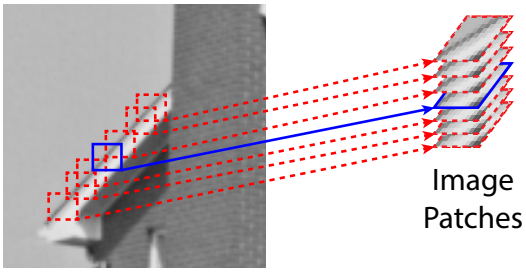


Fig. 4: Image patch space. The set of all overlapping patches in an image can exhibit structure well-represented by manifolds, unions of subspaces, or redundant dictionaries.

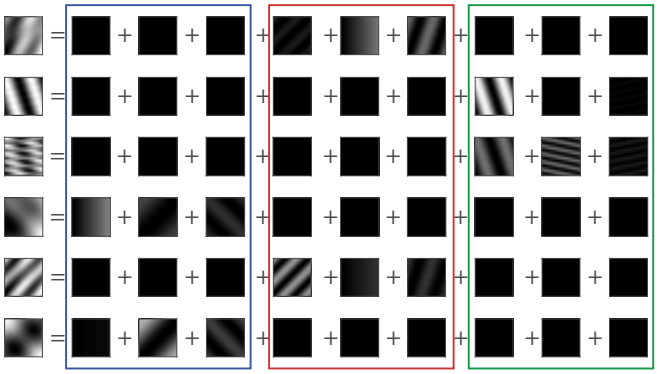


Fig. 5: Image patches represented by a union of three subspaces. The first three patches after the equal sign (blue box) span the first subspace, the second three (red box) span the second subspace, and the final three (green box) span the final subspace. Each patch on the left is a weighted sum of patches, as before, but *only patches from one of the three subspaces*. Black patches correspond to patches with zero weight assigned to them. The first and fifth rows correspond to the second subspace, the second and third rows to the third subspace, and the fourth and sixth rows to the first subspace.

n "training signals" $x_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$, one can learn a union of subspaces that minimizes the sum of squared distances from the training samples to the representation using an alternating minimization algorithm (*cf.,* [15] and references therein). The subspace bases correspond to Krp degrees of freedom and the subspace basis coefficients correspond to $nr \log K$ degrees of freedom for n image patches. In contrast, these K subspaces all lie within a single rank-Kr subspace, so one might consider using principal components analysis to simply estimate this subspace from the n training samples. While the basis for this subspace has Krp degrees of freedom (the same as for the union of subspaces), the subspace basis coefficients correspond to nrK degrees of freedom for n image patches. Thus the single subspace model with the same representation flexibility (and hence same bias) has significantly more degrees of freedom (and hence more variance).

6. REDUNDANT DICTIONARIES

We saw earlier that sparse representations of signals in a particular basis yield an effective trade-off between bias and variance for large families of signals. However, there are some instances where having a large redundant dictionary of representative signals (*i.e.,* a set of a representative signals that are linearly dependent upon one another) can lead to even better representations of signals.

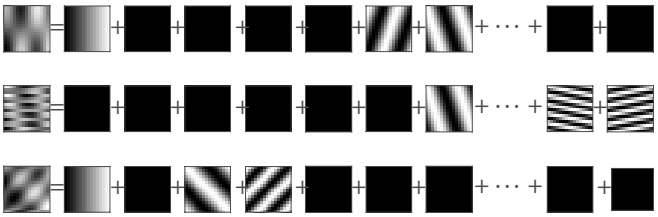


Fig. 6: Image patches represented by a redundant dictionary. Each patch on the left is a weighted sum of a sparse subset of the many patches on the right. The patches on the right may be linearly dependent. Different weights result in different patches. Black patches correspond to patches with zero weight assigned to them.

In general, a large redundant dictionary corresponds to a large number of degrees of freedom; hence learning or leveraging this representation can be sensitive to the number of training samples available or to noise. To sidestep this challenge researchers have considered a variety of mechanisms. Early work considered concatenating a small number of different bases (*e.g.,* a wavelet and Fourier basis together can be used to represent – and separate – edges and texture information [16]). More recent research has focused on dictionary learning [17, 18, 19], in which the dictionary may be highly redundant but each signal can only be represented by a linear combination of a sparse subset of the dictionary elements; Figure 6 illustrates this concept.

Fig. 7: Learned dictionary of patches using boat image as training dataset [20].

Specifically, given a collection of n "training signals" x_i, for $i = 1, \ldots, n$, one can learn a dictionary facilitating sparse representations by solving the following optimization problem:

$$(\hat{D}, \hat{A}) = \arg\min_{D,A} \sum_{i=1}^{n} \|x_i - DA_i\|_2^2 + \lambda \|A_i\|_1 \text{ subject to } \|D\|_F \leq 1$$

where A_i is the i^{th} column of A, $\hat{D}$ is the learned dictionary, and $\hat{A}$ is the collection of sparse weights for the n training signals. Each column of $\hat{D}$ corresponds to one entry in the learned dictionary. In [20], the authors learned a dictionary when the training signals corresponded to small patches of pixels in the "boats" image; the original image and learned dictionary elements are displayed in Figure 7.

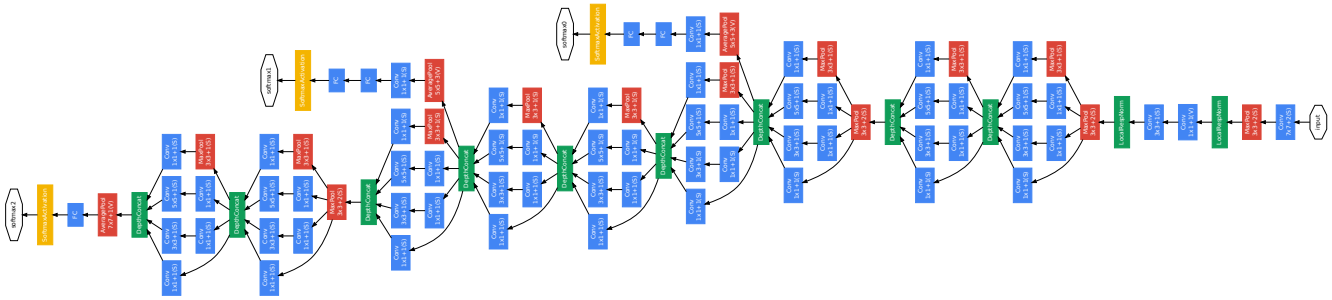


Fig. 8: Architecture of Google Inception deep neural network [21].

7. (DEEP) NEURAL NETWORKS

Many recent representations of signals have been computed using deep neural networks. In contrast to the representations considered earlier in this paper, very little is known about the space of signals that can be represented by a deep neural network with a particular architecture, or how much bias and variance is associated with estimators or classifiers based on the corresponding representations. Despite the current lack of theoretical understanding, deep neural networks have proven effective in a variety of classification and labeling tasks, especially for images.

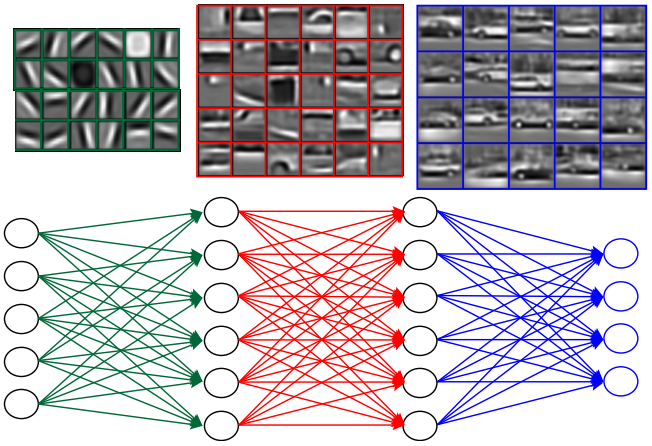


Fig. 9: An illustration of a neural network with two hidden layers. The output of the first layer is highlighted in green, the second layer in red, and the third layer in blue. Learned dictionary elements from [22].

Neural networks, as illustrated in Figure 9, typically have an input layer, and output layer, and one or more intermediate or hidden layers. Each node corresponds to a nonlinear operation that is applied to the incoming representations to produce the node output; for instance, a node might compute a weighted sum of the inputs and then threshold the result. The values of the weights and thresholds used by these nodes must be learned from training data using training algorithms that leverage stochastic gradient descent and a number of engineering innovations. The goal of the training is to learn an effective collection of weights and thresholds so that the resulting mapping from input to output nodes closely resembles the desired classification or encoding. As a result, the outputs of many of the hidden or intermediate nodes in this network correspond to features associated with the signals and hence can be considered a signal representation, as illustrated in Figure 9 [22]. A network architecture developed by Google for image classification and labeling is shown in Figure 8 [21].

Deep neural networks have received widespread attention recently because of their efficacy on a number of image recognition tasks. The core ideas underlying neural networks was first developed in the 1990s, and have seen such a strong recent resurgence. One reason for this comeback is the availability of much richer and larger training data sets, such as Google's ImageNet [23] which contains $O(10^8)$ images. Another reason is the availability of high throughput computing tools that allow computers to train these deep neural networks from a large number of different starting conditions with a large number of training samples in a relatively small amount of time. Despite these successes, the utility of deep neural networks for settings in which much smaller collections of training data are available remains an open question.

8. CONCLUSIONS

This paper has reviewed various signal representations that have evolved over the past decade, often through research supported by NSF programs managed by John Cozzens. As these representations have been developed over the past decade, they have led to significant advances in our ability to perform signal estimation, reconstruction, classification, and recognition. The conventional wisdom that parsimonious representations reduce sensitivity to noise, missing data, and high compression ratios has played a significant role in these developments, aided by novel computational methods for learning nonlinear representations from real data that do not restrict signal processors to linear methods with closed-form expressions. Cozzens' support of this research has led to a new era of signal representation for modern data science and signal processing.

9. REFERENCES

[1] L. L. Scharf, *Statistical signal processing*, vol. 98, Addison-Wesley Reading, MA, 1991.

[2] S. Mallat, *A wavelet tour of signal processing*, Elsevier/Academic Press, Amsterdam, 2009, The sparse way, With contributions from Gabriel Peyré.

[3] E. J. Candes and D. L. Donoho, "Curvelets: A surprisingly effective nonadaptive representation for objects with edges," Tech. Rep., DTIC Document, 2000.

[4] R. M. Willett and R. D. Nowak, "Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging," *IEEE Transactions on Medical Imaging*, vol. 22, no. 3, pp. 332–350, 2003.

[5] D. Labate, W.-Q. Lim, G. Kutyniok, and G Weiss, "Sparse multidimensional representation using shearlets," in *Optics & Photonics 2005*. International Society for Optics and Photonics, 2005, pp. 59140U–59140U.

[6] T. Chan and J. Shen, *Image Processing And Analysis: Variational, PDE, Wavelet, And Stochastic Methods*, Society for Industrial and Applied Mathematics, 2005.

[7] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Communications on pure and applied mathematics*, vol. 42, no. 5, pp. 577–685, 1989.

[8] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[9] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Transactions Image Processing*, vol. 18, no. 11, pp. 2419–34, 2009.

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

[11] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering," *Advances in neural information processing systems*, vol. 16, pp. 177–184, 2004.

[12] W. K. Allard, G. Chen, and M. Maggioni, "Multi-scale geometric methods for data sets ii: Geometric multi-resolution analysis," *Applied and Computational Harmonic Analysis*, vol. 32, no. 3, pp. 435–462, 2012.

[13] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, vol. 2, pp. 60–65.

[14] J. Salmon, Z. Harmany, C. Deledalle, and R. Willett, "Poisson noise reduction with non-local PCA," *Journal of Mathematical Imaging and Vision*, vol. 48, no. 2, pp. 279–294, 2014, arXiv:1206:0338.

[15] D. Pimentel-Alarcón, L. Balzano, R. Marcia, R. Nowak, and R. Willett, "Group-sparse subspace clustering with missing data," in *2016 IEEE Statistical Signal Processing Workshop*, 2016.

[16] J.-L. Starck, Y. Moudden, J. Bobin, M. Elad, and D. L. Donoho, "Morphological component analysis," in *Optics & Photonics 2005*. International Society for Optics and Photonics, 2005, pp. 59140Q–59140Q.

[17] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.

[18] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[19] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, "Supervised dictionary learning," in *Advances in neural information processing systems*, 2009, pp. 1033–1040.

[20] F. Bach et al., "ICCV tutorial on sparse coding and dictionary learning for image analysis," ICCV, 2009, `h3p://lear.inrialpes.fr/people/mairal/tutorial_iccv09/tuto_part2.pdf`.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[22] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 609–616.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.