**Summary**

Sparse coding networks (SCN) are an effective method for computer vision, however are on efficient on a restricted number of dataset. Typically, a dataset is not compatible with an SCN when that dataset exhibits natural noise, such as variations in object orientation. Specifically, the linear representation of sparse coding is unable to capture the nonlinear variations of the data set. As such, an SCN networks poor performance on large data sets, where they're common variabilities and anomalies. However, Deep Neural Networks rely on large data sets, as they thrive on their ability to extract essential features of an object from an abundant diverse set of examples. As such the idea behind this paper is to develop a deep sparse coding networks that can thrive conditions that SCN networks do not.  
 This paper presents a novel deep sparse coding network that relies on nonnegative sparse coding and multilevel optimization. The purpose being able to efficiently extend the conventional sparse coding to multilayer architectures in order to extend the networks learning capacity. Multilayer sparse coding networks are useful for this problem as combines the advantages of both deep-layer architectures and sparse coding networks Deep architectures are useful as they ameliorate any information degeneracy, meanwhile sparse coding networks contain a more intuitive interpretation of the learned features. As such this paper uses a bottleneck module, which uses two specialized sparse coding layers: one with a wide dictionary, and another with a slim dictionary in order to reduce the number of learnable parameters without loss of performance. Furthermore, both the dictionaries and regularization parameter of the  SCN are optimized using an end-to-end supervised learning algorithm based on multilevel optimization.

**Related Work** While the deep sparse coding network this is a novel idea, it was developed by using approached documented in other papers. 

*Deep Neural Networks with wide shapes* Wide networks [12,13,14] are proposed in order to take advantage of the high dimensional latent features, thus having a comparable performance to a slim deep network while using less layers. However, a wide network contains a significantly larger number[10,11,15] of learnable parameters compared to its slim network counterpart. As such, this paper focuses on regularizing the wide SNC via dimensionality reduction and clustering using nonnegative sparse coding-based approach, in order to optimize the number of learnable parameters.   
*Dimensionality reduction and clustering in deep neural networks*   
 The purpose of a bottleneck shaped neural network[10,14] is to reduce the dimensions of the incoming signal this will in turn reduce an overfitting caused by the network. Furthermore, for this problem dimensionality reduction with non-negative sparse coding is equivalent to clustering [16], thus low dimensional hidden features are weighted cluster indicators. The approach proposed in this paper, is to simultaneously learn high dimension discriminative representation and low dimensional clustered features into a single network architecture with end-to-end supervised learning. Alternative approaches employ a deep semi-nonnegative matrix factorization[17] that trains a hierarchical network with the reconstruction loss.   
*Multi-layer Sparse Coding*Multi-layer are typically implemented with a purpose of training the reconstructive dictionaries of each layer in a greedily layer-wise fashion [7,18]. Alternatively, another approach is unfold unfold and approximated the sparse coding process with deep neural networks [5,9]. The approach proposed in this paper differs from these methods, such that they extend the conventional nonnegative sparse coding to a multilayer architecture without unfolding, as such the model size does not increase with each iteration. Furthermore, this approach allows for higher compatibility of off the shelf sparse recover algorithms, without changing the network architecture. Additionally this network, trains the network in an unsupervised, semi-supervised, or supervised method by manually balancing the discriminative and reconstructive loss. This method was inspired by stacked auto-encoders [19] and CNN-based model with auxiliary reconstruction loss [20,21].  
*Supervised Dictionary Learning* The purpose of Supervised Dictionary Learning is to strengthen the power of sparse codes by exploiting the labeled samples. Previous methods relied on applying a fixed point differentiation and bilevel optimization, a supervised dictionary learning scheme for the shallow sparse coding [23,24]. This paper generalizes the single-layer supervised dictionary learning to multilayer network based on multilevel optimization.

**Datasets/ Experiments**

The proposed network was tested using the CIFAR-10, CIFAR-100, STL-10 and MNIST image datasets. This testing demonstrates that the proposed 14-layer SCN model out performs a 164 layer and 1001 layer deep residual network. The architecture resembles ResNet, and contains 7 bottlenecks modules which includes 14 sparse coding layers that are divided into 3 sections. Furthermore, a spatial subsampling with a factor of 2 is applied to the last two bottleneck modules, and following the final sparse coding layer a global spatial average pooling layer and a fully connected layer, which serves as the linear classier, are added. For training, all data except for MNIST are data augmented and pre-processed with random horizontal flipping and random translations. For CIFAR dataset, the image is translated up to 4 pixels, and up to 12 pixels for STL-10. Furthermore, both training and testing data sets are preprocessed with per-pixel-mean subtractions. When doing baseline comparison methods, the proposed SCN is compared with the multilayer sparsity regularized coding (OMP) and nonnegative sparse coding (NOMP), the supervised convolution kernel networks (SCKN), and scattering networks (ScatNet). The deep neural networks are compared to the residual network (ResNet), wide residual network (WRN), and swap out networks (SwapOut). The training the SCN-4 model for: CIFAR, -10 and -100, dataset took around 26 hours, STL-10 dataset took around 21 hours and MNIST dataset took around 3 hours. Testing the SCN-4 model on : CIFAR, -10 and -100, dataset took 9 seconds STL-10 dataset took 65 seconds and MNIST dataset took 7 seconds.

*CIFAR-10,CIFAR-100* The most extensive experiment was conducted on the CIFAR-10 and CIFAR-100 datasets, which consisted 60,000, , color images, which were split into 50,000 training samples and 10,000 samples. The CIFAR-10 and CIFAR-100 contains identically set of images, however CIFAR-10 images are split into 10 classes, while CIFAR-100 is spit into 100 classes. In CIFAR-10 dataset, each class has 5000 training images and 1000 testing images, while the CIFAR-100 has 500 training images and 100 images per class thus making it harder to classify.   
 The proposed network, SCN-4 with 15 layers, reported a classification error of 5.81% for the CIFAR-10 dataset, and 19.93% for CIFAR-100 dataset. Comparative models such as the OMP, two layers, NOMP, four layers, and SCKN, ten layers, reported a classification error of 18.50%, 18.60% and 10.20% respectively. Furthermore, for CIFAR-100 testing, the NOMP model reported a classification error of 39.92%, the OMP and SCKN model was not tested on the CIFAR-100 dataset. Additionally, the SCN-4 networks has comparable results to the ResNet and SwapOut for both CIFAR-10 and CIFAR-100 datasets while using less layers. The ResNet-110, ResNet-1001, and ResNet-29 models, which have 110,1001 and 29 layers respectively, reported a classification error of 6.41%, 4.92% and 3.58% respectively for CIFAR-10 and a classification error of 27.22%, 27.21% and 17.31% respectively for CIFAR-100. The SwapOut-20 and SwapOut-32, which have 20 layers and 32 layers respectively, reported a classification error of 5.68% and 4.76% respectively for CIFAR-10 and a classification error of 25.86% and 22.72% respectively for CIFAR-100.   
*STL-10* The STL-10 dataset contains 13,000, , images that were split into 5,000 training samples and 8,000 testing images. As such when testing the SCN-4 network against the STL-10 dataset, it out preformed comparative models. The SCN-4 network had a classification accuracy of 83.11% while comparative model SWWAE, and Deep-Ten had a classification accuracy of 74.33% and 76.29% respectively.

*MNIST* The MNIST data set containers 70,000, , images of digits, which were split into 60,000 training samples and 10,00 testing samples. Once again, the SCN-4 network out preformed competing models for the MNIST dataset. The SCN-4 dataset had a classification error of 0.36% while the comparative models of CKN, SCTNET , PCANet,S-SC, TDDL had a classification error of 0.39%,0.43%,0.62%,0.84% and 0.54% respectively.

**Bibliography:**

[5] K. Gregor and Y. LeCun. Learning fast approximations of sparse coding. In *ICML*, 2010. [9] T. Moreau and J. Bruna. Understanding trainable sparse coding via matrix factorization. In *ICLR*, 2017.   
[10]  K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.   
[11]  K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016.   
[12]  S. Zagoruyko and N. Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 201 6.  
[13]  S. Singh, D. Hoiem, and D. A. Forsyth. Swapout: Learning an ensemble of deep architectures. In *NIPS*,2016.  
[14]  S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. *arXiv preprint arXiv:1611.05431*, 2016.  
[15]  F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet- level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360, 2016.   
[16] C. H. Q. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE TPAMI*, 32(1):45–55, 2010.   
[17]  G. Trigeorgis, K. Bousmalis, S. Zafeiriou, and B. Schuller. A deep semi-NMF model for learning hidden representations. In *ICML*, 2014.   
[18]  A. Coates and A. Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *ICML*, 2011.   
[19]P.Vincent,H.Larochelle,I.Lajoie,Y.Bengio,andP.Manzagol.Stackeddenoisingautoenco ders:Learning useful representations in a deep network with a local denoising criterion. In *JMLR*, 2010.   
[20]  J Zhao, M. Mathieu, R. Goroshin, and Y. LeCun. Stacked what-where auto-encoders. *CoRR*, abs/1506.02351, 2015.   
[21]  Y. Zhang, K. Lee, and H. Lee. Augmenting supervised neural networks with unsupervised objectives for large-scale image classification. *CoRR*, abs/1606.06582, 2016.   
[22]  J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE TPAMI*, 34(4):791–804, 2012.   
[23]  J. Yang, K. Yu, and T. Huang. Supervised translation-invariant sparse coding. *in CVPR*, Jun. 2010.   
[24]  J. Yang, Z. Wang, Z. Lin, X. Shu, and T. Huang. Bilevel sparse coding for coupled feature spaces. In *CVPR*, 2012.