DEEP MULTIMODAL SPARSE REPRESENTATION-BASED CLASSIFICATION

Mahdi Abavisani† *Vishal M. Patel*⋆

† Rutgers University, Electrical and Computer Engineering, NJ, USA
⋆Johns Hopkins University, Electrical and Computer Engineering, MD,USA

ABSTRACT

In this paper, we present a deep sparse representation based fusion method for classifying multimodal signals. Our proposed model consists of multimodal encoders and decoders with a shared fully-connected layer. The multimodal encoders learn separate latent space features for each modality. The latent space features are trained to be discriminative and suitable for sparse representation. The shared fully-connected layer serves as a common sparse coefficient matrix that can simultaneously reconstruct all the latent space features from different modalities. We employ discriminator heads to make the latent features discriminative. The reconstructed latent space features are then fed to the multimodal decoders to reconstruct the multimodal signals. We introduce a new classification rule by using the sparse coefficient matrix along with the predictions of the discriminator heads. Experimental results on various multimodal datasets show the effectiveness of our method.

***Index Terms*—** Sparse representation, multimodal sparse representation, multimodal sparse representation classification, deep multimodal sparse representation classification

1. INTRODUCTION

Sparse representation is an established technique in signal and image processing with various applications [1, 2, 3, 4]. Among the many applications, Sparse Representation Classification (SRC) methods exploit the discriminative nature of sparse codes and provide robust classification models less sensitive to non-constant variability, outliers, and small data sets [1, 5, 6, 7]. The SRC method uses a sparsity-promoting optimization problem to represent an unlabeled test sample as a sparse linear combination of labeled training samples. This representation is then used in assigning a label to the test sample based on the minimum reconstruction error rule [1]. Various SRC-based methods have been proposed in the literature. These methods include linear models in various applications [5, 6], kernel trick-based nonlinear models [8, 9, 10], and a recent deep neural network-based SRC method [11] (DSRC) that finds an explicit nonlinear mapping for data, while simultaneously obtaining sparse codes that can be used for classification.

Many real-world phenomena involve multiple modalities. Learning from multimodal sources offers the opportunity to gain an in-depth understanding of the phenomena by integrating the complementary information provided in different modalities [12, 13]. In multimodal learning, the model receives the data from multiple modalities and learns to fuse them. The information from different modalities can be fused at feature level (i.e., early fusion), decision level (i.e., late fusion), or intermediately [12, 13, 14].

In this paper, we propose a multimodal deep SRC-based method. We enforce the different modalities to interact through the sparse coefficients of their latent space features. Our deep networks learn the latent space features of different modalities through an autoencoder framework. Our framework encourages different modalities to learn latent features that are discriminative, suitable for sparse coding, and lie in mutual subspaces. The latent space features for the test samples are reconstructed by a linear combination of training samples with a sparse coefficient matrix that is shared among all the modalities.

Sharing the coefficient matrix among all the modalities pushes the test sample to interact with the training samples of all the modalities simultaneously. Therefore a more reliable coefficient matrix, which is calculated by the complementary information from different modalities, is constructed. Since the labels for the training samples are available, we use extra supervision based on labels to develop discriminative latent features. This extra supervision is employed by discriminator heads that are connected to latent space features of different modalities and are trained by a classification loss. At the test time, we combine the prediction of discriminator heads with the minimum reconstruction error rule, and introduce a new classification rule to assign labels to the test samples.

2. RELATED WORK

Sparse Representation-based Classification: In the SRC task, we are given a set of labeled training samples, and the goal is to classify an unseen set of testing samples. Suppose that we collect all the vectorized training samples in the matrix $\mathbf{X}_{\text{train}} \in \mathbb{R}^{d_0 \times n}$, where d_0 is the dimensional size of each sample, and n is the number of training samples.

SRC is based on the assumption that an observed sample $\mathbf{x}_{\text{test}} \in \mathbb{R}^{d_0}$ can be well approximated by a linear combination of samples in $\mathbf{X}_{\text{train}}$ that share the same class label as $\mathbf{x}_{\text{test}}$. Therefore, one can predict class label of a given unseen data such as $\mathbf{x}_{\text{test}}$ by finding the set of few samples in the training set that can better approximate $\mathbf{x}_{\text{test}}$. These samples can be picked out by solving the following optimization problem.

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0 \;\; \text{s.t.} \;\; \mathbf{x}_{\text{test}} = \mathbf{X}_{\text{train}}\boldsymbol{\alpha}. \tag{1}$$

where $\|\boldsymbol{\alpha}\|_0$ counts the number of non-zero elements in $\boldsymbol{\alpha}$. In practice, the ℓ_0 norm is replaced by the ℓ_1 norm [15, 16]. Thus, in the case of noisy obseervations, the SRC solves the following sparsity-promoting problem

$$\min_{\boldsymbol{\alpha}} \|\mathbf{x}_{\text{test}} - \mathbf{X}_{\text{train}}\boldsymbol{\alpha}\|_F^2 + \lambda_0\|\boldsymbol{\alpha}\|_1 \tag{2}$$

where λ_0 is a positive regularization parameter.

The solution, $\boldsymbol{\alpha}$, is used in the minimum reconstruction rule to estimate the class label of $\mathbf{x}_{\text{test}}$ as follows

$$\text{class}(\mathbf{x}_{\text{test}}) = \arg\min_{k} \|\mathbf{x}_{\text{test}} - \mathbf{X}_{\text{train}}\Gamma_k(\boldsymbol{\alpha})\|_F^2 \tag{3}$$

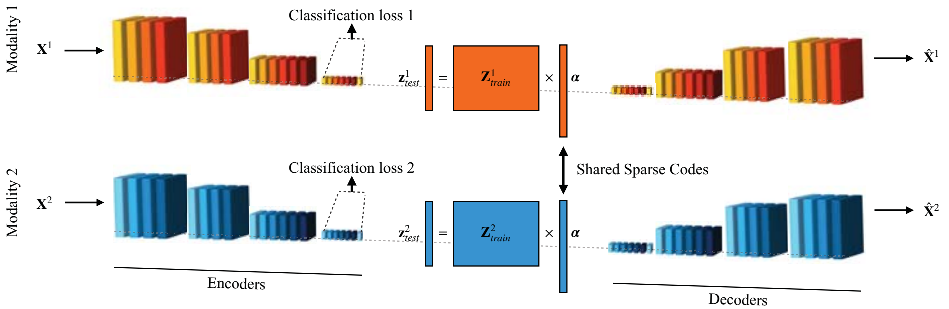


Fig. 1. An overview of the proposed deep multimodal sparse representation-based classification network in a two-modality task. Features of different modalities are fed to their corresponding encoder, where a discriminative criterion is enforced to develop discriminative latent features that are especially suitable for jointly sparse representation. The latent features of different modalities are reconstructed by optimal joint sparse codes and are fed to decoders to reconstruct the raw modality features. The optimal joint sparse codes, along with the predictions of discriminator heads are exploited to predict the class labels of test samples.

where $\Gamma_k(\cdot)$ is the matrix indicator function defined by setting all the rows but those corresponding to the ith class equal to zero.

Linear Multimodal Sparse Representation-based Classification: A number of multimodal extensions for the linear SRC problem have been proposed in the literature [5, 17, 18, 19, 20, 21]. Among those, the methods proposed in [5, 19, 21] are the closest to our model. They impose joint sparsities within and across different modalities. This way, the correlations and coupling the information among modalities are simultaneously taken into account.

Assume the given multimodal data is observed in M modalities, each with n training samples. For each modality $m = 1, 2, \cdots, M$, let's denote $\mathbf{X}^m_{\text{train}} \in \mathbb{R}^{d_m \times n}$ as the dictionary of training samples in m-th modality where d_m is the feature dimension of data in m-th modality. Similarly, the representation of a multimodal test sample in the m-th modality can be represented as $\mathbf{x}^m_{\text{test}} \in \mathbb{R}^{d_m}$.

The SRC model can be applied to the individual modalities. In other words, $\mathbf{x}^m_{\text{test}}$ can be reconstructed by a linear combination of a few atoms in the dictionary $\mathbf{X}^m_{\text{train}}$. Thus, we have

$$\mathbf{x}^m_{\text{test}} = \mathbf{X}^m_{\text{train}}\boldsymbol{\alpha}^m + \mathbf{N}^m, \tag{4}$$

where $\boldsymbol{\alpha}^m \in \mathbb{R}^n$ is a sparse coefficient vector, and $\mathbf{N}^m \in \mathbb{R}^{d_m \times n}$ is the noise matrix. The joint sparsity model argues that $\boldsymbol{\alpha}^m$ has the same sparsity pattern across the different modalities for $m = 1, 2, \cdots, M$. In other words, the matrix $\mathbf{A} = [\boldsymbol{\alpha}^1, \cdots, \boldsymbol{\alpha}^M]$ formed by concatenating the coefficient vectors of an observation across different modalities has the same non-zero rows in its different columns. The matrix $\mathbf{A}$ can be found by the following ℓ_1/ℓ_q-regularized least square problem

$$\mathbf{A} = \arg\min_{\mathbf{A}} \frac{1}{2}\sum_{m=1}^{M} \|\mathbf{x}^m_{\text{test}} - \mathbf{X}^m_{\text{train}}\boldsymbol{\alpha}^m\|_F^2 + \lambda_0\|\mathbf{A}\|_{1,q}. \tag{5}$$

where $q > 1$. Here, $\|\mathbf{A}\|_{1,q}$ is a norm defined as $\|\mathbf{A}\|_{1,q} = \sum_{k=1}^{n}\|\gamma^j\|_q$, where γ^j's are the rows in $\mathbf{A}$.

Once $\mathbf{A}$ is found, similar to problem (3), one can predict the class label of the test observation by solving the following problem

$$\min_k \|\mathbf{x}^m_{\text{test}} - \mathbf{X}^m_{\text{train}}\Gamma_k(\boldsymbol{\alpha}^m)\|_F^2 \tag{6}$$

3. DEEP MULTIMODAL SPARSE REPRESENTATION-BASED CLASSIFICATION NETWORKS

Joint sparse representation-based classification methods [5, 19, 21] are able to extend the SRC model to a model that incorporates multiple modalities while keeping the benefits of sparse representation such as being less sensitive to outliers, and small data sets. However, they still rely on the assumption that the data points across different modalities show linear similarities within samples of the same class. This provides a strong motivation to incorporate deep neural networks to capture complex underlying structures of data across different modalities. We bridge multimodal SRC models and deep neural networks by proposing a transductive multimodal classification model based on deep sparse representation. A transductive model is a model in which both training and test sets are observed, and the learning process pursues reasoning from the specific training samples to a specific set of test cases [22].

We use stacked multimodal autoencoders to exploit the nonlinear relations between the data points. In particular, we have a set of encoder and decoder per each available modality. The encoders and decoders are trained together to find latent space features that are discriminative, lie into a union of linear subspaces, and are constructed with the integrated information from all the available modalities.

To meet these properties, the autoencoder in each modality is trained according to both the training labels and the underlying structures of data in other modalities. The autoencoders of different modalities interact with each other by invoking the same linear relation between data points of different modalities in the training process. The same linear relation is imposed by sharing the same sparse codes in a sparsity-promoting reconstruction loss. As will be described in detail, the sparse codes can be modeled with a fully-connected layer in the deep neural networks framework. Figure 1 shows an overview of our framework.

Thus, the training objective of our model can be divided into reconstruction criteria and discriminative criterion.

Reconstruction Criteria: Reconstruction criteria itself consists of the reconstruction criterion in sparse coding and the reconstruction constraint for autoencoders.

Let $\{\mathbf{X}^m_{\text{train}}\}_{m=1}^M$ and $\{\mathbf{X}^m_{\text{test}}\}_{m=1}^M$ be the given set of training and test samples across the different modalities. We concatenate the training and test samples of each modality and construct the input matrices. The input matrix of m-th modality is denoted by $\mathbf{X}^m \in \mathbb{R}^{d_m \times n}$, and is constructed by the concatenation of $\mathbf{X}^m_{\text{train}} \in \mathbb{R}^{d_m \times n_{\text{train}}}$ and $\mathbf{X}^m_{\text{test}} \in \mathbb{R}^{d_m \times n_{\text{test}}}$, which are respectively the available training and testing samples in the m-th modality.

We feed $\mathbf{X}^m$ to its corresponding encoder and develop the embedding features $\mathbf{Z}^m$. The matrix $\mathbf{Z}^m$ consists of two types of embedding features. Those that are associated with the training samples and those that are corresponded to the testing samples. We respectively indicate to them with $\mathbf{Z}^m_{\text{train}}$ and $\mathbf{Z}^m_{\text{test}}$.

The SRC problem can be employed in the embedding feature space of each individual modality. However, if we couple information among different modalities, richer representations can be learned. We propose to tie the embedding features of different modalities by enforcing them to share the same sparse code solutions in the SRC problem across the embedding space of all the different modalities. This way, the complementary information across different modalities are integrated without imposing an extra burden on the networks for explicitly learning to represent a joint representation.

Thus, we propose to find common sparse codes by solving the following optimization problem

$$\mathbf{A}_c = \arg\min_{\mathbf{A}_c} \frac{1}{2} \sum_{m=1}^{M} \|\mathbf{Z}^m_{\text{test}} - \mathbf{Z}^m_{\text{train}}\mathbf{A}_c\|_F^2 + \lambda_0\|\mathbf{A}_c\|_1, \quad (7)$$

where $\mathbf{A}_c$ is the common sparse coding matrix. This matrix can be modeled by parameters of a set of M fully-connected layers with shared parameters. Note that the reconstruction term $\|\mathbf{Z}^m_{\text{test}} - \mathbf{Z}^m_{\text{train}}\mathbf{A}_c\|_F^2$ in the m-th modality is equivalent to the penalty term of a fully-connected layer with the input $\mathbf{Z}^m_{\text{train}}$, the output $\mathbf{Z}^m_{\text{test}}$ and the parameters $\mathbf{A}_c$. We use this in the implementation of our model and refer to the fully-connected layer as *joint sparse coding layer*.

The joint sparse coding layer is located between encoder and decoder of different modalities. This layer performs an identical task across all the modalities. It passes the training features to the corresponding decoder, and uses the parameters $\mathbf{A}_c$ to reconstruct the testing features, and passes the reconstructions to the decoders.

Assuming that $\hat{\mathbf{Z}}^m_{\text{train}}$ and $\hat{\mathbf{Z}}^m_{\text{test}}$ are respectively outputs of the common sparse coding layer for the training set and testing set in the m-th modality, we have

$$\hat{\mathbf{Z}}^m_{\text{train}} = \mathbf{I}_{n_{\text{train}}}\mathbf{Z}^m_{\text{train}}, \quad \hat{\mathbf{Z}}^m_{\text{test}} = \mathbf{Z}^m_{\text{train}}\mathbf{A}_c, \quad (8)$$

where $\mathbf{I}_{n_{\text{train}}} \in \mathbb{R}^{n_{\text{train}} \times n_{\text{train}}}$ is the identity matrix. Therefore, if for the mth decoder the input is $\hat{\mathbf{Z}}^m = [\hat{\mathbf{Z}}^m_{\text{train}}, \hat{\mathbf{Z}}^m_{\text{test}}]$, from (8) we can calculate $\hat{\mathbf{Z}}^m$ as $\hat{\mathbf{Z}}^m = \mathbf{Z}^m\mathbf{\Theta}_{sc}$, where

$$\mathbf{\Theta}_{sc} = \begin{bmatrix} \mathbf{I}_{n_{\text{train}}} & \mathbf{A}_c \\ \mathbf{0}_{n_{\text{train}} \times n_{\text{test}}} & \mathbf{0}_{\text{test}} \end{bmatrix}, \quad (9)$$

where $\mathbf{0}_{n_{\text{train}} \times n_{\text{test}}} \in \mathbb{R}^{n_{\text{train}} \times n_{\text{test}}}$ and $\mathbf{0}_{\text{test}} \in \mathbb{R}^{m \times m}$ are zero matrices.

Combining the criteria in sparse coding and training of the encoder-decoders, one can write the reconstruction objective as

$$\mathcal{L}_{rec} = \sum_{m=1}^{M} \|\mathbf{Z}^m - \mathbf{Z}^m\mathbf{\Theta}_{sc}\|_F^2 + \lambda_0\|\mathbf{\Theta}_{sc}\|_1 + \sum_{m=1}^{M} \lambda_1\|\mathbf{X}^m - \hat{\mathbf{X}}^m\|_F^2 \quad (10)$$

Discriminative Criterion: We aim to train encoders by which the embeddings of different classes are best discriminated against. This property can be enforced on the encoders by incorporating labels for the training set. We plug discriminator heads to the output of encoders and train them to discriminate embedding features of different classes. Let $\mathbf{Y}$ represent the labels of the training samples; we define the discriminative criterion as follows

$$\mathcal{L}_{cls} = \sum_{m=1}^{M} \text{CE}\left(D_m(\mathbf{Z}^m_{\text{train}}), \mathbf{Y}\right) \quad (11)$$

where D_m denotes the discriminator head that is dedicated to classifying the embedding features of m-th modality and $\text{CE}(\cdot,\cdot)$ is the cross-entropy loss. We let the error of these discriminators be backpropagated to the encoders so that the encoders learn to produce separable embedding features.

$$\mathcal{L}_{cls} = \sum_{m=1}^{M} \|\mathbf{Z}^m - \mathbf{Z}^m\mathbf{\Theta}_{sc}\|_F^2 + \lambda_0\|\mathbf{\Theta}_{sc}\|_1 + \sum_{m=1}^{M} \lambda_1\|\mathbf{X}^m - \hat{\mathbf{X}}^m\|_F^2 \quad (12)$$

The discriminative criterion aims to train encoders that produce separable embedding features.

Full Objective: Combining the reconstruction and the discriminative criteria, our full objective function for training our networks is as follows:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \beta\mathcal{L}_{\text{cls}}, \quad (13)$$

where $\beta > 0$ is a regularization parameter. Note that with our formulation, it is possible to train the networks in an end-to-end manner, and yet find the optimal sparse codes and encoder-decoder parameters, simultaneously.

Classification Rule: Once the networks are trained and the common sparse coding matrix $\mathbf{A}_c$ is found, we can use them for associating class labels to the test samples. Each test sample comes with m modalities as $\{\mathbf{x}^m_{\text{test}}\}_{m=1}^M$. They have the corresponding embedding features as $\{\mathbf{z}^m_{\text{test}}\}_{m=1}^M$, and the corresponding sparse code column as $\boldsymbol{\alpha}$ in the common sparse code matrix $\mathbf{A}_c$. We can estimate the label of the sample by

$$\text{class}(\{\mathbf{x}^m_{\text{test}}\}_{m=1}^M) = \arg\min_k \sum_{m=1}^{M} \|\mathbf{z}^m_{\text{test}} - \mathbf{Z}^m_{\text{train}}\Gamma_k(\boldsymbol{\alpha})\|_F^2 \quad (14)$$

with $\Gamma_k(\cdot)$ similar to the equation (3).

In addition to the solution of (14), in our model, the discriminator heads in our framework can provide extra class label predictions for the test samples. Thus, we determine the final label estimates by ensembling the predictions of discriminator heads and the solution of (14). This is done by averaging the normalized scores.

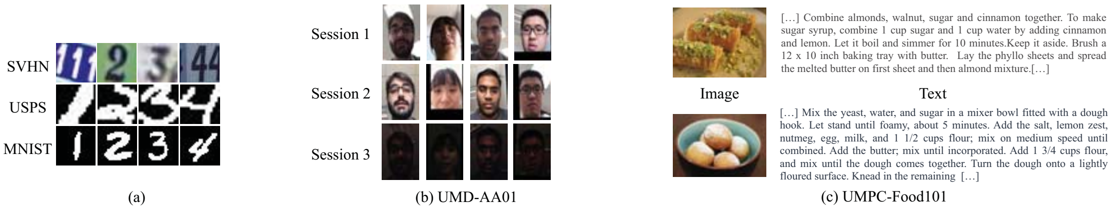
4. EXPERIMENTAL RESULTS

We evaluate our method on three multimodal datasets for digit classification, face recognition, and food categorization. We evaluate our method against state-of-the-art unimodal SRC methods, multimodal SRC methods, and the commonplace fusion methods. In particular, we compare against SRC [1] and DSRC [11] as unimodal baselines. For multimodal SRCs, we compare against Joint Sparse representation (J-SRC) [5] as well as the classical SRC performed on the concatenation of individual modalities denoted as SRC-C. Finally, in the last category of our baselines, we compare against the late feature fusion (feature-fusion), and score-fusion methods. Feature-fusion and score-fusion are of the most effective approaches in deep multimodal learning [12, 13].

We use the following datasets in our experiments:

Digits: We combine SVHN [23], USPS [24] and MNIST [25] digits datasets to assemble a multiview digit dataset. Here, we view images from the individual datasets as different views of the same digit. Since the number of parameters in the sparse coding layer of our model scales quadratically with the size of the data, we randomly select 200 samples per digit to keep the networks to a tractable size. In total, we have 2000 multiview digits.

Faces: We view different Sessions of the UMD mobile faces dataset (UMDAA-01) [26] as different modalities. We randomly select 50 facial images per subject from each Session.



[…] Combine almonds, walnut, sugar and cinnamon together. To make sugar syrup, combine 1 cup sugar and 1 cup water by adding cinnamon and lemon. Let it boil and simmer for 10 minutes.Keep it aside. Brush a 12 x 10 inch baking tray with butter. Lay the phyllo sheets and spread the melted butter on first sheet and then almond mixture.[…]

[…] Mix the yeast, water, and sugar in a mixer bowl fitted with a dough hook. Let stand until foamy, about 5 minutes. Add the salt, lemon zest, nutmeg, egg, milk, and 1 1/2 cups flour; mix on medium speed until combined. Add the butter; mix until incorporated. Add 1 3/4 cups flour, and mix until the dough comes together. Turn the dough onto a lightly floured surface. Knead in the remaining […]

(a) (b) UMD-AA01 (c) UMPC-Food101

Fig. 2. Samples from different modalities of datasets used in our experiments. (a) Digits from MNIST, SVHN and USPS. (b) Face images from different Sessions of UMDAA-01. (c) Food images and their recipe from UMPC-food101.

	SRC-C	J-SRC	score-fusion	feature-fusion	DMSRC (ours)	Unimodal SRC			Unimodal DSRC		
						M1	M2	M3	M1	M2	M3
Digits	91.87	92.34	18.25	18.13	96.25	11.98	88.13	92.25	62.75	94.25	95.37
Faces	83.45	84.76	14.37	15.17	94.13	78.32	77.21	75.83	91.21	90.56	89.12
Foods*	63.42	65.12	90.62	87.31	92.75	55.18	49.81	n/a	91.16	76.66	n/a

* All the methods for food-101 dataset use deep features extracted from DenseNet and BERT (for images and texts, respectively).

Table 1. Classification accuracy of different methods. M1 is SVHN in digits, Session1 in Faces and Images in Foods. M2 is USPS in digits, Session2 in Faces and texts in Foods. M3 is MNIST in digits, Session3 in Faces.

UMPC Food-101 [27]: The dataset contains images of 101 different foods along with recipes found from the web for these datasets. We keep the first 10 classes and randomly select 200 samples per class in our experiments. For text normalization, we remove double spaces, lower case all characters, and remove any character other than the English alphabets.

Figure 2 (a), (b), and (c) show samples from the digits, UMDAA-01 and UMPC Food-101 datasets, respectively. We use 60% of the samples in each dataset as the training set, and the remaining 20% as the testing set.

Training details: We implemented our method with Tensorflow-1.4. We use the adaptive momentum-based gradient descent method (ADAM) [28] to minimize our loss functions, and apply a learning rate of 10^{-3}. Before we start training on our objective function, in each experiment, we pre-train our encoder and decoder on the dataset without the sparse coding layer. We set the regularization parameters as $\lambda_0 = 1$, $\lambda_1 = 8$ and $\beta = 1000$ in all the experiments.

4.1. Digits and Faces

For Digits and Faces datasets, we adopt the same architecture as described in [11]. That is using stacked autoencoders of four convolutional layers for the encoder and three deconvolution layers for the decoder per each modality. The first two rows in Table 1 compare the performance of our method against unimodal and multimodal classifiers on digits and faces datasets. In the first row of Table 1, M1, M2, and M3 refer to SVHN, MNIST, and USPS datasets, respectively. In the second row, M1, M2 and M3 respectively refer to Session 1, Session 2 and Session 3 of UMDAA-01.

We observe multimodal SRC-based methods outperform the unimodal methods. This clarifies the benefits of integrating multiple modalities. However, score-fusion and feature fusion perform poorly here since the networks are shallow and are trained from scratch. Our DMSRC provides the best performance in both the datasets by using both the benefits of deep multimodal learning and the robustness of SRC-based methods.

4.2. Deep Networks with State-of-the-art Architectures

In this experiment, we evaluate our method against state-of-the-art deep neural networks. We adopt DenseNet [29] and BERT [30] networks that are of the most efficient deep architectures for processing images and texts, respectively. We use Wikipedia pre-trained BERT and pre-train DenseNet on Imagenet.

For both the networks, we add a fully-connected layer with 100 hidden nodes before the final classifier layer to provide a low-dimensional in-depth feature space in which the experiments of SRC-based methods are conducted. This layer is fine-tuned by the training samples of our UMPC-Food101 subset. Score-fusion and feature-fusion methods also use the same architecture.

For unimodal DSRC and our method, we use two fully-connected layers with 40 hidden state nodes as encoder and decoder.

In Table 1, we refer to the image modality as M1 and denote the text modality as M2. The results of Table 1 are interesting to compare with unimodal accuracies of 90.12% for DenseNet in the image modality, and 72.33% for BERT in the text modality. As Table 1 reveals, score-fusion and feature-fusion methods perform quite well, and on contrast, linear SRC-based method does not show a strong performance. It shows that although the in-depth features provide discriminative features, the learned features are not suitable for a linear sparse representation. DSRC and our MDSRC, on the other hand, can successfully map the features to a latent space in which the benefits of SRC-based methods can be exploited. Our MDSRC, outperforms all the baselines.

5. CONCLUSION

We presented a deep sparse representation-based fusion method for classifying multimodal signals. We use autoencoders to develop features that may be different across different modalities but share the same sparse codes in sparse representation classification problems that are applied to separate modalities. The training objective for encoders consists of reconstruction criteria and discriminative criterion. We proposed a classification rule that uses sparse codes as well as the prediction of classification heads in different modalities to determine an accurate estimate for classifying the test samples.

Acknowledgment: This research is supported in part by the Northrop Grumman Mission Systems Research in Applications for Learning Machines (REALM) initiative and the NSF grant 1618677.

6. REFERENCES

[1] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Yi Ma, “Robust face recognition via sparse representation,” *IEEE Trans-*

actions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp. 210 –227, feb. 2009.

[2] Michal Aharon, Michael Elad, Alfred Bruckstein, et al., "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311, 2006.

[3] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[4] M. Joneidi, A. Zaeemzadeh, S. Rezaeifar, M. Abavisani, and N. Rahnavard, "Lfm signal detection and estimation based on sparse representation," in *2015 49th Annual Conference on Information Sciences and Systems (CISS)*, 2015, pp. 1–5.

[5] Sumit Shekhar, Vishal M Patel, Nasser M Nasrabadi, and Rama Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 113–126, 2014.

[6] Lei Zhang, Meng Yang, and Xiangchu Feng, "Sparse representation or collaborative representation: Which helps face recognition?," in *Computer vision (ICCV), 2011 IEEE international conference on*. IEEE, 2011, pp. 471–478.

[7] M. Abavisani, M. Joneidi, S. Rezaeifar, and S. B. Shokouhi, "A robust sparse representation based face recognition system for smartphones," in *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2015, pp. 1–6.

[8] Li Zhang, Wei Da Zhou, Pei Chann Chang, Jing Liu, Zhe Yan, Ting Wang, and Fan Zhang Li, "Kernel sparse representation-based classifier," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1684 –1695, april 2012.

[9] S. Gao, I. W. Tsang, and L. T. Chia, "Kernel sparse representation for image classification and face recognition," in *European Conference on Computer Vision*, 2010, vol. 6314.

[10] X. T. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[11] Mahdi Abavisani and Vishal M Patel, "Deep sparse representation-based classification," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 948–952, 2019.

[12] Dhanesh Ramachandram and Graham W Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[13] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *International Conference on Machine Learning*, 2011, pp. 689–696.

[14] Mahdi Abavisani and Vishal M Patel, "Deep multimodal subspace clustering networks," *Journal of special topics in signal processing*, 2018.

[15] Emmanuel J Candes and Terence Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[16] David L Donoho, "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, vol. 59, no. 6, pp. 797–829, 2006.

[17] S. Shafiee, F. Kamangar, V. Athitsos, J. Huang, and L. Ghandehari, "Multimodal sparse representation classification with fisher discriminative sample reduction," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 5192–5196.

[18] Gaurav Goswami, Paritosh Mittal, Angshul Majumdar, Mayank Vatsa, and Richa Singh, "Group sparse representation based classification for multi-feature multimodal biometrics," *Information Fusion*, vol. 32, pp. 3–12, 2016.

[19] Haichao Zhang, Nasser M Nasrabadi, Yanning Zhang, and Thomas S Huang, "Multi-observation visual recognition via joint dynamic sparse representation," in *International Conference on Computer Vision*. IEEE, 2011, pp. 595–602.

[20] Xiao-Tong Yuan, Xiaobai Liu, and Shuicheng Yan, "Visual classification with multitask joint sparse representation," *IEEE Transactions on Image Processing*, vol. 21, no. 10, pp. 4349–4360, 2012.

[21] Bin Cheng, Guangcan Liu, Jingdong Wang, Zhongyang Huang, and Shuicheng Yan, "Multi-task low-rank affinity pursuit for image segmentation," in *International Conference on Computer Vision*. IEEE, 2011, pp. 2439–2446.

[22] Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik, "Learning by transduction," in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 148–155.

[23] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, 2011, vol. 2011, p. 5.

[24] Jonathan J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 16, no. 5, pp. 550–554, 1994.

[25] Yann LeCun and Corinna Cortes, "Mnist handwritten digit database," *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2010.

[26] Heng Zhang, Vishal M Patel, Sumit Shekhar, and Rama Chellappa, "Domain adaptive sparse representation-based classification," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. IEEE, 2015, vol. 1, pp. 1–8.

[27] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso, "Recipe recognition with large multimodal food dataset," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6.

[28] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.