Cross-Modal Prediction of Superclasses Using Cortex-Inspired Neural Architecture

Olcay Kursun
Department of Computer Science
Auburn University at Montgomery
Montgomery, AL 36117, USA
okursun@aum.edu

Hoa T. Nguyen
Department of Computer Science
University of Central Arkansas
Conway, AR 72035, USA
hnguyen4@cub.uca.edu

Oleg V. Favorov
Joint Department of Biomedical Engineering
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA
favorov@email.unc.edu

***Abstract*—The concept of stimulus feature tuning is fundamental to neuroscience. Cortical neurons acquire their feature-tuning properties by learning from experience and using proxy signs of tentative features' potential usefulness that come from the spatial and/or temporal context in which these features occur. According to this idea, local but ultimately behaviorally useful features should be the ones that are predictably related to other such features either preceding them in time or taking place side-by-side with them. Inspired by this idea, in this paper, deep neural networks are combined with Canonical Correlation Analysis (CCA) for feature extraction and the power of the features is demonstrated using unsupervised cross-modal prediction tasks. CCA is a multi-view feature extraction method that finds correlated features across multiple datasets (usually referred to as views or modalities). CCA finds linear transformations of each view such that the extracted principal components, or features, have a maximal mutual correlation. CCA is a linear method, and the features are computed by a weighted sum of each view's variables. Once the weights are learned, CCA can be applied to new examples and used for cross-modal prediction by inferring the target-view features of an example from its given variables in a source (query) view. To test the proposed method, it was applied to the unstructured CIFAR-100 dataset of 60,000 images categorized into 100 classes, which are further grouped into 20 superclasses and used to demonstrate the mining of image-tag correlations. CCA was performed on the outputs of three pre-trained CNNs: AlexNet, ResNet, and VGG. Taking advantage of the mutually correlated features extracted with CCA, a search for nearest neighbors was performed in the canonical subspace common to both the query and the target views to retrieve the most matching examples in the target view, which successfully predicted the superclass membership of the tested views without any supervised training.**

Keywords*—Multi-view Feature Extraction, Dimensionality Reduction, Recommendation Systems, Deep Learning, Transfer Learning, Contextual Guidance.

I. INTRODUCTION

Cerebral cortex is a complex dynamical system dominated by feedback circuits and cross-modal interactions (e.g., among sensory and behavioral components such as visual, auditory, tactile, and motor system modalities). These modalities are organized in a modular and hierarchical architecture [1]. The central pathway in the feed-forward elaboration of cortical neurons' properties in these multi-layered architectures typically proceeds through a repeating sequence of two cortical layers: the input Layer 4 and the output Layer 3. The operation of these layers resembles a convolutional block (convolution, rectification, and pooling) of deep convolutional neural networks [2]. Each column-shaped module in a higher-level cortical area builds its more complex features using as input the features of a local set of modules in the lower-level cortical area. Thus, as we go into higher areas these features become increasingly more global and nonlinear, and thus more descriptive [1-5]. It has been proposed that interactions among the modules provide contextual guidance for feature tuning and cross-modal prediction (prediction of missing information using information available in other modules or modalities). Such interactions are akin to maximization of mutual information via Canonical Correlation Analysis (CCA) [2, 5]. In this paper, we propose a neural architecture inspired by the mutual information maximization model of this cross-modal cortical architecture.

In artificial neural networks, cross-modal learning – akin to learning via contextual guidance in cerebral cortical networks described above – refers to learning to infer information in one view/modality from the information in another view/modality relying on mutual information among views/modalities [6-16]. Cross-modal learning is used in feature extraction and prediction in many real-world problems, where data are frequently multimodal. Cross-modally chosen features can enhance predictive models by helping to eliminate view-specific noisy distractors and emphasize dominant underlying causal factors. That is, extraction of mutual information between modalities can guide feature extraction towards modality-invariant but class-specific features and creation of a common space of such features for all modalities/views. Such a common space of features for all modalities/views would allow one view's samples to be predicted by projecting other views' samples onto this space, which would be emphasizing the underlying sources of correlations and class-related features [8, 11].

One approach to reveal correlations between different views/modalities is Canonical Correlation Analysis (CCA), which maximizes correlation between modalities [9, 12, 13], enabling their representation using a common feature subspace. Although CCA is originally proposed to tune to features that maximize linear correlations between two views [12], other variations have been proposed for its application to more than two views [14] and for discovering discriminative and nonlinear relations among the views, such as kernel-CCA [8], discriminative CCA [9], and deep-CCA [15].

Cross-modal prediction is a powerful application to demonstrate the usefulness of the CCA extracted features. For example, in [11], the pseudo inverse of the CCA features is used to achieve cross-modal prediction. In this paper, we propose a novel cross-modal recommendation framework that performs a search for nearest neighbors to recommend in the canonical subspace learned by CCA as a common subspace of the query and target views. We tested our method on a noisy-labelled image dataset, in which the images formed one modality and the noisy labels yielded the superclass information that formed the other modality. To incorporate noisy labels into its feature tuning and image recommendations, CCA is applied to discover the common subspace between these two modalities. In other words, we

This work is supported, in part, by NSF under Grant No. 2003740.

979-8-3503-3849-2/23/$31.00 ©2023 IEEE

use image features extracted by deep networks and a set of noisy tags describing the class/superclass of the images to train our CCA-based algorithm [16].

II. Multiview Dataset Construction

We used CIFAR-100 dataset [17] contains 60,000 images categorized into 100 classes and these classes are further grouped into 20 superclasses. We used the image as one view and a noisy representation of the superclass information as the second view. Thus, in this setup, we have each image with a tag that correlates to its superclass. If the superclass information was provided accurately in a 20-dimensional vector simply as a one-hot-encoding, then CCA would be equivalent to LDA (linear discriminant analysis) [18]. Instead we create a 100-dimensional noisy version of the superclass. Thus, we construct a binary 60,000x100 tag-view as the second view, where a 1-value in the dataset means that the image corresponding to that row may belong to the class corresponding to that column. The 100-dimensional encoding of the super-class information is probabilisticly defined as follows: the component corresponding to the true class has very high probability of being 1, components corresponding to classes in the same superclass has high probability of being 1, and the rest of the components have very low probabilities of being 1. For example, if the class of an image was "beaver", then the 100-dimensional tag would likely contain a 1 for the beaver class but also contain 1 for other classes under the "aquatic mammals" superclass (such as otter, seal, and dolphin). When using the proposed noisy encoding scheme that favors but not exactly identifies the class and the superclass, the demand on CCA would be to find the most matching linear combinations of image features with linear combinations of these noisy tags. As multiple 1's are associated with an image (with more probability of being 1 for the classes that belong to the superclass of the image), CCA captures superclass information, and the proposed CCA-based cross-modal prediction learns to suggest which superclass the given query image belongs to. Note that this cross-modal prediction is achieved without explicitly training a supervised classifier for predicting the superclasses.

The tag-related view is referred to as CIFAR-100-tag. Deep learning models were used as feature extractors for images in CIFAR-100. The CIFAR-100-tag data and the deep-learning features were the two modalities fed to the proposed cross-modal prediction model. The image feature extractors used were pre-trained CNN models. Their abilities in discovering distinguishable patterns in images were compared. The application of pre-trained CNNs is part of transfer learning, a subfield of machine learning and artificial intelligence that exercises the knowledge gained from a source task to a different but similar target task, which is one of the benefits of deep learning systems (Fig. 1) [19, 20]. Pre-trained CNNs are models that are already built on very large datasets for image classification. Pre-trained CNN models provide a shortcut to training a CNN from scratch, which may take up time and resources depending on the size of the dataset. Three pre-trained CNNs were investigated: AlexNet, ResNet, and VGG [21]. All three models were trained on the ImageNet database, which has more than 14 million images grouped into about 22 thousand classes (according to statistics recorded on ImageNet's homepage). AlexNet architecture includes eight layers, five of which are convolutional layers, and three are fully connected layers. The input to AlexNet must be RGB images of size 256×256 [22]. ResNet (short for Residual Network) was built and trained on one million 224x224 colored images from ImageNet [23]. There have been multiple versions of ResNet depending on the number of layers; we used ResNet101 that has 101 layers. VGG requires RGB images of dimensions 224x224 [24]. Similar to ResNet, there are multiple versions of VGG; we used VGG11_bn.

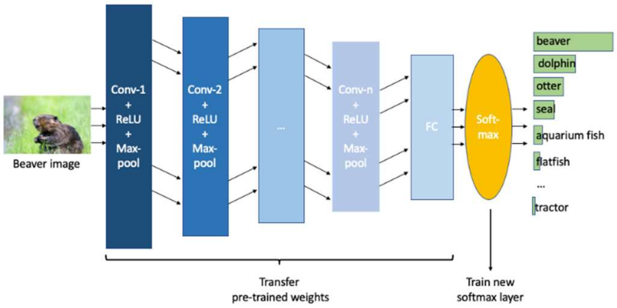

Fig. 1. An exemplary deep CNN application via transfer learning.

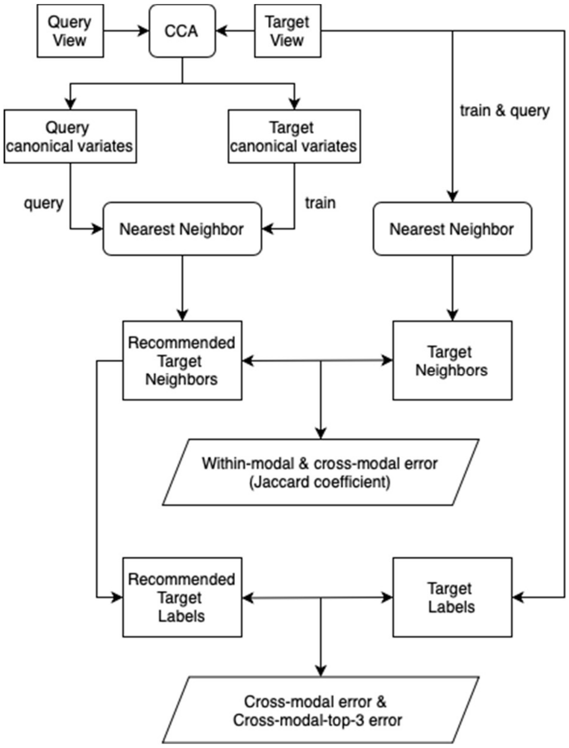
Query View
CCA
Target View
train & query
Query canonical variates
Target canonical variates
query
train
Nearest Neighbor
Nearest Neighbor
Recommended Target Neighbors
Target Neighbors
Within-modal & cross-modal error (Jaccard coefficient)
Recommended Target Labels
Target Labels
Cross-modal error & Cross-modal-top-3 error

Fig. 2. Proposed method for cross-modal prediction.

III. Proposed Method

The cross-modal prediction approach proposed in [11] worked by inverting the canonical weight matrix of the target view. This essentially means transforming the query view to the canonical subspace and then, using the inverted weights, converting the canonical subspace to obtain the most fitting (but artificial) representations in the target view. We propose that for cross-modal prediction, performing the search within the canonical subspace is better than generating artificial representations. Even if the sample reconstruction generates a realistic target-view example, it is not a real example in the target-view dataset. To find the real examples, a nearest neighbors search would be performed for the retrieval task of relevant items from the target view. Therefore, instead of inverting the CCA features of the query to obtain a reasonable representation in the target-view space, the proposed approach applies a nearest neighbors search within the canonical subspace. As the CCA features are generally much fewer than the original dimensionality, the proposed approach offers higher performance in the accuracy rate (due to the phenomenon known as curse-of-dimensionality) and also in the search time.

Our method first projects the data on CCA dimensions (the numbers of dimensions is selected based on the elbow point of the canonical correlations on the validation set) and then applies an nearest neighbors model to acquire similar examples using the covariate space; that is, the feature vector obtained from the query (first) view is used to find nearest neighbors of the CCA transformations of the other (second) view. These best matching examples have representations in both views and the search can be used to return the first or the second views of these best matches. A flowchart of this cross-modal prediction method is given in Fig. 2.

IV. Experimental Results

As a quick demonstration of the proposed method, we first use the well-known optdigits dataset that contains 1797 handwritten digits down-sampled to 8x8=64 pixels split into two sets (similar to [15]) as shown in Fig. 3. We also split the dataset randomly in a class-stratified way (thus preserving the class priors) into 10% for the training set and 90% for the test set. For each test example, we find *k* nearest neighbors in each view (note that each test example has two representations in these two views). We measure the percentage of the common nearest neighbors (for example, if all *k* were common to both sets, that would indicate 100% agreement; and if the two sets were disjoint, that would indicate 0% agreement). We compared the agreement when using the original pixels (32 dimensions per view), CCA with 5 components, and principal component analysis (PCA) with 5 components. As shown in Fig. 4, using CCA's covariate features lead to higher agreement in average.

To demonstrate the power of CCA features for cross-modal tasks on the CIFAR dataset, we used three pre-trained CNN models, AlexNet, ResNet, and VGG, as image feature extractors. This also allowed us to compare them with respect to their abilities in discovering distinguishable patterns in images, addressed by the corresponding CCA-based cross-modal prediction performance. Both AlexNet and VGG extract 4,096 image features at its last layer and feeds them into the final classifier (softmax) layer, while ResNet extracts 512 features. These features are highly descriptive and are suitable to be transferred to this CIFAR image-domain classification task. The features were transferred and used in the CCA-based cross-modal prediction. To help CCA's convergence and to reduce its training runtime, principal component analysis (PCA) was applied to the features extracted by all three models to reduce the dimensionality and eliminate collinearities within these features: The number of PCA features was fixed at 500 for comparisons among models. 500 PCA features covered more than 80% of the total variance for all deep learning models.

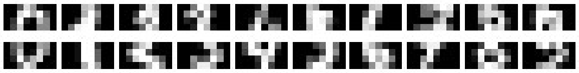

Fig. 3. Proposed method for cross-modal prediction.

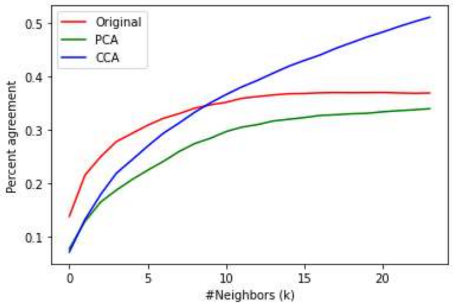

Fig. 4. Agreement between two sets of nearest neighbors

The number of CCA components used in the cross-modal prediction process with the CIFAR-100 datasets was chosen empirically to be 20. The highest correlation learned between the modalities' 20 components for the testing data was 0.5279, 0.5355, and 0.5350, respectively for AlexNet, ResNet, and VGG. The cross-modal prediction aimed at recommending 100 most fitting examples with the deep-learning image features and tags representations taking turns to be the query/target modality. Table I compares the test errors produced by the prediction process using canonical variates and those output by the pseudoinverse method.

TABLE I. Cross-modal Prediction Test Errors Produced by the Proposed and the Pseudoinverse Methods

Query-View	Target-View	CNN Architecture	Cross-modal classification error		Cross-modal-top-3 classification error	
			Proposed Method	Pseudoinverse Method	Proposed Method	Pseudoinverse Method
Images	Tags	AlexNet	0.3465	0.3553	0.1664	0.1793
		ResNet	0.2925	0.3015	0.1336	0.1469
		VGG	0.2989	0.3039	0.1271	0.1359
Tags	Images	AlexNet	0.2438	0.5045	0.1106	0.2914
		ResNet	0.2437	0.4459	0.1027	0.244
		VGG	0.2495	0.4861	0.1040	0.2688

With the query and target views as images and tags, respectively, the cross-modal prediction task was analogous to finding the tag representations that might label a given image. The proposed method was more accurate in the superclasses of the recommended tag representations. With AlexNet image features as the query, the proposed method output a cross-modal classification error of 0.3465, which means that about 65 out of 100 tag representations recommended were correct. This error was comparable to the pseudoinverse method. However, the proposed method was more beneficial in the way that it performed its search in a low dimensional space: It eliminated the additional computation to reconstruct the 500-dimensional image features. Moreover, the search in a low dimensional space was slightly more accurate than the pseudoinverse method. When considering the top-3 classification error, the proposed method had much smaller error. Furthermore, for both the proposed and pseudoinverse methods, cross-modal prediction using image features extracted by ResNet and VGG produced better results than those by AlexNet. With ResNet and VGG, the top-3 tag representations contained the true superclass approximately 87% of the time (the top-3 error of VGG was slightly lower). Overall, when comparing the pre-trained CNNs based on the corresponding CCA-based cross-modal prediction performance, ResNet and VGG were comparable in their ability to extract discriminative image features, and they were better than AlexNet.

When the tag representations were used as the query view, the cross-modal recommendation aimed at retrieving the most fitting images associated with the given noisy tag vector. For this cross-modal prediction task, the improvement in accuracy using the proposed method was more apparent. The proposed method performed significantly better than the pseudoinverse method, cutting the classification errors down by almost half. All three pre-trained CNNs worked well, with ResNet yielding the best performance.

V. Conclusion

In this paper, we propose a novel CCA-based cross-modal prediction method, which is built on using CCA to extract canonical features and find a common (canonical) subspace between the two views of the training examples. During the test phase, one of the views is chosen to be the query view and the other one to be the target view. The cross-modal prediction is performed by computing the canonical variates using the data in the query view, and then applying a nearest-neighbors search in the canonical space of the target view to retrieve the most matching examples in the target view. The proposed method was compared with an alternative method referred to as the pseudoinverse method [11], which reconstructs the representations in the target view from the canonical space and then performs a nearest-neighbors search in a much higher dimensional space. The two alternative methods were applied to CIFAR-100 dataset of 60,000 images categorized into 100 classes and further into 20 superclasses. The experimental results showed that the proposed nearest-neighbor search in the canonical space is more effective than the pseudoinverse method.

More generally, this study offers an experimental support to the neuroscientific conjecture that neurons on the cerebral cortex acquire their feature tuning properties under the contextual guidance from neighboring processing modules in the same cortical area or from other cortical areas.

References

[1] Hawkins, J., Ahmad, S., & Cui, Y. (2017). A theory of how columns in the neocortex enable learning the structure of the world. Frontiers in Neural Circuits, 11:81.

[2] Kursun, O., Dinc, S., Favorov, O.V. (2022) Contextually Guided Convolutional Neural Networks for Learning Most Transferable Representations, 24th IEEE International Symposium on Multimedia (IEEE-ISM), Naples, Italy, December 2022.

[3] Clark, A. & Thornton, C. (1997). Trading spaces: Computation, representation, and the limits of uninformed learning. Behavioral and Brain Sciences, 20(1):57-66.

[4] Phillips, W.A., Singer, W. (1997) In search of common foundations for cortical computation. Behavioral and Brain Sciences 20: 657-722.

[5] Becker, S. (1996) Mutual information maximization: models of cortical self-organization. Network 7(1): 7-31.

[6] Chen, Z., Lu, F., Yuan, X., & Zhong, F. (2017). TCMHG: Topic-based cross-modal hypergraph learning for online service recommendations. IEEE Access, 6, 24856-24865.

[7] Chen, N., Zhu, J., & P Xing, E. (2010). Predictive subspace learning for multi-view data: a large margin approach. In Proceedings of the Neural Information Processing Systems (pp. 361–369).

[8] Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. Neural computation, 16(12), 2639-2664.

[9] Sakar, C. O., & Kursun, O. (2017). Discriminative feature extraction by a neural implementation of canonical correlation analysis. IEEE Transactions on neural networks and learning systems, 28(1), 164-176.

[10] Zhou, Y., Mishra, S., Verma, M., Bhamidipati, N., & Wang, W. (2020, April). Recommending themes for ad creative design via visual-linguistic representations. In Proc. of The Web Conference 2020 (pp. 2521-2527).

[11] Bilenko, N. Y., & Gallant, J. L. (2016). Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. Frontiers in neuroinformatics, 10, 49.

[12] Hotelling, H. (1992). Relations between two sets of variates. In Breakthroughs in statistics (pp. 162-190). Springer, New York, NY.

[13] Gong, Y., Ke, Q., Isard, M., & Lazebnik, S. (2014). A multi-view embedding space for modeling internet images, tags, and their semantics. International journal of computer vision, 106(2), 210-233.

[14] Kettenring, J.R. (1971) Canonical analysis of several sets of variables. Biometrika 58:433-451.

[15] Benton, A., Khayrallah, H., Gujral, B., Reisinger, D. A., Zhang, S., & Arora, R. (2017). Deep generalized canonical correlation analysis. arXiv preprint arXiv:1702.02519.

[16] Nguyen, H.T. (2021) Cross-Modal Prediction Using Canonical Correlation Analysis with Privacy Preservation, Master's Thesis, Supervisor: Kursun, O., University of Central Arkansas.

[17] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

[18] Bartlett, M. S. (1938). Further aspects of the theory of multiple regression. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 34, No. 1, pp. 33-40). Cambridge University Press.

[19] Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.

[20] Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. arXiv preprint arXiv:1411.1792.

[21] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703.

[22] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[23] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[24] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.