2015 IEEE International Conference on Data Mining

Deep Convolutional Neural Networks for Multi-Instance Multi-Task Learning

Tao Zeng
School of Electrical Engineering and Computer Science
Washington State University
Pullman, USA
tzeng@eecs.wsu.edu

Shuiwang Ji
School of Electrical Engineering and Computer Science
Washington State University
Pullman, USA
sji@eecs.wsu.edu

***Abstract*—Multi-instance learning studies problems in which labels are assigned to bags that contain multiple instances. In these settings, the relations between instances and labels are usually ambiguous. In contrast, multi-task learning focuses on the output space in which an input sample is associated with multiple labels. In real world, a sample may be associated with multiple labels that are derived from observing multiple aspects of the problem. Thus many real world applications are naturally formulated as multi-instance multi-task (MIMT) problems. A common approach to MIMT is to solve it task-by-task independently under the multi-instance learning framework. On the other hand, convolutional neural networks (CNN) have demonstrated promising performance in single-instance single-label image classification tasks. However, how CNN deals with multi-instance multi-label tasks still remains an open problem. This is mainly due to the complex multiple-to-multiple relations between the input and output space. In this work, we propose a deep leaning model, known as multi-instance multi-task convolutional neural networks (MIMT-CNN), where a number of images representing a multi-task problem is taken as the inputs. Then a shared sub-CNN is connected with each input image to form instance representations. Those sub-CNN outputs are subsequently aggregated as inputs to additional convolutional layers and full connection layers to produce the ultimate multi-label predictions. This CNN model, through transfer learning from other domains, enables transfer of prior knowledge at image level learned from large single-label single-task data sets. The bag level representations in this model are hierarchically abstracted by multiple layers from instance level representations. Experimental results on mouse brain gene expression pattern annotation data show that the proposed MIMT-CNN model achieves superior performance.**

Keywords*—*Deep learning, multi-instance learning, multi-task learning, transfer learning, bioinformatics

I. Introduction

In conventional supervised learning, we learn a classifier based on a training set of feature vectors, where each feature vector is assigned a single label from a predefined set of classes. Such problem is referred to as single-instance learning when considering its input space, or single-task (single-label) learning if the focus is on the output space. When the inputs are images, a common approach is to design hand-crafted local features such as the SIFT [1], along with unsupervised learning schemes such as the bag-of-words to build image level features. These features are then used in classifiers such as support vector machine or random forests [2]. Such schemes, known as shallow learning, usually require substantial prior domain knowledge.

In contrast to the shallow models, recent advances in deep models are revolutionizing various image-related tasks. Among these models, deep convolutional neural network (CN-N) archived state-of-the-art performance in large-scale single-label object recognition tasks [3]. Despite of their satisfactory performance on these tasks, the application of CNN to more complex learning scenarios is largely unexplored. For example, in many real world problems a label is usually given by observing multiple instances of samples. For example, in the case of drug discovery, a drug is considered effective if any of its molecules has the binding ability to the target proteins. Also in image classification, a scene could be analyzed by combining multiple sub-regions of the image. Such learning settings are referred to as multi-instance learning (MIL). In MIL, we learn a classifier based on a training set of bags, where each bag contains multiple feature vectors known as instances. In addition, labels are only associated with bags while instance labels remain unknown and might be inferred during learning.

In contrast to MIL that focuses on the input space, multi-label or multi-task learning focus on output space in which multiple labels are assigned to a given sample. For example, an image may contain more than one object of different categories. Multi-task learning can be solved by converting it to a set of independent single-task learning problems [4]. Other sophisticated solutions have attempted to model correlations of the labels simultaneously [5], [6], which have shown to improve the performance of tasks. Currently, MIL methods mainly focus on single-label scenarios and it requires multi-task problems to be solved task-by-task independently. Hence, the correlations between labels are usually not taken into account.

In this work, we propose a novel deep CNN architecture to solve multi-instance and multi-task learning problems simultaneously. In addition to taking advantage of the state-of-the-art performance achieved by CNN, it overcomes the aforementioned limitations of MIL and MTL when they are performed separately in scenarios that require multiple instances as input and multiple tasks as output. We apply our deep model to brain image data set, where multiple images are usually produced by sectioning across 3D tissue in order to detect certain properties inside the tissue. These properties are annotated for each individual anatomical subregions in the

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.92

given tissue. We attempt to automate the task of annotating gene expression patterns at various structures in the mouse brain from *in situ* hybridization (ISH) image sets provided by the Allen Developing Mouse Brain Atlas (ADA).

There are several significant challenges to learn a model from such data sets. First, information of brain structures are presented across several images and the correspondence between ISH images and brain structures is unknown in current data sets. This is, there exist a complex internal correlation between instances (ISH images). Secondly, the gene expression annotation for brain structures are collectively provided to the entire set of ISH section images across the brain. Thirdly, even for the same brain structures, its corresponding location, size, shape, inner structures in ISH images may vary in different experiments. All these challenges make the task very difficult to tackle as compared to traditional machine learning tasks.

Hence, to solve these problems of annotating gene expression patterns for brain structures, our deep convolutional neural networks are constructed as follows. First, multiple CNN models that share the same set of parameters were employed to build lower layers of the networks. This allows for the transfer of knowledge learned from large natural image set and thus is able to form instance level representations for each of the ISH images obtained across brain sections. We then combine all feature maps from the last max pooling layers of each shared-weight CNN networks, followed by appending several convolutional layers and fully connected layers at the end of the networks. The newly added layers are designed to capture global representations of the bags at brain-wide. To enable model training with a small number of labeled examples, we transfer parameters from pre-trained CNN networks to those layers that are shared. The parameters of the newly added layers are initialized with standard randomization approach. We then fine-tune this model on ISH image data sets. Finally, features extracted from various layers of our deep models were used to train classifiers in order to compare the discriminative power of features at different deep CNN layers.

Experimental results show that our approach of constructing deep CNN networks outperforms the bag-of-word and pre-trained CNN model for annotating gene expression patterns at all developing stages. In addition, bag-level expression features yield higher performance than the simple combination of instance level features. These results indicate that our model can capture correlations between brain structures and ISH images.

II. Background and Related Work

A. Multi-Instance Learning

In contrast to the standard supervised learning setting in which each feature vector has an associated class label, multi-instance learning (MIL) considers a set of bags, each containing multiple feature vectors referred to as instances. In MIL, the available label information is only assigned to entire bags. Thus the labels of the individual instances in the bag are not known. Moreover, not all the instances are necessarily relevant; this is, some instances in the bag might not be relevant to certain labels. In standard settings of MIL, it is commonly assumed that a bag is labeled positive if at least one instance in that bag is positive. A bag is labeled negative if all the instances in it are negative. In general, the task of MIL is trying to correctly predict the labels of unseen bags through learning from training bags with known labels. Therefore, the key challenge in MIL is to deal with the ambiguity of not knowing which of the instances in a positive bag are the actual positive examples and which ones are not.

Research on handwriting recognition in [7] was an early attempt to deal with MIL problem. As a variation of supervised learning, the standard framework of MIL was initially proposed by [8] to predict the binding ability of a drug to the protein structure. In their experiment, each drug molecule may exhibit a number of conformations, having potential ability to bind to certain protein. However, such ability for individual conformation cannot be differentiated by chemical experiment. Hence, the binding ability (label) is only assigned to a drug (a bag), which consists of multiple conformations of its molecules (instances). The drug is positive if any of its constituent conformations has the binding ability. Following these initial studies, many MIL methods have been proposed, such as the diverse density [9], extended citation kNN [10], etc. MIL has achieved considerable success in solving a wide range of learning tasks ranging from image semantic learning and text categorization to financial market prediction. The standard MIL has also been widely used in bioinformatics studies [11].

Although many real-world problems can be appropriately modeled under the standard MIL, the standard MIL implicitly assumes that each instance has a hidden class label that identifies it as either a positive or a negative instance. This makes the qualification of a bag being positive entirely relies on the existence of one positive key instance. Such assumption, however, might not hold in other problem domains. For example, sometimes there are MIL problems in which a positive label is determined by the joint effects of several instances. For example, given multiple regions in a image, the problem of detecting beach can fall into a task of finding the existence of both sand and water in the subregions of the image. A positive label could be given under multiple scene combinations of subregions, such as that there are multiple regions that all have just sands while another set of non-overlapping regions containing only water scene, or both sand and water scenes exist in some subregions of image. To tackle such MIL problems that differ from the definition of standard MIL, researchers have proposed different generalized multi-instance learning frameworks to accommodate more sophisticated interactions among instances in the bag [12], [13].

B. Deep Convolutional Neural Networks

Deep learning models are based on the idea that representations of observed data are the results of hierarchical abstraction at many different levels [3], [15]. As level moves further up, more abstract information is generated by building on lower level features. Hence, such model can learn a hierarchy of features by building high-level features from low-level ones.

Convolutional neural networks (CNN) are a class of deep models that were inspired by information processing in the brain. In the visual neocortex of the brain, each neuron has a receptive field capturing information from certain local neighborhood in visual space. Inspired by such mechanism, CNN mimics the receptive field of biological neuron, and each

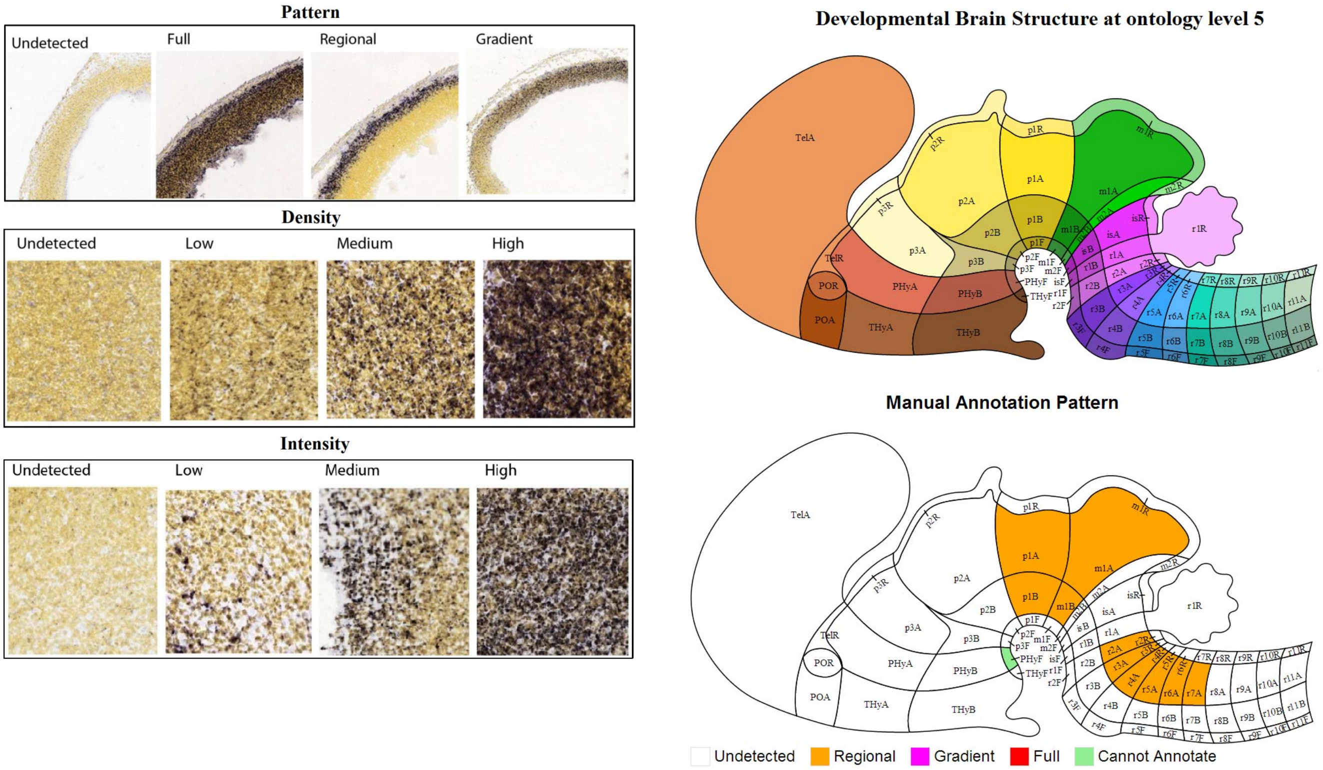

Fig. 1. Illustration of the manual annotations. The figure on the left illustrates the three different metrics. The top-right figure shows the brain structure ontology at level 5. The bottom-right figure denotes the corresponding manual annotations. The figures were reproduced from [14] with permission.

unit in CNN receives local inputs from lower level. CNN also uses replicated weight matrix for all units in the same feature map to compute the same feature from all locations on the inputs [3], [15].

CNN models usually consist of alternating combination of convolutional layers with trainable filters and local neighborhood pooling layers, resulting in a complex hierarchical representations of the inputs. CNNs are intrinsically capable of capturing highly nonlinear mappings between inputs and outputs. When trained with millions of labeled images, they have achieved superior performance on many image-related tasks [16]–[19]

III. The Proposed Models

A. *Problem Setup*

In the following, we will use the Allen Developing Mouse Brain Atlas (ADA) as an application example, though the proposed methods can be applied to generic multi-instance multi-task learning problems. The ADA provides a framework for exploring the spatiotemporal dynamics of gene expression over the course of mouse brain development [20]. In ADA, *in situ* hybridization (ISH) image data are available for about 2,000 genes in the sagittal section across seven developing stages. For each gene, ISH was applied to multiple sections of the brain to detect a specific gene expression covering the entire brain.

To characterize structural level gene expression, ADA created reference atlases for each of developing stages which allows the ISH images to be aligned to a standardized 3D reference brain platform, enabling segmentation of the brain into structures. The subsequent annotation of gene expression at structural levels was then determined manually using the expertise of neuroscientists [14]. Specifically, the experts first attempted to identify whether a given gene expression was detectable in a specified structure. They subsequently annotated this gene expression using pattern, intensity and density metrics if it is detectable. Otherwise, the gene was labeled as “undetected” for all three metrics. The pattern metric was scored as full, reginal, and gradient, whereas density and intensity metrics were scored as low, median and high, respectively (Figure 1). Currently, manual annotations have been generated for four (E11.5, E13.5, E15.5, and E18.5) out of the seven developing stages by Allen Institute for Brain Science. Such annotations enable neuroscientists to explore the intrinsic mechanism as to how genes regulate the development of brain at fine structure levels. However, manually annotating gene expressions over an enormous number of ISH images is labor-intensive and may result in inconsistence among different experts [14].

In ADA, an ISH experiment generated 12-15 sagittal image sections across the brain. A structure in the brain could be sectioned into multiple section images and some image sections may not contain a given structure. Information regarding the relationship as to how each structure maps to which ISH image

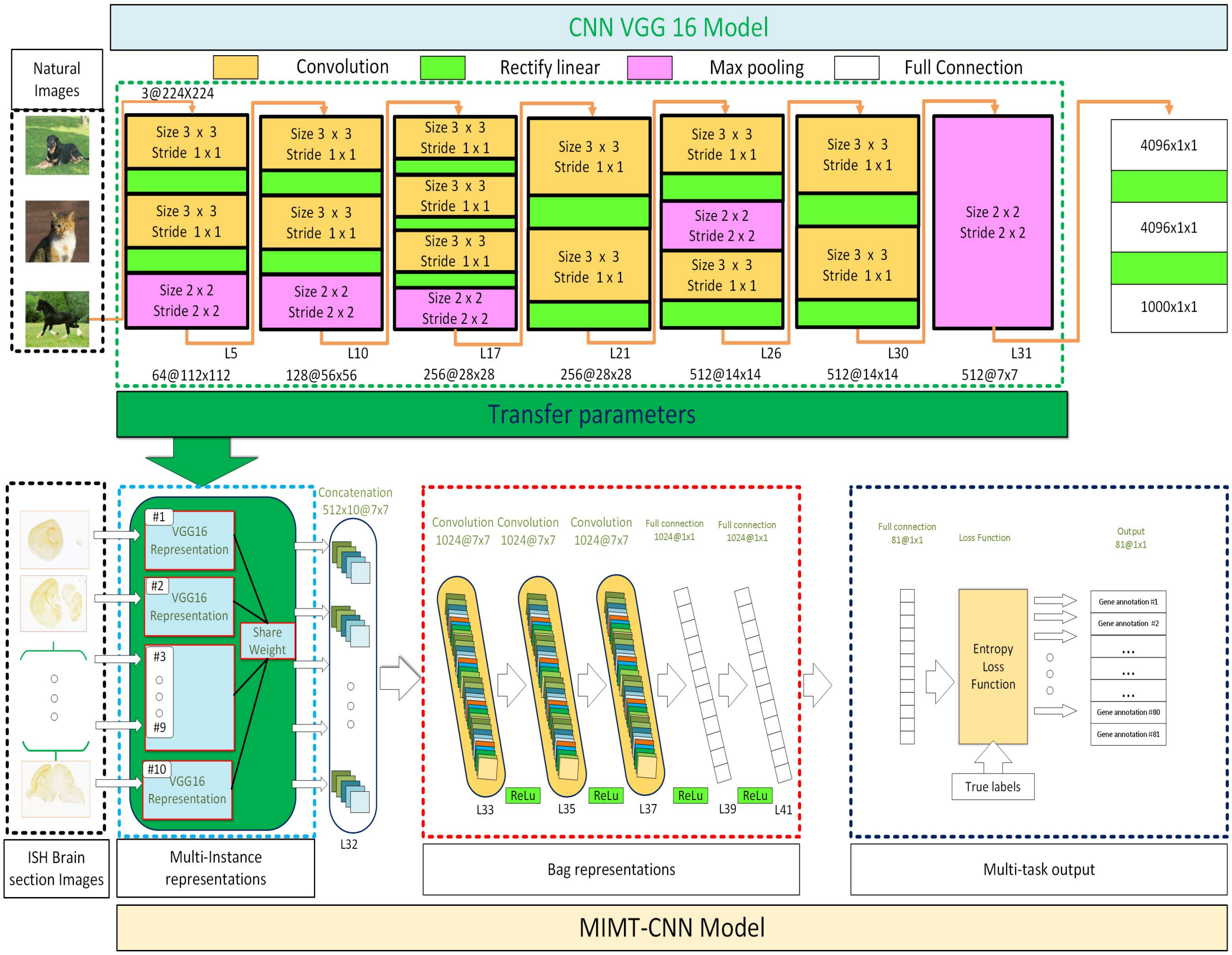

Fig. 2. Illustration of the proposed deep models for multi-instance multi-task learning. The network contains multiple CNN models that all share the same set of parameters pre-trained on the ImageNet data set (top row). The pre-trained parameters are transferred to our deep model and then fine-tuned with the domain-specific biological images (bottom row). This model accepts multiple ISH section images as inputs. Each image is processed by a sub-network to produce instance-level features. These sub-networks share the parameters that are initially transferred from the pre-trained model and then fine-tuned using the target data sets.

sections is not given. Therefore, the learning task requires a model to learn the mapping from a serial of raw ISH image sections to gene expression annotations for each of the anatomical structures in the brain. We considered image sections as "instances" and the set of sections representing a brain corresponds to a "bag". In this problem, gene expression annotation for each of the brain structure is only assigned to bag, and individual instance dose not convey explicit information for such annotation. Such task can be considered as a generalized form of multi-instance learning. In addition, it is also a multi-task learning problem since it involves multiple brain structures. Based on such relationship between multiple section images of the brain and annotation terms of brain structures, we thereby tackle this problem under the framework of multi-instance multi-task learning (MIMT). The proposed deep learning framework is illustrated in Figure 2, and the complete details are given in the following sections.

B. Deep Models for Transfer Learning

One advantage of CNN is that the entire network is capable of learning features from raw image data without prior knowledge. The main disadvantage of this model, however, is that it requires a large number of training samples in order to achieve competitive performance. Thus, a key challenge in applying CNNs to biological image data is that the available labeled training samples are very limited.

Recent studies used the ImageNet, a data set with thousands of categories and millions of labeled natural images, to train a CNN model. The learned model was then applied to other image data sets for features extraction. Such approach of transferring knowledge from one image data set to another yields superior performance on a wide variety of object recognition tasks [16]–[19]. Such idea of transfer learning aims to achieving competitive performance with limited number of samples by fully leveraging the existing knowledge learned from different but related tasks with a large sample size.

Hence we propose to employ transfer learning to transfer knowledge from well-known models trained on large natural image set to overcome this difficulty. We explore whether this transfer learning property of CNNs can be generalized to biological images. Specifically, the CNN model was trained on the ImageNet data containing millions of labeled natural images with thousands of categories and used directly as feature extractors to compute representations for ISH images. In this work, we apply the pre-trained 16-layer "VGG" model [21].

C. Deep Models for Multi-Instance Learning

In this paper, we propose a novel deep model for multi-instance learning. This model is then applied to annotate a given set of ISH section images across the developing mouse brain with gene expression terms for a set of brain structures. Two simplifications are applied to ease the computational cost of assessing our proposed model. First, we limit the number of brain structure by choosing the annotation at ontology level 5 that contains 81 brain regions. Second, we convert the multi-class problem (undetected, full, reginal, and gradient classes) into binary class problem; namely, "undetected" versus other detectable gene expression classes due to the abundance of the "undetected" class. Note that our problem is still a multi-task learning problem in which each task is a binary-class classification task.

Multi-instance learning problems require that the learning model is able to receive multiple instances as inputs while leveraging the knowledge of other pre-trained model. Note that all existing well-known pre-trained models on large data sets were trained based on single-instance basis, and the labels are associated with each image. However, the gene expression annotation for a structure is assigned to a set of ISH section images. This poses a key challenge when attempting to transfer knowledge learned by conventional architecture to multi-instance architecture. To overcome such difficulty, we embed a well-known model into our deep learning architecture. Specifically, we propose to replicate the pre-trained model to form multiple sub-CNN architectures running inside our deep learning architecture with weights being shared among them as in the Siamese architecture used in signature and face verification applications [22], [23]. Thus, each of such sub-models is able to form representations for individual instance in the bag. In addition, they are capable of transferring knowledge learned by conventional architecture that was trained on large image set. More specifically, we embed 10 parallel architectures of the VGG model with the last two fully connected layers being removed into our proposed network model. Such architecture allows for a bag of 10 instances as input corresponding to 10 ISH section images of the brain. It produces instance-level representations based on the knowledge learned from the ImageNet for each individual ISH image.

To learn bag-level representations from instance-level representations, we further combine all output feature maps produced by the last max pooling layers of all shared-weight VGG sub-networks. We then add 3 convolutional layers and 2 fully connected layers to our deep model. We expect such additional layers to be capable of capturing the complex relationship between instances and thus form global representations for the gene expression annotation at brain structure level.

D. Deep Models for Multi-Task Learning

In ADA, the manual gene expression annotation is given to each of brain structures, producing multiple annotations covering the brain. Similarly, a learning model for such task is required to learn multiple annotations at brain structure level when presented with a set of ISH section images. Considering the relationship between the input of multiple section images and the output of multiple brain structural level gene annotations, the annotation problem is a multi-instance multi-task (MIMT) problem.

Note that in multi-instance learning, a sample has many input descriptions (instances), and the objective is to study the relations in the input space. In contrast, multi-task learning studies problem in the output label space, where an sample has many output descriptions (labels). MIMT integrates multi-instance learning and multi-task learning by considering the problem in the input and output space simultaneously.

As depicted in the above section, we employed the transfer learning approach to capture instance-level representations and added additional neural network layers to form global representations of bags from outputs of those instance-level representations. Such global representations are designed to learn two types of complex relations; namely, the relations between instances and the multiple gene expression annotations at brain structure level.

The pre-trained model inside our deep learning model was originally trained to recognize objects in natural images, while our deep model is designed to study gene expression annotations at brain structure level. Although the leveraged knowledge from the source task could reflect some common characteristics shared in different image sets such as corners or edges, extra efforts are needed to capture the properties of individual ISH images and the global properties of the brain. Hence, we first initialized the parameters of the pre-trained model through transfer learning approach, and set the parameters of the last 5 neural network layers with standard initialization. Subsequently, we trained the entire model on the ISH image data set. This resulted in fine-tuned sub-networks capable of capturing instance level representations specific to the properties of individual ISH image, and concurrently, the bag representations capturing multiple-to-multiple relation between a set of image instances and gene expressions of brain structures.

In the last layer of our CNN model, we have multiple outputs, each of which corresponds to a gene expression annotation task for a given brain structure. Since these outputs are fully connected to a hidden layer that they share, the internal representations learned by one task could be used by other tasks. Note that the back-propagation is done in parallel on these outputs in the network. For each task, we used its individual loss function to measure the difference between outputs and the ground truth. In particular, we are given a training set of k tasks $\{X_i, y_i^j\}_{i=1}^m$, $j = 1, 2, \ldots, k$, where $X_i \in R^n$ denotes the i-th training sample, m denotes the total number of training samples. The output label y_i^j denotes the gene expression annotation status of training sample, which is binary with the form

$$y_i^j = \begin{cases} 1 & \text{if } X_i \text{ is annotated with the “undetected”} \\ & \text{class for } j\text{-th brain structure.} \\ 0 & \text{otherwise.} \end{cases}$$

To quantitatively measure the difference between the predicted annotation results and ground truth from annul annotation, we used the following multi-task loss function:

$$loss(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_{i=1}^{m}\sum_{j=1}^{k}\Big(y_i^j \log \hat{y}_i^j + (1 - y_i^j)\log(1 - \hat{y}_i^j)\Big)$$

where $\mathbf{y} = \{y_i^j\}_{i,j=1}^{m,k}$ denotes the ground truth label matrix over different tasks, and $\hat{\mathbf{y}} = \{y_i^j\}_{i,j=1}^{m,k}$ is the output matrix of the softmax layer [24]. Note that our multi-task loss function is the sum of the loss functions of individual tasks, and all these functions are optimized jointly during training. Also, all tasks share the same internal representation.

IV. Experimental Evaluation

A. Experimental Setup

To train and evaluate our proposed model, we built four data sets, one for each developing stage. Each data set consists of ISH section images corresponding to approximately 2,000 genes. We observed that the class “undetected” dominates in manual annotation. To alleviate this class imbalance problem, we simplified the classification tasks to a set of binary-class problems, where one class corresponds to the undetected category, and the other class includes all remaining categories. Additionally, training samples were selected by maximizing the class balance. That is, we randomly selected training samples from the data set and examined whether the ratios between two classes among all structures were above a certain threshold. We repeated this process for a maximum of 5,000 times and then decreased the threshold if the ratio was not satisfied. Thus, the final thresholds are different for different data sets. By converting the annotation tasks into a two-class problems, the three metrics (pattern, intensity, and density) resulted in the same set of binary classification tasks.

In ADA, each ISH experiment generated 12-15 sagittal image sections across the brain to detect a specific brain-wide gene expression pattern. It is highly desirable that all the ISH image sections associated with an experiment are presented to the model for learning simultaneously in order to fully exploit the available information for each sample. Technically, the weight sharing architecture of our deep CNN allows the networks to take an arbitrary number of instances as input. However, since our CNN architecture takes advantage of GPU machine, a large number of instances could result in large intermediate data, leading to insufficient GPU memory problem. Hence, we fixed the number of input instances in the bag. Specifically, we divided the transverse plane into 10 evenly-distributed intervals from lateral to central, and assign each with a sagittal section image that has coordinates closest to it. This is, for each gene, we obtained a bag with fixed number of 10 ISH images and assigned the manual annotations of all brain structures as multiple labels to the bag. Moreover, since we limit our experiment at brain ontology level 5 which contains 81 regions, a bag in the final sample data contains 10 ISH images and is assigned with 81 binary labels corresponding to gene annotations for the 81 brain regions.

In this work, we aim at automating the gene expression annotation task for the ADA data sets. To this end, we trained the aforementioned deep learning models using the ADA data sets and used the models as feature extractors. The classification models using these features are subsequently trained to perform such tasks. To compare image representations from the proposed models with other methods, we also obtained the image representations of bag-of-words coding and those of the CNN model pre-trained on the natural image set. Given the training set and test set, the goal of our tasks is to detect a set of gene expression for all brain structures. We partitioned the entire data set into training and test sets so that 2/3 of the data were in the training set, and the remaining 1/3 were in the test set. For all different features, the same training and test sets were used.

Using the feature vectors generated from our proposed deep models, our task is to automate the annotations of gene expression for a given developing stage. We trained a classification model to annotate gene expression patterns for each brain structure. The ℓ_2-norm regularized logistic regression classifier was used in the annotation. For each annotation task, we used the area under the ROC curve (AUC) as the our primary performance measure. We first compare the AUC values achieved using features of various layers in our model. We then used features from the best layer from our deep model and the pre-trained model to compare with the bag-of-word representations. We report this comparison using AUC, accuracy, sensitivity and specificity metrics. Moreover, we reported the AUC achieved for each individual brain structure and the overall AUC for gene annotation tasks across all brain structures.

B. A Baseline Image Representation

To compare the proposed model with other method, we considered a baseline approach based on the bag-of-words representation that has been widely used in modeling natural and biological images [25]–[29]. To obtain robust representations that are invariant to various distortions on the images, scale-invariant feature transform (SIFT) descriptors were applied on local patches of ISH images that were down-sampled by a factor of 4 to reduce the computational cost [1]. We applied dense SIFT feature descriptors on the ISH images implemented in the VLFeat software package [30]. This generated approximately 20,000 SIFT feature vectors from each ISH image section.

The bag-of-words approach requires a visual codebook for vector quantization. To this end, we randomly sampled the nonzero descriptors for each image to obtain a descriptor pool of size 100,000. The K-means algorithm was applied to cluster the SIFT descriptors in this pool, and the resulting cluster centers were considered as visual words in the codebook. We repeated the K-means algorithm multiple times with random initializations and used the one with the smallest within-cluster distance, since initialization may impact the results of K-means algorithm. For a given ISH image, we counted the number of occurrences of each visual word to form a global histogram representing an entire image.

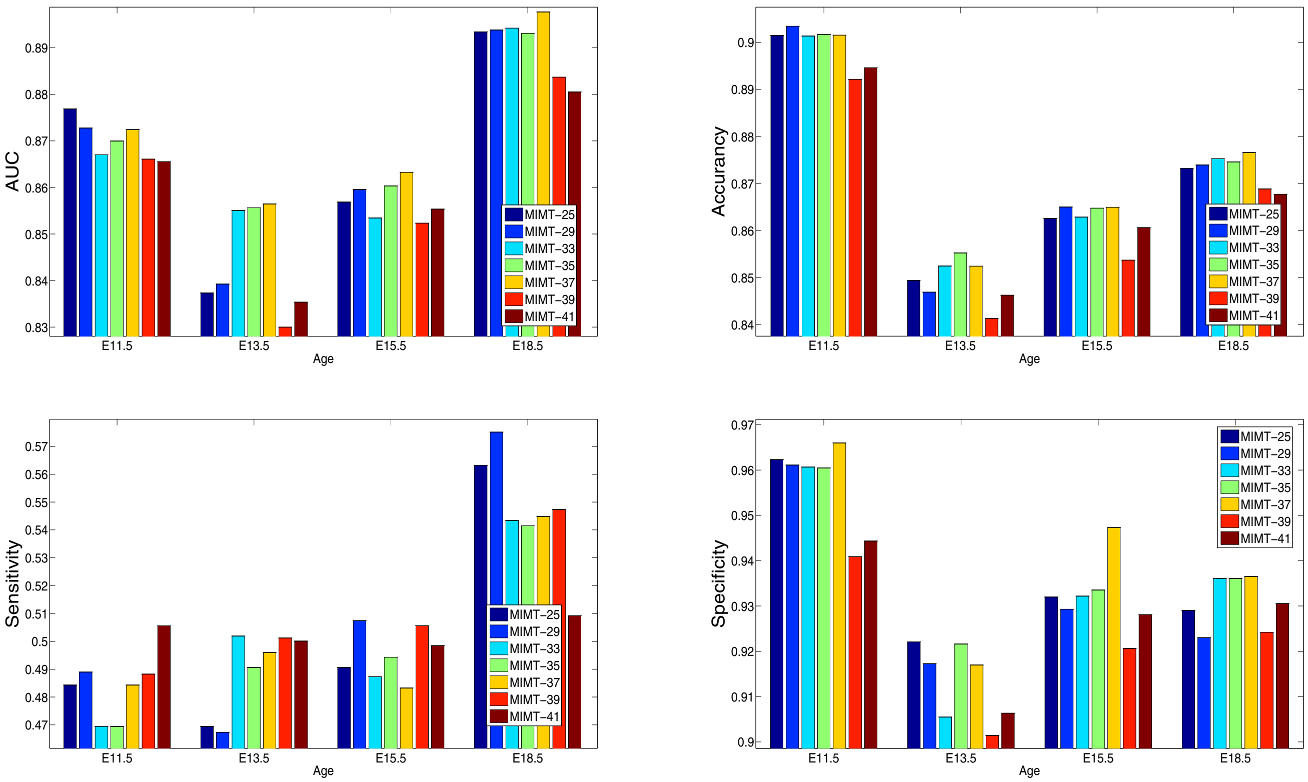

Fig. 3. Performance achieved for gene expression annotation tasks averaged over all brain structures using different layer features in our deep learning model. Each figure shows the annotation performance for one developing stage. The numbers indicate the network layer from which the features were extracted.

To represent gene expression patterns covering the entire 3-D brain, we divided the brain sagitally into seven intervals. Each ISH image section was assigned to one of the seven intervals based on its spatial location. The corresponding bag-of-words representations of ISH images assigned to the same interval were averaged to reflect regional sagittal gene expression. The global representation was built by concatenating the seven regional bag-of-words vectors.

C. *Comparison of Different Network Layers*

The deep learning model is constructed in such way that hierarchical feature representations are formed from low level to high level as the depth of network increases. When such model is used as a feature extractor, a natural question is which layer has the most discriminative power to capture the characteristics of input images. When such networks are trained with natural images such as the ImageNet data, the feature representations in the lower layers are expected to be generic features such as edge and corner detectors. In contrast, the features in higher layers are expected to represent objects specific to the training set. Hence, for the task of natural object recognition, the features extracted from higher levels usually yielded better discriminative performance [19].

In the architecture of our deep model, lower layers of neural networks carry the prior knowledge learned from large natural image data set. On the other hand, the higher layers abstract global representations from multiple instance-level representations yielded by lower layers. More specifically, lower layers consist of multiple sub-CNNs similar to the architecture of the pre-trained models. These sub-CNNs, each represents an instance of ISH image, share the same parameters that are initially transferred from the pre-trained models. In addition, the output feature maps from sub-CNNs are concatenated and used as inputs to several convolutional layers and full connection layers. These newly added layers combine the representations from each individual instance and are expected to form bag level representations; namely, the brain-wide gene expression representations in our study. Thus, we constructed our deep models by applying transferring learning approach for lower layers and initialized the parameters in the higher layers with standard randomization approach. We then fine-tuned the model using the ISH images from ADA.

Several interesting questions arise as we construct our deep model. First, how the discriminative power of higher level features changes from being specific to natural images to being specific to ISH images after fine-tuning? Second, can the newly added extra layers, that are designed to capture complex global representations of bags (in our case, the entire brain), yield features that have better discriminative power than the simple combination of all features at instance level?

To identify the most discriminative features for the tasks of gene expression annotation at brain structure level, we

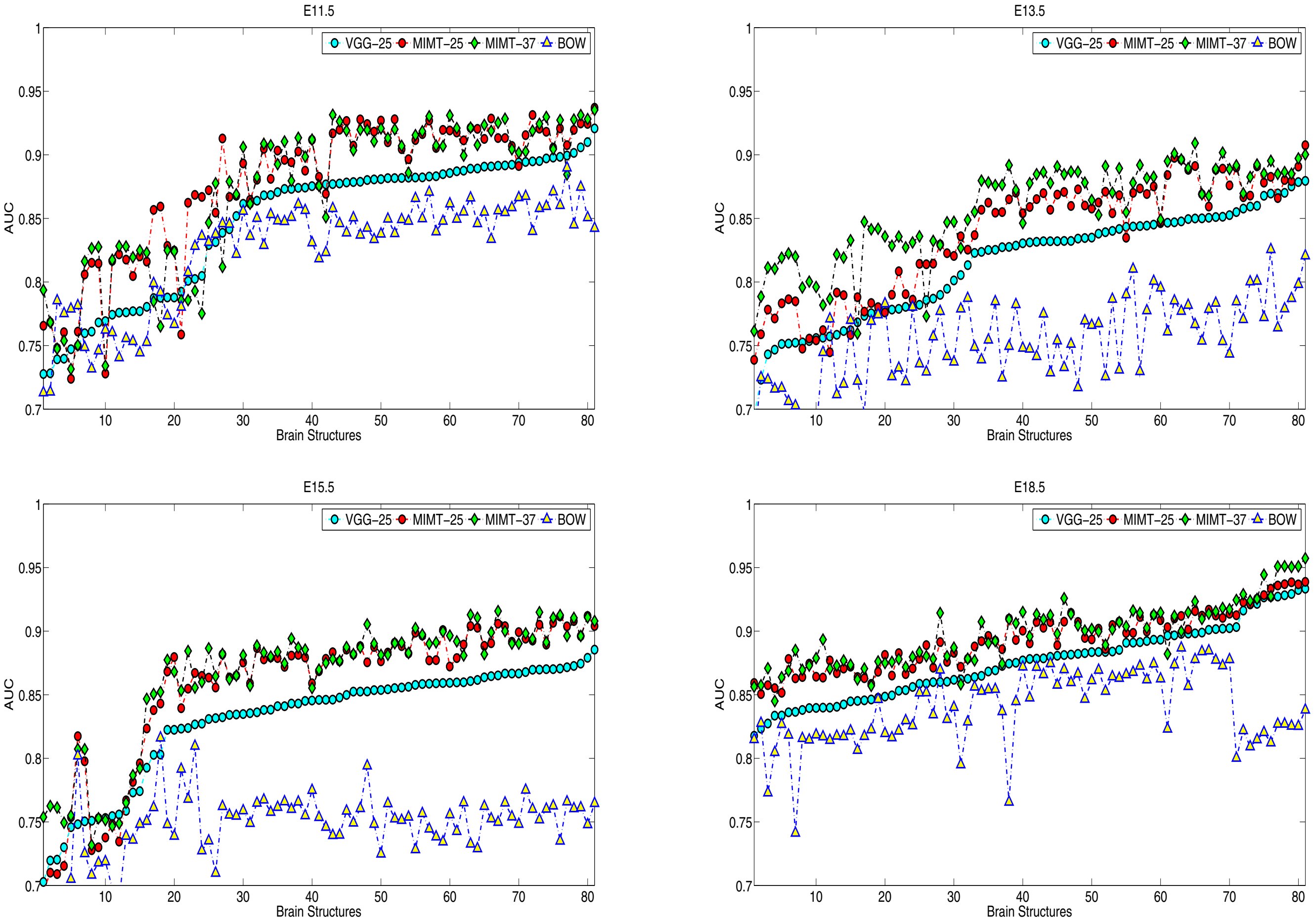

Fig. 4. Performance of different features achieved at each individual brain structure. Each figure shows the annotation performance for one developing stage. "BOW", "VGG", and "MIMT" denote the performance achieved by the bag-of-words, the pre-trained VGG model and our proposed multi-instance multi-task model, respectively. The numbers indicate the network layer from which the features were extracted. Note that, for better visualization, the brain structures along the x-axis were sorted based on the AUC values achieved by the VGG-25 features.

compared the features extracted from various layers from the pre-trained VGG network and the MIMT-CNN model that was fine-tuned from VGG. For each gene annotation task, we selected 10 ISH section images out of those available as depicted in Section IV-A, and each was used as input to the pre-trained VGG network. As lower layers produced large feature vector in VGG, we only extracted such image-level features from layers 25 and 29 to reduce the computation cost. The element-wise maximum operation was applied to those features to produce brain-wide representation. Similarly, the same 10 ISH section images were used as multi-instance inputs to our proposed deep models that produced instance-level and bag-level features. For features extracted from instance level representations, we again applied the maximum operation on features extracted from layers 25 and 29 to form brain-wide representations, respectively. We can observe from Figure 3 that the bag-level features from layer 37 outperform those at instance level (ISH image level) at ages E13.5, E15.5 and E18.5. They achieved slightly lower performance at age E11.5. Note that features from the last two full connection layers (39, 41), however, possess much less discriminative powers. This indicates that, in our models, bag-level representations in convolution layers can extract useful brain-wide representations from multiple instances, which are specific to gene annotations tasks of multiple brain structures. Overall, such representations have better discriminative power than those of brain-wide representation generated by simple combination of instance-level representations.

D. Evaluation of Annotation Performance

To assess the performance of automated gene expression pattern annotation tasks for our deep models, we compared the classification results averaged for all brain structures with those yielded by the pre-trained model and the bag-of-word representations using four metrics. Here, we selected feature extracted from layer 37 since it possess the best overall discriminative power among other layers in our MIMT-CNN model. Similarly, we used features from layer 25 in the pre-trained model due to their higher performance as compared

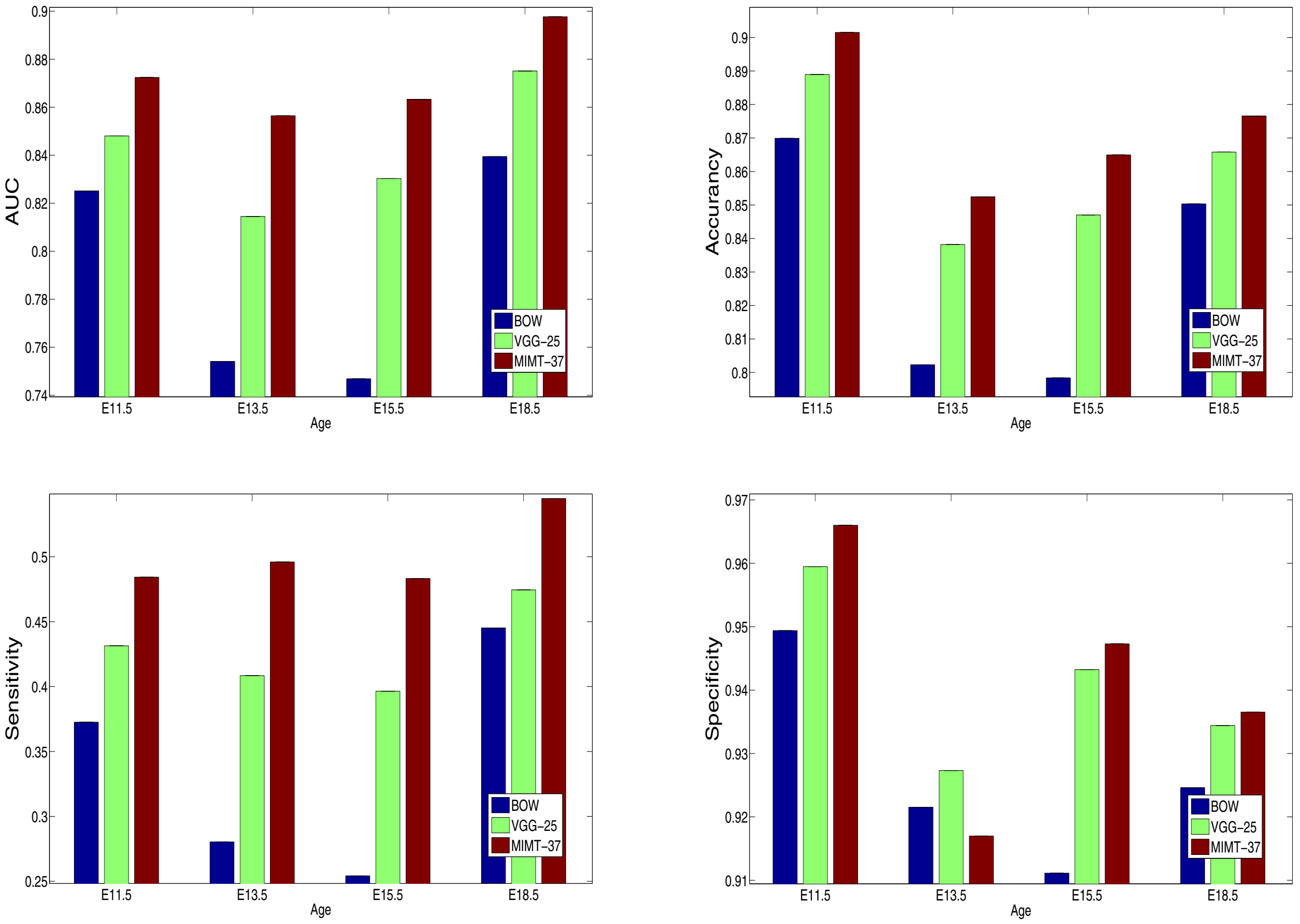

Fig. 5. Performance comparison for the proposed MIMT deep models, the pre-trained model and the bag-of-words approach. Each figure shows the annotation performance for one developing stage. "BOW", "VGG", and "MIMT" denote the performance achieved by the bag-of-words, the pre-trained VGG model and our proposed model, respectively.

with those from layer 29. Results in Figure 5 show that our proposed MIMT-CNN model outperforms the pre-trained and the bag-of-word models using the metrics of AUC, accuracy, and sensitivity for all four ages. However, the pre-trained model yielded the highest specificity for data from age E13.5. Overall, our MIMT-CNN deep modal achieved significant improvement for gene expression pattern annotation tasks with 3-4 percent increase in terms of AUC metric.

In Figure 4, we show how the performance of gene expression pattern annotation task changes for each individual brain structure using the pre-trained and our MIMT models. We noticed that for those structures that the pre-trained models achieved AUC lower than approximately 0.8, the performance of our MIMT models tend to fluctuate. For structures that the pre-trained model achieved relatively high AUC values, our MIMT models can improve the performance consistently and significantly. These observations indicate that some structures may possess properties that are significantly different with majority of other structures such as volume and morphology. It would be interesting to carry out further experiments to address questions such as what are the factors that affect the results of classification tasks. However, it requires the manual segmentation of structure for ISH images and such information is currently not available.

V. Conclusion and Future Work

In this work, we propose a deep CNN model to tackle the problem of multi-instance multi-task learning. We embed multiple well-known CNN models into the architecture of our model by replicating all layers of the network. We then propose to transfer the corresponding image-level knowledge learned on large natural image sets. Such sub-CNNs share the parameters and each computes instance-level representations for an input image (instance) in the bag. Subsequently, the outputs of all sub-CNNs are aggregated as inputs to an additional network consisting of convolutional and fully connected layers to produce the multi-task predictions. We train our models on the Allen Developing Mouse Brain Atlas data, which naturally fit the problem of multi-instance multi-task learning. This leads to fine-tuned features specific to the target data set at instance level and bag level features representing the complex multiple-to-multiple relations. Experimental results show that the feature extracted from our models achieve higher performance in comparison with those of the pre-trained VGG model and bag-of-words representations. We also show that the bag-level representations from hierarchical multiple neural network layers generate more discriminative features than those formed by simply combing instance level representations. Overall, our new deep CNN models demonstrate the potential of CNN to capture ambiguous multiple-to-multiple relation in multi-instance multi-task learning on a data set with a limited number of labeled samples.

In the current work, we merge the multiple detectable expression pattern classes and treat the annotation task of each region as a binary-class problem. It would be interesting to consider each of the detectable expression pattern classes as a separate class, thereby leading to multi-class, multi-task learning problems. We will explore such learning scenarios in the future. In addition, the multiple detectable expression pattern classes are ordinal; namely some ordered relationship exists among the classes. We plan to modify the loss function to incorporate such ordinal class information in the future. In the Allen Developing Mouse Brain Atlas, the brain regions are organized into a direct acyclic graph structure. In the current study, we do not consider this hierarchical relationship in formulating the multi-task learning. We will explore hierarchical multi-task learning models in the future to incorporate this information into the deep models.

Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research. This research is supported by NSF DBI-1147134, and NSF CAREER Award DBI-1350258.

References

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1106–1114.

[4] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[5] F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1719–1726.

[6] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *Proceedings of the 15th international conference on Multimedia*. ACM, 2007, pp. 17–26.

[7] J. D. Keeler, D. E. Rumelhart, and W. K. Leow, "Integrated segmentation and recognition of hand-printed numerals," in *Advances in neural information processing systems*, 1991, pp. 557–563.

[8] O. Maron and A. L. Ratan, "Multiple-instance learning for natural scene classification." in *The 15th International Conference on Machine Learning*, vol. 98. Citeseer, 1998, pp. 341–349.

[9] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," *Advances in neural information processing systems*, pp. 570–576, 1998.

[10] J. Wang and J.-D. Zucker, "Solving multiple-instance problem: A lazy learning approach," in *Proceedings of the 17th International Conference on Machine Learning*. ACM, 2000, pp. 1119–1125.

[11] Y.-X. Li, S. Ji, S. Kumar, J. Ye, and Z.-H. Zhou, "Drosophila gene expression pattern annotation through multi-instance multi-label learning," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 9, no. 1, pp. 98–112, 2012.

[12] S. Scott, J. Zhang, and J. Brown, "On generalized multiple-instance learning," *International Journal of Computational Intelligence and Applications*, vol. 5, no. 01, pp. 21–35, 2005.

[13] N. Weidmann, E. Frank, and B. Pfahringer, "A two-level learning method for generalized multi-instance problems," in *Proceedings of the 14th european conference on machine learning*. ACM, 2003, pp. 468–479.

[14] Allen Institute for Brain Science. (2013) ALLEN Developing Mouse Brain Atlas Technical White Paper: Expert Annotation of ISH Data. http://developingmouse.brain-map.org.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[16] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31st International Conference on Machine Learning*, 2014, pp. 647–655.

[17] M. Oquab, I. Laptev, L. Bottou, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[18] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2014, pp. 512–519.

[19] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision 2014*. Springer, 2014, pp. 818–833.

[20] Allen Institute for Brain Science. (2013) Allen Developing Mouse Brain Atlas [internet]. http://developingmouse.brain-map.org.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[22] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a Siamese time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 669–688, 1993.

[23] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1735–1742.

[24] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[25] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proceedings of the European Conference on Computer Vision*, 2006, pp. 490–503.

[26] S. Ji, L. Sun, R. Jin, S. Kumar, and J. Ye, "Automated annotation of drosophila gene expression patterns using a controlled vocabulary," *Bioinformatics*, vol. 24, no. 17, pp. 1881–1888, 2008.

[27] N. Liscovitch, U. Shalit, and G. Chechik, "Funcish: learning a functional representation of neural ish images," *Bioinformatics*, vol. 29, no. 13, pp. i36–i43, 2013.

[28] S. Ji, Y.-X. Li, Z.-H. Zhou, S. Kumar, and J. Ye, "A bag-of-words approach for drosophila gene expression pattern annotation," *BMC bioinformatics*, vol. 10, no. 1, p. 119, 2009.

[29] L. Kirsch, N. Liscovitch, and G. Chechik, "Localizing genes to cerebellar layers by classifying ish images," *PLOS computational biology*, vol. 8, no. 12, p. e1002790, 2012.

[30] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1469–1472.