COMPACT REPRESENTATION LEARNING USING CLASS SPECIFIC CONVOLUTION CODERS - APPLICATION TO MEDICAL IMAGE CLASSIFICATION

Uddeshya Upadhyay[†], *Biplab Banerjee*[⋆]

[†]Department of Computer Science and Engineering, Indian Institute of Technology-Bombay
[⋆]Centre of Studies in Resources Engineering, Indian Institute of Technology-Bombay

ABSTRACT

Medical image classification using deep learning techniques rely on highly curated datasets, which are difficult and expensive to obtain in real world due significant expertise required to annotate the dataset. We propose a novel framework called Class Specific Convolutional Coders (CSCC) to tackle the problem of learning highly discriminative, compact and non-redundant feature space from a relatively small amount of labelled images. We design separate attention-driven convolution network based feature extractors for the categories. These feature learning modules are further intuitively combined so as to make the whole image recognition system end-to-end trainable. Results on different medical image classification tasks show the advantages of our contributions, where our proposed methods outperforms the benchmark supervised deep convolutional networks (CNNs) trained from scratch.

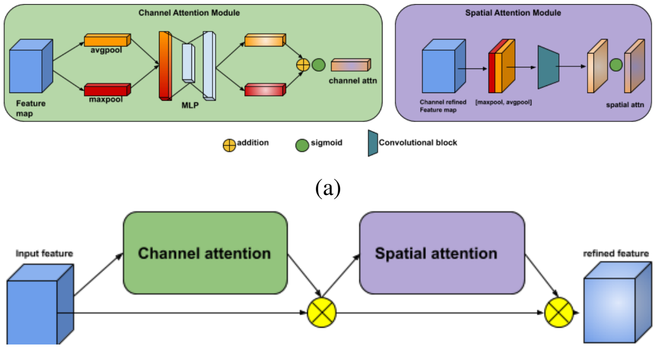
Index Terms— Deep Learning, Class Specific Convolutional Coders, Medical Imaging, Classification, Attention

1. INTRODUCTION AND RELATED WORK

The recent times have witnessed the rise of deep learning (DL) based models for several visual inference tasks [1, 2]. Success of DL models can largely be attributed to their data-driven hierarchical feature learning capability which is primarily made possible through the availability of a large volume of labeled training examples. In this regard, we note that DL models can be supervised, unsupervised or self-supervised, as far as training strategies are concerned. Popular self-supervised techniques usually follow the encoder-decoder architecture and focus to learn abstract feature representations corresponding to the input data. Typical examples of self-supervised DL models include different variant of regularized auto-encoders (AE): sparse, denoising, contractive, to name a few [3]. It is highly desired that the feature representations obtained from deep learning model should be class-wise distinctive, compact, and highly non-redundant in nature so as to produce improved classification performance. However, none of the aforesaid AE based models ensure the class compactness in the learned feature space since the label information is not explicitly utilized while training the models. A straight-forward solution in this respect is to train AE model both with reconstruction and classification loss functions under a multi-task learning setup, but this does not solve the problem to its entirety since classification loss alone cannot handle separation of fine-grained classes. Besides, several works rely on the deep metric learning techniques [4, 5] for discriminative feature space learning in terms of the triplet or Siamese architectures. However, such metric learning approaches require efficient mining of the data triplets or pairs which may be non-trivial at times. Another line of research in this respect considers the adaptive temperature scaling based soft-max [6] formulation while optimizing the cross-entropy type error measure. However, setting-up the temperature hyper-parameter is crucial and highly data dependent. Finally, the notion of non-redundant feature learning is largely ignored in these models.

Similar to other fields, DL models have shown impressive performance in the field of medical image analysis on myriad of tasks such as classification, segmentation [7, 8, 9], regression, etc. [10, 11, 12, 13, 14]. Unfortunately, obtaining annotations for a medical image dataset is often time-consuming and expensive since it requires highly experienced domain experts, for example, pathologists in case of digital pathology slides or radiologists in case of X-ray, MRI, CT scans. It is although observed that transfer learning and domain adaptation can enhance the performance of DL methods in presence of limited data [15, 16], however such methods may not be useful if there is a huge difference in the source and target domain, for example, source domain being natural images and target domain being medical images consisting of chest X-ray. Hence such a field will see great benefits from DL based methods where robust feature representations can be learned even with limited data.

Our contribution: Inspired by aforementioned issues in representation learning, we propose solution to the medical image classification problem by designing a novel set of class-specific feature extractors which are subsequently combined to produce an end-to-end image classification system. Each of the class specific convolutional coders (CSCC) follows a typical convolutional encoder-decoder architecture and learns a compact, discriminative, and highly non-redundant feature representations for a given class even from relatively small



(a)

(b)

Fig. 1: (a) Components of convolutional block attention module (CBAM). (b) pipeline showing application of CBAM.

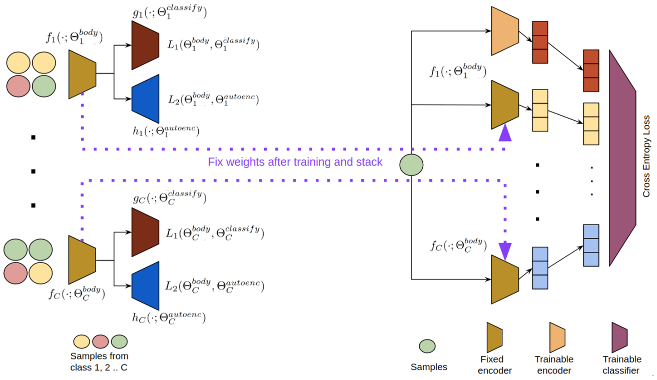

$g_1(\cdot;\Theta_1^{classify})$
$f_1(\cdot;\Theta_1^{body})$
$L_1(\Theta_1^{body},\Theta_1^{classify})$
$L_2(\Theta_1^{body},\Theta_1^{autoenc})$
$h_1(\cdot;\Theta_1^{autoenc})$
Fix weights after training and stack
$g_C(\cdot;\Theta_C^{classify})$
$L_1(\Theta_C^{body},\Theta_C^{classify})$
$L_2(\Theta_C^{body},\Theta_C^{autoenc})$
$f_C(\cdot;\Theta_C^{body})$
$h_C(\cdot;\Theta_C^{autoenc})$
$f_1(\cdot;\Theta_1^{body})$
$f_C(\cdot;\Theta_C^{body})$
Cross Entropy Loss
Samples from class 1, 2 .. C
Samples
Fixed encoder
Trainable encoder
Trainable classifier

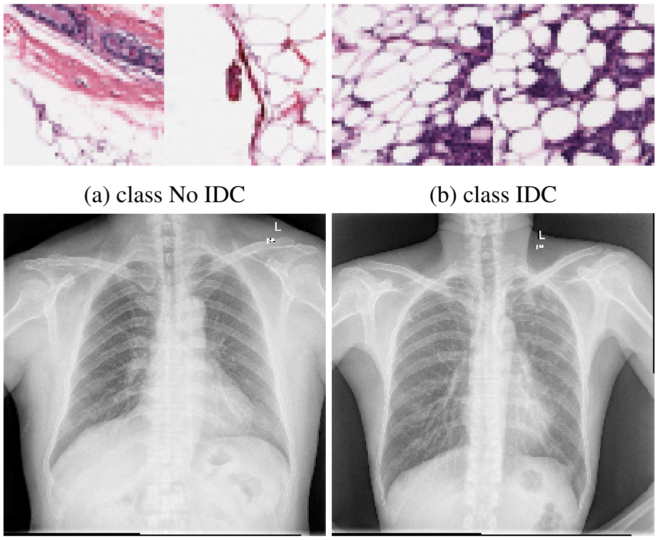
Fig. 2: Joint Class Specific Convolutional Coders (CSCCs) based framework. First Joint CSCCs are trained, one for each class. Trained CSCCs (with their weights fixed) are stacked together along with a secondary convolutional block with trainable parameters to generate the final features which is passed through a fully connected classifier.

training set. Besides, we judiciously incorporate spectral-spatial attention modules [17] in each of the convolution layers for improved feature learning. We conduct extensive experimentation on histopathology and chest X-ray image datasets using various DL models to showcase the efficacy of CSCC framework in comparison to more general classifiers.

2. PROPOSED METHODOLOGY

We first discuss the working principle of a given CSCC model and elaborate the proposed two-stage training procedure. We then discuss the inference scheme.

Training: Formally, let there be a training set $\mathcal{X} = \{x_i, y_i\}_{i=1}^{N}$ with images x_i and labels $y_i \in \{1, 2, \ldots, C\}$ where C defines the number of semantic categories. Further, let $N = N_1 + N_2 + \ldots + N_C$ where N_c accounts for the number of training samples corresponding to class c. Under this setup, we propose the c^{th} deep CSCC (referred to as Joint CSCC) $F_c(\Theta_c^{body}, \Theta_c^{classify}, \Theta_c^{autoenc})$ to be a convolutional encoder-decoder network based model directed to learning compact feature representations for the samples with class label c. The model broadly consists of three major sub-modules: a feature extractor (f_c) with learnable parameters Θ_c^{body} whereas the two decoder modules (aka heads) which carry out the following tasks simultaneously:



(a) class No IDC

(b) class IDC

(c) class No Tuberculosis

(d) class Tuberculosis

Fig. 3: Samples from training datasets. **(a)-(b)** IDC breast histopathology dataset with two classes. **(c)-(d)** Pulmonary chest X-ray abnormalities dataset with two classes.

head-1 (g_c) with parameters $\Theta_c^{classify}$ to perform binary classification, i.e, whether a given input image x belongs to class c or not.

head-2 (h_c) with parameters $\Theta_i^{autoenc}$ which performs cross-image reconstruction from a given input image of the c^{th} class to the mean of all the images belonging to the c^{th} class within a training iteration (or mini-batch).

While the classification loss ensures proper discrimination of the c^{th} class samples from all the other classes, the reconstruction loss, on the other hand, encourages high intra-class compactness by reducing the variance for the c^{th} class data. In order to train the CSCC, we sample a mini-batch of size $2 * K$ in every iteration as follows: i) images from the c^{th} class: $B_c = \{x_k^c, y_k^c = 1\}_{k=1}^{K}$, and ii) images from outside the c^{th} class: $B_{nc} = \{x_k^{nc}, y_k^{nc} = 0\}_{k=1}^{K}$, respectively. Nonetheless, it is ensured that B_{nc} proportionately contains samples from all the classes except the c^{th} class given the entire dataset. Considering the high within-class data variations, we further consider to incorporate attention modules within the CSCC model. Since we prefer to capture both the *what* and *where* aspects while highlighting the attentive regions within the images, the CBAM [17] method is considered in conjunction with the convolution blocks both along the encoder and decoder sides of the CSCC model. We define the loss functions to be minimized at **head-1**, **head-2**, and the latent space for each CSCC, respectively.

Classification loss at head-1: We consider a binary classification loss (L_1) using standard cross-entropy formulation,

$$L_1(\Theta_c^{body}, \Theta_c^{classify}) = - E_{x \in B_c \cup B_{nc}}[y \log(g_c(f_c(x))) + (1-y)\log(1-g_c(f_c(x)))] \quad (1)$$

where y is 1 for $x \in B_c$ and 0 otherwise. This loss directs proper discrimination between the c^{th} class samples and the rests, thus making the c^{th} CSCC to effectively learn the class boundary for the c^{th} class data.

Reconstruction loss at head-2: The cross-sample reconstruction loss (L_2) is defined specifically for the samples in B_c. However as opposed to [18] which performs reconstruction for a pair of training samples, here we propose to perform reconstruction of the mean of all the sample in the mini-batch B_c from a given $x \in B_c$. This, in effect, ensures a discriminative latent space. Besides, the proposed reconstruction loss provides two more advantages: i) the CSCC model can now be apparently trained with less training data, and ii) the idea of mini-batch compactness in B_c subsequently helps in achieving more compact class density in the latent space as a whole. In this respect, let $\mu_{B_c} = E_{x \in B_c}[h_c(f_c(x))]$ define the mean feature vector for B_c at head-2 and L_2 is defined as,

$$L_2(\Theta_c^{body}, \Theta_c^{autoenc}) = E_{x \in B_c}[||h_c(f_c(x)) - \mu_{B_c}||_2^2] \quad (2)$$

Orthogonality constraint in the latent representations: In order to ensure the non-redundancy in the learned feature encodings, we further incorporate a soft orthogonality loss (L_3) in the latent feature space as follows,

$$L_3(\Theta_c^{body}) = E_{x \in B_c}[||f_c(x)^T f_c(x) - I||_2^2] \quad (3)$$

where I denotes the standard identity matrix.

Total loss: The cumulative loss for training the c^{th} CSCC model is henceforth given by,

$$L(\Theta_c^{body}, \Theta_c^{classify}, \Theta_c^{autoenc}) := \lambda L_1(\Theta_c^{body}, \Theta_c^{classify}) + L_2(\Theta_c^{body}, \Theta_c^{autoenc}) + L_3(\Theta_c^{body}) \quad (4)$$

Note that λ is the hyper-parameter controlling the relative contribution of the classification error in L. The network is trained through the standard alternate optimization approach given that L is convex separately with respect to the individual parameter set. The C CSCC models are trained following the same experimental protocols. We also note that for a given sample, the class specific CSCC model produce low reconstruction error in the head-2 whereas all the remaining CSCC modules provide high reconstruction error at the head-2 decoder end. Once the individual CSCC modules corresponding to each class are trained, we further introduce a second network to learn the global image encoding and subsequent multi-class classification as shown in figure 2.

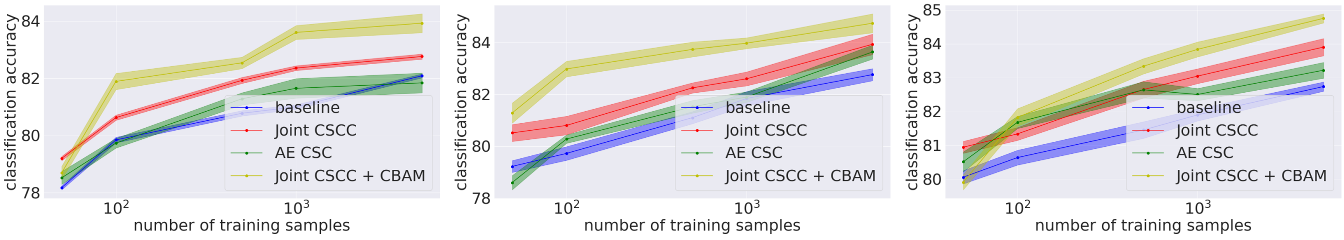
Given the trained CSCC modules, each of the training samples are passed through all the CSCC feature extractors (f_cs) and a feature concatenation is subsequently performed considering a pre-defined order of the CSCC modules. This is followed by another secondary backbone network with learnable parameters and the obtained feature representations are passed through a multi-class classification layer. This secondary network is trained end-to-end with the categorical cross-entropy based classification loss. Since the feature concatenation already produces a discriminative feature space, thanks to the discriminative characteristics of each CSCC module, the output of the secondary backbone ideally boosts the between-class separation effectively. A depiction of the entire training process can be obtained in Figure 2.

Inference: The inference can be accomplished in two ways: i) assessing the reconstruction error of each of the CSCC modules, and ii) propagating the test images through the primary CSCC modules and the secondary classification network, respectively. In the experimental study, we opt for the second approach.

3. RESULTS AND DISCUSSION

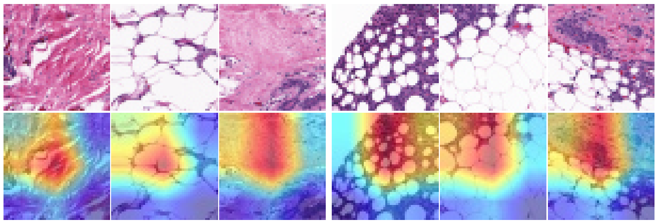
We study the performance of our method on two challenging medical imaging datasets dealing with breast histopathology and pulmonary chest X-ray. Typically, one has access to limited labeled training data in these cases, hence training a very deep traditional CNN model is non-trivial. In the following, the dataset descriptions and the experimental protocols followed are elaborated.

Invasive Ductal Carcinoma (IDC) breast histopathology: This dataset [19] consists of breast cancer histology images in the form of image patches from histology slides of various patients and the task is to predict whether a given image patch depicts the case for Invasive Ductal Carcinoma (IDC) or not. The original dataset consists of $277,524$ RGB image patches of resolution 50×50 which are derived from 162 HE-stained breast histopathology samples. Out of them $10,253$ images are positively labeled (i.e positive case for IDC) and remaining $198,738$ samples are negatively labeled (Figure 3). We consider multiple training-test splits (training set consisting of 50, 100, 500, 1000, and 5000 images evenly distributed between the classes in different splits while the rests are used for testing) to carry out the experiments. This setup aids in assessing the robustness of the model under different levels of supervision. For each CSCC, we consider three backbone architectures mimicking the design of VGG-11, VGG-16, and VGG-19 at the encoder end, respectively [20] while head-2 in the decoder reflects the design of the encoder in the reverse order. We compare the efficacy of the CSCC based model (Joint CSCC) with that of i) deep CNN models trained from scratch (Baseline Network), and ii) the CSC model of [18] (AE CSC). We also showcase the effect of CBAM in better feature learning by comparing our attention based CSCC



(a) (b) (c)

Fig. 4: Results on IDC histopathology dataset with different backbone models. **(a)** VGG-11. **(b)** VGG-16. **(c)** VGG-19.



(a) class No IDC (b) class IDC

Fig. 5: Visualisation of attention maps for IDC histopathology samples using the respective CSCC module

Model	Baseline	AE CSC	Joint CSCC	Joint CSCC + CBAM
ResNet-18	84.38	85.07	85.64	**86.12**
ResNet-34	85.13	85.64	86.25	**86.73**
ResNet-50	83.32	83.14	84.23	**85.14**
ResNet-101	83.26	83.21	84.14	**85.56**

Table 1: Performance (accuracy in %) comparison for the Pulmonary dataset with different ResNet based backbone architectures. Joint CSCC + CBAM model performs the best for all the models.

model (Joint CSCC + CBAM) with the one without any attention. Figure 4 showcases the performance measures of four techniques corresponding to the *Baseline network*, *CSC*, *Joint CSCC* and *Joint CSCC + CBAM* ($\lambda = 1$ is considered in Equation 4). In our experiment we vary the number of labeled samples in the training set and measure the corresponding changes in performance (accuracy in %) of the respective techniques. As the amount of labeled data increases the performance of all the models also improve. However, for each of the VGG-11, VGG-16, VGG-19 backbones, CSCC and CSCC+CBAM always outperforms the baseline and the CSC models. This clearly showcases that the compact representations learned using the CSCC model are better than other models. In addition, we perform grad-CAM [21] based visualization to showcase the effects of the attention model in highlighting important class-specific regions in each image. Figure 5 shows the output of Grad-CAM analysis on a few samples for histopathology dataset. For a sample belonging to one of the classes, the class activation map is obtained by using the CSCC trained for that class (only considering the backbone/body and classification head). Figure 3-(a,b) shows that the patches belonging to class "IDC" has a dominant presence of dark violet color in them, where as the patches belonging to class "No IDC" are predominantly light pink and white in color. The outputs of Grad-CAM in figure 5 show a similar trend where for the class "No IDC" the activation map focuses on light pink/white regions in the patch and for class "IDC" the activation map focuses on region with dark violet pigments.

Pulmonary chest X-ray abnormalities: This publicly available dataset (Figure 3-c,d) consists of Chest X-rays from Shenzhen and Montgomery [22]. The task is to detect whether given chest X-ray is a positive case for tuberculosis or not. The dataset consists of 800 scans in total and each scan has been labeled by trained radiologists as positive or negative case for Tuberculosis. We use 80% of the data as training set and 20% of data for testing. We follow experimental and evaluation protocols similar to IDC dataset. Since the dataset is very small we do not perform ablation study similar to IDC dataset and instead of VGG models, we follow the Resnet 18, 34, 50, 101 network structures [23] in our feature extractors in this case. However, any backbone/body model structure can be considered. Table 1 shows the results of our experiments on pulmonary chest X-ray dataset using different ResNet based backbone/body architectures. Amongst the different backbone models, we observe that ResNet-34 performs the best for this dataset while models with more parameters (i.e ResNet-50 and and ResNet-101) leads to overfitting on the training data. Although this dataset is comparatively small, it still demonstrates that the proposed CSCC and CSCC+CBAM outperforms the baseline and the CSC [18] models consistently.

4. CONCLUSIONS

We propose *Class Specific Convolutional Coders* (CSCC). The proposed CSCC takes inspiration from multi-task networks and representation learning ability of encoder-decoder [3]. We also incorporate spectral-spatial attention mechanism [17] in our CSCC framework which further improves our framework. We also propose new loss functions to learn highly discriminating features even in low data regime. We achieve state of the art performances.

5. REFERENCES

[1] Suraj Srinivas, Ravi Kiran Sarvadevabhatla, Konda Reddy Mopuri, Nikita Prabhu, Srinivas SS Kruthiventi, and R Venkatesh Babu, "A taxonomy of deep convolutional neural nets for computer vision," *Frontiers in Robotics and AI*, vol. 2, pp. 36, 2016.

[2] Naveed Akhtar and Ajmal Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[3] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[4] Elad Hoffer and Nir Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.

[5] Sofia Ira Ktena, Sarah Parisot, Enzo Ferrante, Martin Rajchl, Matthew Lee, Ben Glocker, and Daniel Rueckert, "Distance metric learning using graph convolutional networks: Application to functional brain networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 469–477.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[7] Thorsten Falk, Dominic Mai, Robert Bensch, Özgün Çiçek, Ahmed Abdulkadir, Yassine Marrakchi, Anton Böhm, Jan Deubner, Zoe Jäckel, Katharina Seiwald, et al., "U-net: deep learning for cell counting, detection, and morphometry," *Nature methods*, p. 1, 2018.

[8] Patrick Esser, Ekaterina Sutter, and Björn Ommer, "A variational u-net for conditional appearance and shape generation," in *CVPR*, 2018.

[9] Korsuk Sirinukunwattana, Josien P.W. Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J. Matuszewski, Elia Bruni, Urko Sanchez, Anton Bhm, Olaf Ronneberger, Bassem Ben Cheikh, Daniel Racoceanu, Philipp Kainz, Michael Pfeiffer, Martin Urschler, David R.J. Snead, and Nasir M. Rajpoot, "Gland segmentation in colon histology images: The glas challenge contest," *Medical Image Analysis*, vol. 35, pp. 489 – 502, 2017.

[10] Rodney LaLonde, Dong Zhang, and Mubarak Shah, "Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information," .

[11] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei, "Relation networks for object detection," .

[12] Longyin Wen Xiao Bian Zhen Lei Zhang, Shifeng and Stan Z. Li., "Single-shot refinement neural network for object detection," in *Computer Vision and Pattern Recognition (CVPR), IEEE*, 2018.

[13] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, vol. 1, no. 2, pp. 4, 2017.

[14] Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko, "Parallel feature pyramid network for object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 234–250.

[15] Shixing Chen, Caojin Zhang, and Ming Dong, "Coupled end-to-end transfer learning with generalized fisher information," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira, "Analysis of representations for domain adaptation," in *Advances in neural information processing systems*, 2007, pp. 137–144.

[17] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[18] Sanatan Sharma, Akashdeep Goel, Omkar Gune, Biplab Banerjee, and Subhasis Chaudhuri, "Class specific coders for hyper-spectral image classification," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3304–3308.

[19] ," https://www.ncbi.nlm.nih.gov/pubmed/27563488.

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

[21] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[22] ," www.ncbi.nlm.nih.gov/pmc/articles/PMC4256233/.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.