DEEP CONVOLUTIONAL ACTIVATION FEATURES FOR LARGE SCALE BRAIN TUMOR HISTOPATHOLOGY IMAGE CLASSIFICATION AND SEGMENTATION

Yan Xu[1,2], *Zhipeng Jia*[2,3], *Yuqing Ai*[2,3], *Fang Zhang*[2,3], *Maode Lai*[4], *Eric I-Chao Chang*[2*]

[1]Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education, Beihang University

[2]Microsoft Research, Beijing, China

[3]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China

[4]Department of Pathology, School of Medicine, Zhejiang University, China

{eric.chang, v-zhijia, v-yuai,v-fangz}@microsoft.com, xuyan04@gmail.com, lmd@zju.edu.cn

ABSTRACT

We propose a simple, efficient and effective method using deep convolutional activation features (CNNs) to achieve stat- of-the-art classification and segmentation for the MICCAI 2014 Brain Tumor Digital Pathology Challenge. Common traits of such medical image challenges are characterized by large image dimensions (up to the gigabyte size of an image), a limited amount of training data, and significant clinical feature representations. To tackle these challenges, we transfer the features extracted from CNNs trained with a very large general image database to the medical image challenge. In this paper, we used CNN activations trained by ImageNet to extract features (4096 neurons, 13.3% active). In addition, feature selection, feature pooling, and data augmentation are used in our work. Our system obtained 97.5% accuracy on classification and 84% accuracy on segmentation, demonstrating a significant performance gain over other participating teams.

***Index Terms*—** deep convolutional activation features, deep learning, feature learning, segmentation, classification

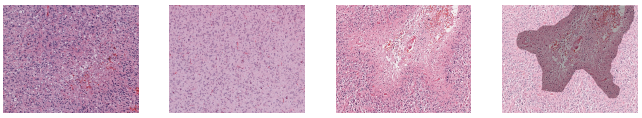
1. INTRODUCTION

Feature representation plays an important role in the medical image field. Since there are different image modalities, such as MRI, CT and digital histopathology images in the medical domain, even though images are acquired from the same patient for a certain disease, their morphologies, textures and color distributions vary significantly. For example, brain tumor scan images from MRI and histopathology show distinct patterns, making it hard to apply a general pattern for brain tumor detection on both image sources. Therefore, feature representation [1] is a top priority in high-level medical tasks such as classification and segmentation. A lot of research has focused on feature design of various types, such as object-like features and texture features. However, their applications are limited due to the special designs.

In addition, an insufficient amount of training data is another major concern in the medical domain. Since training data depends on the number of disease incidences, it is usually harder to collect than images of natural scenes. Also, the detailed annotation of medical images is a challenging task. Manual annotation is not only labor-intensive and time-consuming, but also intrinsically ambiguous even when labeled by clinical experts. Therefore, a limited amount of available training data is common in medical image tasks and indeed poses a great challenge to solving real-world problems using feature learning. In our case, there are only 45 images for classification and only 35 images for segmentation.

Deep convolutional activation features have achieved great success in computer vision in recent years [2, 3, 4, 5, 6, 7, 8]. The emergence of large image databases such as ImageNet, comprising more than 10 million images and more than 20,000 classes [6], makes it possible for CNNs to provide sufficient feature description for general images. In this paper, we explore the potential of using ImageNet knowledge via deep convolutional activation features to extract features for classification and segmentation, as highlighted in the MICCAI 2014 Brain Tumor Digital Pathology Challenge [9].

Glioma is a kind of brain tumor with several subtypes based on their glioma grade. High grade glioma includes anaplastic astrocytomas and glioblastoma multiforme [10]. The characteristics that distinguishes high grade glioma from low grade glioma (LGG) is the presence of necrotic regions in the glioblastoma multiforme and the presence of hyperplastic blood vessels and megakaryocytes [11]. Figures 1 (a) and (b) show samples from GBM and LGG histopathology images.



(a) (b) (c) (d)

Fig. 1. Samples of (a) GBM and (b) LGG; and samples of necrosis images. (c) Raw image. (d) Ground truth image; Gray mask represents necrosis.

In sub-challenge I, the task was the classification of glioblastoma multiforme (GBM) and low grade glioma (LGG) using digital histopathology images. A standard histopathology image can be scanned at a resolution as big as $100,000 \times 100,000$ pixels, which can contain about 1 million descriptive objects. Therefore, it is difficult to design special pathological features for distinguishing GBM and LGG. We introduced deep convolutional activation features to describe pathological features of brain tumors. In our method, the inputs need to be resized to 224×224 pixels to fit our CNN model trained by ImageNet. If an original image is resized to 224×224 pixels, pathologists cannot recognize it correctly. One key step in our approach for classification requires some tailoring to fit the properties of CNN features, namely feature pooling. Similar to the activity of visual neurons in the mammalian primary visual cortex, our CNN

*Corresponding author. This work was supported by Microsoft Research under eHealth program, Beijing National Science Foundation in China under Grant 4152033, Beijing Young Talent project in China, and the Fundamental Research Funds for the Central Universities of China.

activation feature vector is fairly sparse (4096 neurons, 13.3% active), which, according to [12] and cross-validation, indicates that 3-norm pooling may be more suitable to our task. Therefore, we adopt 3-norm pooling to integrate the final features for each image. In addition, feature selection is used to select a subset of more relevant features and to reduce redundant or irrelevant features. Finally, selected features are passed to a linear SVM [13] for classification.

Necrosis is a significant indicator to distinguish LGG from GBM. In sub-challenge II, the task was a segmentation of necrosis and non-necrosis regions from GBM histopathology images (See Figures 1 (c) and (d)). We cast the segmentation problem as a classification problem. An image is split into many patches as either necrosis or non-necrosis. Necrosis patches are considered as positive samples while non-necrosis are considered as negative. The features of patches are extracted by CNNs. A linear SVM [13] is applied to classify these patches of necrosis and non-necrosis. The discriminative probability (or classification confidence) maps for each pixel are created by the mean of the confidences of all the patches containing the pixel.

2. RELATED WORK

Feature representation design is a popular topic in histopathology images. Expert designed features include morphometric features [14], fractal features [15], texture features [16] and object-like features [17]. However, study [18] has pointed out that features learned by a two-layer network with non-linear responses using unlabeled image patches are superior to expert designed representations when it comes to histopathology images. Nayak [19] introduces sparse features learning using the restricted Boltzmann machine (RBM) to describe histopathology features in GBM and clear cell kidney carcinoma (KIRC). These two methods show that feature learning is better than special feature designs. However, a limited amount of training data is a universal challenge to feature learning. In our case, similar problems arose because there were only 45 images for classification and only 35 images for segmentation. Based on the above two points, we used CNN features trained by ImageNet to represent features in brain tumor histopathology images.

Deep CNN features used as generic descriptors are a growing trend. Some publicly available CNN models have been used to extract features: Caffe [20] is utilized in works [20, 3, 2] and OverFeat [21] is used by [8]. The CNN features are usually used in classification and object detection tasks [20, 3, 2, 8]. The above-mentioned related work only focuses on nature images. To the best of our knowledge, this is the first attempt to transfer CNN features to histopathology images and achieve state-of-the-art performance in a pathology image challenge.

3. ALGORITHMS

3.1. CNN architecture

The CNN model we used is generously provided by the CognitiveVision team in ImageNet LSVRC 2013 [5]. The CNN architecture is similar to the one used in [6], but without the GPU split, since modern GPUs have enough memory for the entire model. The graphical representation of the architecture is shown in Figure 2. Note that this model was trained on the entirety of ImageNet; thus it is not the same one the CognitiveVision team used in ILSVRC 2013. The code used for training and extracting features is based on [6]. During training, the data pre-processing and data augmentation techniques introduced in [6] were used, turning input images of various resolutions to 224×224 input for the network. For feature extraction, since input patches are already 224×224, no rescaling or cropping was needed. In this paper, 4096-dimensional output of the last hidden layer, i.e., the second-to-last fully connected layer, is used as our extracted feature vector (highlighted in Figure 2).

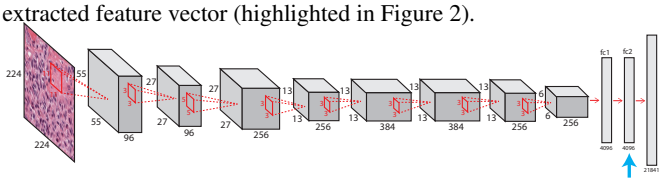


Fig. 2. The architecture of the CNN model used in this work. The blue arrow indicates the layer whose output is our CNN feature.

3.2. Classification framework

The vast size of histopathology images makes it necessary to extract features locally. To this end, we divide each histopathology image into a set of overlapping square patches with a size of 336×336 pixels determined by cross-validation. The patches form a rectangular grid with which adjacent patches have 25% overlap (i.e. 84 pixels). To reduce the number of patches with only white background, if the RGB values of all pixels in a patch are all greater than 200, that patch is discarded. All patches are then resized to 224×224 pixels to form 4096-dimensional CNN feature vectors. All the feature vectors of an image are computed over a 3-norm pool based on the theoretical analysis of feature pooling [12] and cross-validation, which yield a final single feature vector for the whole image. The equation of 3-norm pool is computed by $f_P(\mathbf{v}) = (\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_i^P)^{\frac{1}{P}}$, where P is 3, N is the number of patches of an image and $\mathbf{v}_i$ is the 4096 dimensional feature vector of the ith patch.

Feature selection is necessary in order to select a subset of more relevant features and to reduce redundant or irrelevant features. Features are selected based on the rank of difference between GBM and LGG. The difference of the kth feature dimension is computed as follows: $diff_k = \left|\frac{1}{N_{\text{GBM}}}\sum_{i=1}^{N_{\text{GBM}}} v_{ik} - \frac{1}{N_{\text{LGG}}}\sum_{i=1}^{N_{\text{LGG}}} v_{ik}\right|$ (for $k = 1, \ldots, 4096$, where N_{GBM} and N_{LGG} are the number of GBM and LGG in the training set, and v_{ik} is the kth dimensional feature of the ith image.). The top 100 features with the largest differences were chosen as our final features.

Finally, a one-vs-one linear SVM is used to classify GBM and LGG. The regularization parameters C of SVM are determined by cross-validation. Figure 3 shows the pipeline of our classification framework.

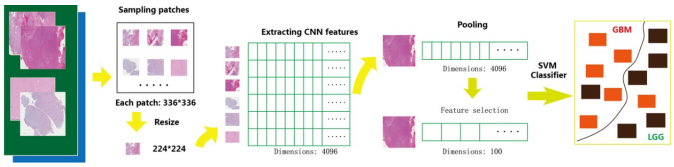


Fig. 3. Flow diagram of classification framework. The inputs include both GBM (positive) and LGG (negative) images. We sample patches of 336×336 pixels on a regular grid. Because the inputs of the CNN model are 224×224 pixels, all patches are resized to 224×224 pixels. A 4096-dimensional CNN feature vector is extracted for each patch. Feature pooling and feature selection are used to obtain a 100-dimensional feature vector for each image. A linear SVM automatically classifies GBM and LGG. Orange square: GBM; Brown square: LGG.

3.3. Segmentation framework

The segmentation methods of medical images can mainly be divided into three categories: unsupervised [17], weakly supervised [22] and

supervised learning [15]. Compared with supervised learning, unsupervised and weakly supervised methods often lead to inferior results. Therefore, we have chosen a supervised method to segment necrosis and non-necrosis in GBM.

In our work, we pose the segmentation problem of an image as a collection of classification problems on its patches. Figure 4 describes the pipeline of our segmentation framework. Patches are sampled on a regular grid at a size of 112×112 pixels in 8-pixel strides. If the necrosis area of a patch is greater than 50% of the patch area, the patch is labeled as a necrosis patch, and vice versa. The classification problem is to distinguish necrosis from non-necrosis. Necrosis patches are considered as positive instances while non-necrosis are considered as negative instances. All patches are resized to 224×224 pixels to obtain a 4096-dimensional CNN feature vector similar to the previous classification workflow. A linear SVM is used to learn the segmentation model. Since a pixel can be covered by overlapped patches with their respective labels, a confidence score of the pixel is computed based on the mean of the confidence scores of these patches containing the pixel from the SVM classifier. The discriminative probability (or classification confidence) maps for each pixel are created by the corresponding confidence scores. Next, we obtain segmentation results of necrosis by the threshold generated by cross-validation. We then conduct a post process, such as removing very some amounts of very tiny noise and filling some small holes.

In addition, we further make two changes to the training data for the last submission model. (1) We have observed that hemorrhage tissues appear in both necrosis and non-necrosis regions. Therefore, hemorrhage patches in necrosis regions are relabeled as non-necrosis patches. In the prediction stage, necrosis regions containing hemorrhages are predicted to be non-necrosis patches. The phenomenon usually appears in the interior area of necrosis. Thus, the post process recognizes these patches as necrosis. (2) We have observed that training images are nonuniform and have various sizes. The training data is not evenly distributed. In the last model for submission, we augmented the instances of missed regions and false regions generated by leave-one-out cross-validation on the training data.

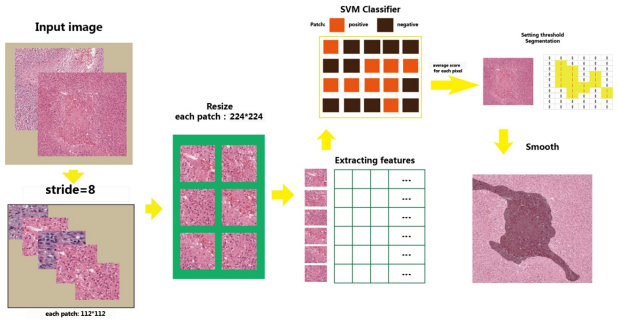


Fig. 4. Flow diagram of the segmentation framework. The inputs include both necrosis (positive) and non-necrosis (negative) images in GBM. We sample 112×112 patches on a regular grid (stride=8). All patches are resized to 224×224 pixels. A 4096-dimensional CNN feature vector for each patch is extracted. A linear SVM classifier distinguishes positive and negative. Probability mapping images are generated using all predicted confidences. After smoothing, necrosis segmentations are formed. Orange square: necrosis; Brown square: non-necrosis.

4. VISUALIZATION OF CNN ACTIVATION FEATURES

We visualize individual components of the responses of neurons in the last hidden layer (4096 dimensions) to observe the properties of CNN features. The degree of relevance of activated neurons is determined via the high-to-low ranking of weight values from the classification training model. For the relevant neurons, we select patches that activate them the most. Figure 5 shows some samples (each row stands for a relevant neuron).

One of the most significant hallmarks is distinct image appearances and similar semantic features, e.g. clinical features. For example, the property of cell heteromorphism (1st row), cell hemorrhage (3rd row), angiogenesis (5th row) and uniform cell size (9th row) can be discerned from the neurons in Figure 5 respectively. These clinical features can adequately distinguish GBM and LGG. The visualization demonstrates that we can successfully transfer CNN features to capture semantic features of medical images in nature.

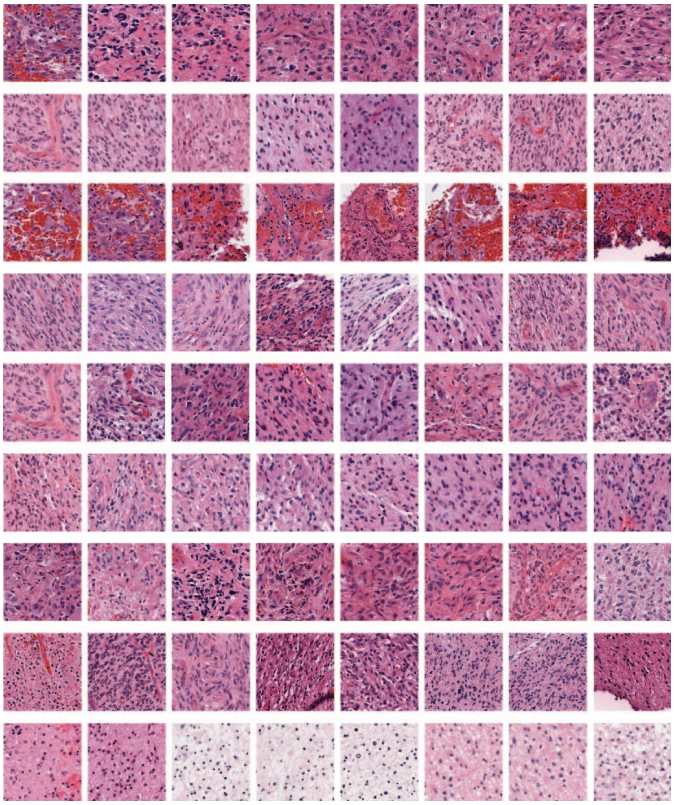


Fig. 5. Sample discriminative patches found with individual components (neurons) of the CNN activation features. Each row of patches causes a high response in one of the 4096 neurons. Note the variance of appearance properties in each row. All the images come from 45 training datasets in the classification task.

5. EXPERIMENTS

5.1. Datasets

The training data is provided by the organizers from the TCGA web [23]. In sub-challenge I, the training set includes 23 GBM images and 22 LGG images. The test set includes 40 images. In sub-challenge II, the training set includes 35 images. The average number of necrosis pixels of a training image is $1,330,000 \pm 1,520,000$. The average number of non-necrosis pixels of a training image is $2,900,000 \pm 3,790,000$. The test set includes 21 images.

5.2. Comparison

We made a comparison with our methods in both classification and segmentation. We compared CNN features with manual features for

generic object recognition in our two tasks. *Manual Feature (MF):* Generic object recognition features were chosen, including SIFT, LBP, and L*a*b color histogram. The feature dimension is 186. *CNN-F:* The last full connection layer was used to extract features (4096 dimensions).

Classification: *MCIL [22]:* The patch extraction setting is the same as our method. The softmax function here is the GM model and the weak classifier is the Gaussian function. The parameters in the algorithm are the same [22]. *Image-level SVM:* A whole pathology image is directly resized into 224×224 pixels. The CNN-F is used. *SVM-MF:* The features used are manual features. The rest is the same as our method. *SVM-CNN:* Our method.

Segmentation: *GraphRLM [17]:* The method is an unsupervised method to distinguish cancer or non-cancer in colon histopathology. The parameters in our experiment are set as: $r_{\min} = 8$, $r_{\text{strel}} = 2$, $win_{\text{size}} = 96$, $dist_{\text{thr}} = 1.25$, and $comp_{\text{thr}} = 100$. *SVM-MF:* The features used are manual. The rest is the same as our method. *SVM-CNN:* Our method.

5.3. Evaluation

In classification, accuracy is used as the evaluation method. In segmentation, given the ground truth map G_i and the probability map P_i generated by the algorithm, the score of an image is $S_i = \frac{2|P_i \cap G_i|}{P_i + G_i}$. The evaluation score (called accuracy) is the mean of S_i (for $i = 1, ..., K$, where K is the number of images).

5.4. Results

Classification Our final submission in the challenge achieves an accuracy of 97.5% on the test data, ranking first among other participants. The increase in accuracy compared to the second place is 7.5%. Table 1 summarizes the performances of some of the top-performing approaches. We also conduct the experiment only in the training data using 3-fold cross-validation. Table 2 compares the performances from the experiments. Compared with manual features (MF), CNN features are powerful in improving performance. Due to the characteristics of large scale images, the image-level SVM method was the worst at recognizing LGG and GBM. Compared with the MCIL algorithm, our method was 6.7% better. Compared with SVM-MF, SVM-CNN improved from 77.8% to 97.8%.

Table 1. Classification performance in the challenge

	Accuracy	Place
Anne Martel	75.0%	4th
Hang Chang	85.0%	3rd
Jocelyn Barker	90.0%	2nd
Our method	**97.5%**	1st

Table 2. Comparison with other methods for classification

MCIL	Image-level SVM	SVM-MF	SVM-CNN
91.1%	62.2%	77.8%	**97.8%**

Segmentation Our final submission to the challenge achieved first place with an accuracy of 84% on the test data. Our top performing model surprisingly improved by 11 points compared with the second team. Table 3 shows the top performances from the other participating teams. In addition, we made a comparison of the experiments in segmentation including GraphRLM [17], SVM with manual features and our submitted method. Figure 6 shows some results from leave-one-out cross-validation only using 35 training data. The GraphRLM method is an unsupervised method to segment histopathology tissue images. It is not suitable for segmenting necrosis and non-necrosis regions. The performance of manual features showed 64% accuracy while the performance of CNN features showed 84% accuracy. A gain of CNN features over manual features was 31%.

Table 3. Segmentation performance in the challenge

	Accuracy	Place
Anne Martel	63%	4th
Hang Chang	68%	3rd
Siyamalan Manivannan	73%	2nd
Our method	**84%**	1st

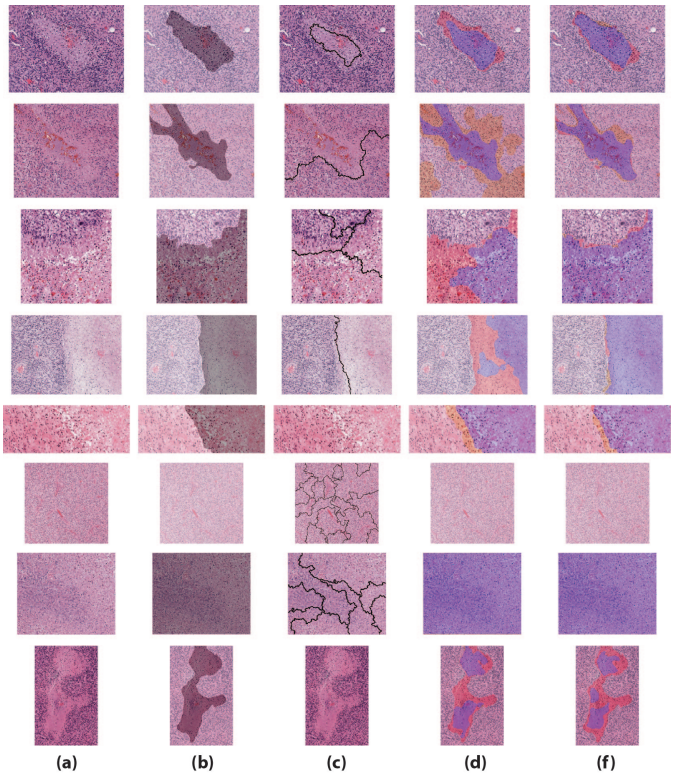


Fig. 6. Image Types: (a): The original images. (b): The gray mask represents necrosis. (c),(d),(e): GraphRLM, SVM-MF, SVM-CNN. Purple: true segmentation; Pale red: missed segmentation; Orange: false segmentation.

6. CONCLUSION

In this paper, we have introduced deep convolutional activation features trained by ImageNet knowledge in the MICCAI 2014 Brain Tumor Digital Pathology Challenge. We successfully transferred ImageNet knowledge as deep convolutional activation features to histopathology image classification and histopathology image segmentation with relatively little training data. CNN features are significantly more powerful than manual features (an improvement of 20 points in both classification and segmentation). In addition, due to the large size of histopathology images, feature pooling is used for a single feature vector in our classification method. Experimentations have demonstrated that this efficient method can achieve a-state-of-the-art accuracy of 97.5% for classification and 84% for segmentation in the brain tumor challenge. In the future, we will attempt this method on other image tasks in the medical image field.

7. REFERENCES

[1] Y. Xu, T. Mo, Q. W. Feng, P. L. Zhong, M. D. Lai, and E. I. Chang, "Deep learning of feature representation with multiple instance learning for medical image analysis," in *ICASSP*, 2014.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.

[3] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *ECCV*, 2014.

[4] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580*, 2012.

[5] O. Russakovsky, J. Deng, J. Krause, A. Berg, and F. Li, "ILSVRC-2013," 2013.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[7] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *CVPR*, 2014.

[8] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," *arXiv preprint arXiv:1403.6382*, 2014.

[9] Brain Tumor Digital Pathology Challenge, "http://pais.bmi.stonybrookmedicine.edu/," .

[10] Glioma, "http://en.wikipedia.org/wiki/glioma/," .

[11] Glioblastoma Multiforme, "http://en.wikipedia.org/wiki/glioblastoma multiforme/," .

[12] Y. L. Boureau, J. Ponce, and Y. LeCun, "A theoretical analysis of feature pooling in visual recognition," in *ICML*, 2010.

[13] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid, "Good practice in large-scale learning for image classification," in *CVPR*, 2012.

[14] H. Chang, A. Borowsky, P. Spellman, and B. Parvin, "Classification of tumor histology via morphometric context," in *CVPR*, 2013.

[15] P. W. Huang and C. H. Lee, "Automatic classification for pathological prostate images based on fractal analysis," *TMI*, vol. 28, no. 7, pp. 1037–1050, 2009.

[16] O. Sertel, J. Kong, H. Shimada, U. V. Catalyurek, J. H. Saltz, and M. N. Gurcan, "Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development," *Pattern Recognition*, vol. 42, no. 6, pp. 1093–1103, 2009.

[17] A. B. Tosun and C. Gunduz-Demir, "Graph run-length matrices for histopathological image segmentation," *TMI*, vol. 30, no. 3, pp. 721–732, 2011.

[18] Q. V. Le, J. Han, J. W. Gray, P. T. Spellman, A. Borowsky, and B. Parvin, "Learning invariant features of tumor signatures," in *ISBI*, 2012.

[19] N. Nayak, H. Chang, A. Borowsky, P. Spellman, and B. Parvin, "Classification of tumor histopathology via sparse feature learning," in *ISBI*, 2013.

[20] 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 *ICML*, 2014.

[21] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *ICLR*, 2014.

[22] Y. Xu, J. Y. Zhu, E. I. Chang, M. D. Lai, and Z. W. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Medical image analysis*, vol. 18, no. 3, pp. 591–604, 2014.

[23] The Cancer Genome Atlas, "http://cancergenome.nih.gov/," .