Classification of MRI Migraine Medical Data using 3D Convolutional Neural Network*

Hwei Geok Ng[1], Matthias Kerzel[1], Jan Mehnert[2],
Arne May[2], and Stefan Wermter[1]

[1] Universität Hamburg, Department of Informatics, Knowledge Technology,
Vogt-Kölln-Str. 30, 22527 Hamburg, Germany
{5ng, kerzel, wermter}@informatik.uni-hamburg.de
[2] Institut für Systemische Neurowissenschaften,
Universitätsklinikum Hamburg-Eppendorf,
Martinistraße 52, 20246 Hamburg, Germany
{j.mehnert, a.may}@uke.de

Abstract. While statistical approaches are being implemented in medical data analyses because of their high accuracy and efficiency, the use of deep learning computations can potentially provide out-of-the-box insights, especially when statistical approaches did not yield a good result. In this paper we classify migraine and non-migraine magnetic resonance imaging (MRI) data, using a deep learning method named convolutional neural network (CNN). 198 MRI scans, which were obtained equally from both data groups, resulted in the maximum classification test accuracy of 85% (validation accuracy: $\bar{x}$=0.69, σ=0.06), compared to the baseline statistical accuracy of 50%. We then used class activation mapping (CAM) method to visualize brain regions that the CNN model took to distinguish one data group from the other and the visualization pointed at the parietal lobe, corpus callosum, brain stem and anterior cingulate cortex, of which the brain stem was mentioned in the medical findings for white matter abnormalities. Our findings suggest that CNN and CAM combined can be a useful image-based data analysis tool to add inspiration or discussion in the medical problem-solving process.

Keywords: Convolutional Neural Network · Class Activation Mapping · Migraine · Magnetic Resonance Imaging.

1 Introduction

Statistical approaches are used in medical data analyses because they are efficient to be implemented and return precise results. Nevertheless, given sufficient meaningful data and computational power, deep learning approaches can also assist in the data analytics process. Convolutional neural networks (CNNs) are a useful deep learning approach, known for their high accuracy in learning relevant features for arbitrary classification tasks, especially for image classification. CNNs

* The authors[1] gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169).

have been utilized in solving numerous medical data problems, such as multiple sclerosis lesion detection [9], Alzheimer's disease recognition [4] and neuronal structure segmentation [3]. Given a balanced dataset, deep learning approaches provide insights that are unbiased to the medical domain knowledge and potentially suggest out-of-the-box findings. Besides, in situations where conventional statistical analyses did not yield good results, deep learning approaches can be a helpful alternative in getting suggestions and inspiration in the problem-solving process.

A trained CNN classification model can be further combined with the class activation mapping (CAM) approach to visualize discriminative regions, which contributed to the classification of the given data. CAM utilizes the learned spatial information of the CNN model and displays discriminative regions of a given image with respect to a chosen class label. The resulted map shows the locations of discriminative features of the image, which the model used to make the classification decision. For an example, a small part of an image showing a toothbrush can be identified as having the strongest contribution to the image being classified as 'brushing teeth' [10]. Applying CNN for classification means that we get to know 'what' is in the image and applying CAM for feature localisation means that we get to know 'where' in the image are the relevant parts that contributed to the classification.

CNN and CAM combined as a medical data analysis tool can be applied to any image-based data. In this paper, we evaluated the classification performance of a CNN specifically on migraine magnetic resonance imaging (MRI) data and used CAM to point out respective discriminative regions. Migraine is a common headache disorder that originates in the trigeminal nervous system which influences 12 - 14% of the world's human population [6]. Nevertheless there is no clear evidence which cortical structures are causing the disorder. MRI image analysis of migraineurs might give a hint of which structures are involved in the development of a migraine as well as providing insights into long-term structural changes caused by migraine.

198 white matter MRIs were obtained equally from migraine and non-migraine participants and preprocessed by the authors from the medical domain. The data was then analyzed using CNN and CAM by the authors from the computer science domain. The CNN classification result was evaluated by executing the best-performing model ten times with random data shuffling. The frequently-occurred and sample-based CAMs were reported. We aim to explore whether an outcome that is free from medical knowledge bias could bring insight to the current medical research as well as to foster scientific exchange and collaborations between medical and computer science domains.

Section 2 explains the experimental setup and methodologies used. Section 3 reports the classification and feature localisation results of the best model. Section 4 discusses the experiment outcome and concludes the study with suggestions for future work.

2 Experimental Setup and Neural Network Architecture

The experimental setup was divided into three stages: 1) dataset acquisition and data preprocessing, 2) CNN training, optimising and testing, and 3) CAM visualization of discriminative regions. The raw MRI images were preprocessed by isolating only white matter regions and discarding all other parts of the images. A three-dimensional CNN architecture was implemented and hyperparameters were modified to get the most optimised validation and test accuracies. The best CNN model was executed ten times with random data shuffling and its accuracy was evaluated. The weights from the best trial were further used for activation maps generation. The regions of activation maps were reported and discussed.

2.1 Dataset Acquisition and Data Preprocessing

The MRI dataset is provided by the authors from the Headache and Pain Research Group at the University Medical Center Hamburg-Eppendorf (UKE). All migraineurs were categorized by a team of trained physicians at the Headache Ambulance of the UKE, while healthy controls reported neither psychiatric nor neurological disorder and no headache disorder in first degree relatives. Raw structural (MPRAGE) images were preprocessed using the Computational Anatomy Toolbox (CAT12[3]) for SPM12 which was implemented in MATLAB. Hereby, each image was segmented into its compartments (grey matter, white matter and cerebrospinal fluid) and normalized to a standardized template space (Montreal Neurological Institute space), as shown in Fig. 1. The images were modulated to keep the volumetric information during this non-linear transformation. The chosen 'mwc2' dataset consisted of 99 white matter MRIs respectively from different migraine patients and the same number from non-migraine people, making a total of 198 images in NIfTI[4] format. Each sample was warped to the same dimensions of (x:121, y:145, z:121).

The preprocessed dataset was handed over to the authors from the Knowledge Technology Group, University of Hamburg, for CNN and CAM implementation. An initial visual inspection showed that every sample looked different from each other, yet there was no obvious feature to distinguish the dataset into the migraine and non-migraine categories, as shown in Fig. 2. A preliminary statistical t-processing done by the medical authors has found no discriminant feature from the dataset, therefore we assumed a baseline accuracy of 50% from the t-test. That means any result from this study that yielded a higher-than-random baseline accuracy can be seen as an improvement to the statistical approach.

The Nibabel[5] Python library was used to load the NIfTI dataset as multidimensional arrays. The arrays from both the migraine and non-migraine classes were assigned to their respective labels. The arrays were then shuffled within their own classes, before being assigned to the train, validation and test sets.

[3] http://www.neuro.uni-jena.de/cat/
[4] https://nifti.nimh.nih.gov/
[5] http://nipy.org/nibabel/

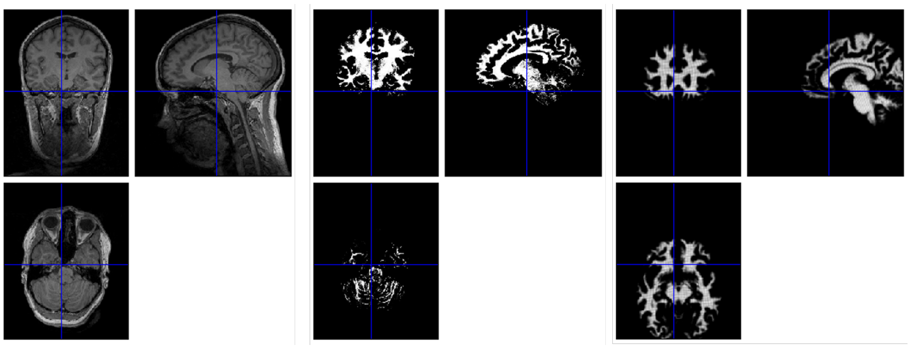

Fig. 1. An example of MRI data: (left) original image sample, (middle) white matter segment and (right) modulated and normalized white matter compartment.

An approximation of 80%, 10% and 10% data proportion were assigned: 158 images to the train set, 20 images to the validation set and 20 images to the test set, with equal data proportion from both the classes to achieve an unbiased outcome. The arrays were then shuffled again, separately within each set.

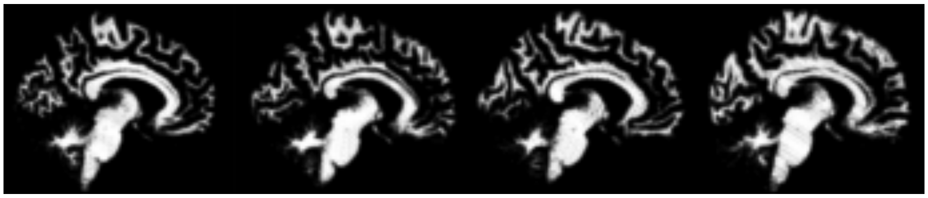

Fig. 2. Four different MRI samples sliced at X: 66/121, Y: 73/145, Z:59/121. Two MRIs from the left are of non-migraineurs and two MRIs from the right are of migraineurs. From visual inspection, all images have different structures but there is no obvious feature to categorise them into migraine and non-migraine groups.

2.2 Network Architecture

A CNN was implemented in Keras[6] with Tensorflow[7] as backend, trained with two 8GB Nvidia GeForce GTX 1080 GPUs. Fig. 3 shows the final network architecture and hyperparameter configuration for the CNN after extensive testing and principled grid search for hyperparameters.

The search range for the best hyperparameters started by taking the minimum values that made the network converge, up until the maximum values that

[6] https://keras.io/

[7] https://www.tensorflow.org/install/

could be allocated by the GPU memory. A batch size of two was assigned to cope with the limited GPU memory. A larger batch size with a smaller network did not yield good accuracies from empirical analyses. The CNN training phase was set to 50 epochs with categorical cross-entropy as the model loss function, Adam [2] with 0.0001 learning rate as the optimizing function and categorical accuracy as the accuracy measure. The input shape for the first convolutional layer was (2, 121, 145, 121). All convolutional layers have a filter size of (3,3,3) with zero padding, rectified linear units as the activation function and glorot uniform as the kernel initializer. The small filter size was used to retain many low-level features from the input. The number of filters differs in each convolutional layer: 20, 10 and 5 in ascending layer order. Max-pooling layers were only applied to the second and third convolutional layer. The reason not to pool the first layer was to retain as much information as possible from the input to the first feature maps. Each max-pooling layer has the same pool size and pool stride of (2,2,2) to decrease the size of network parameters. The dense layers have rectified linear units as the activation function, with the first and the second dense layers having 300 and 100 units respectively. The output layer has two units indicating a two-classes classification and softmax was used as the non-linearity function.

2.3 Discriminative Regions Visualization

The CNN model that achieved the best validation and testing accuracies was further examined with CAM. The steps to generate discriminative activation maps from the original approach by Zhou et al. [10] are: 1) pass an input image to a CNN, which has no dense layers, 2) the feature maps of the final convolutional layer are global-average-pooled (GAP) and influence the output layer prediction, 3) get a classification prediction, 4) the weights between the GAP and the output layer are multiplied with their respective feature maps to identify important regions, 5) sum all the feature maps into one class activation map, and 6) transform the map into heatmaps. The reason to replace the dense layers with a GAP was that the learned spatial information in a CNN will be lost in the dense layers.

A slight modification was done to the CAM approach that we apply in this paper to preserve the dense layers in our network architecture: feature maps from the final convolutional layer were obtained, summed, normalized and transformed into heatmaps, as shown in Fig. 3. This implied that we did not use the weights between the GAP and the output layer. The reason why the modification did not affect the discriminative region accuracy was because of the binary-class CNN classification. Without relying on the weights, the discriminative regions shown from activation maps were class-invariant. It might be problematic for a multi-classes type classification as each class has different features to prioritize. However, for a binary migraine/non-migraine classification, the discriminative regions showed in any activation map separate one class from another: the same region distinguishes migraine from non-migraine class and vice versa.

The final convolutional layer was chosen for CAM visualisation as its activations contain the most detailed features compared to the earlier convolutional

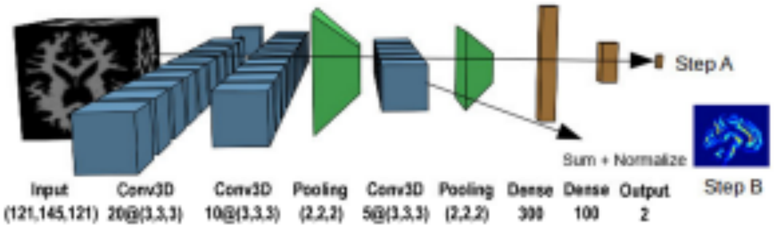

Fig. 3. The three-dimensional CNN architecture and hyperparameter configuration: three convolutional layers, two max-pooling layers, two dense layers and one output layer. Step A: the normal network architecture with dense layers producing a classification prediction. Step B: a modified version of class activation maps. At the final convolutional layer, the feature maps were obtained, summed and normalized to form a class-invariant activation map out of a binary classification.

layers. For each test sample and each activation unit of a convolutional layer, a three-dimensional activation map was generated by summing and normalizing the activations across activation units. Each map was sliced at the x-axis (sagittal brain section) to obtain 60 slices of two-dimensional activation maps. Each slice was resized to five times its current dimensions for better visual inspection. Overall, the total number of generated activation maps was 20 test samples x 60 slices = 1,200 activation maps. The maps visualized how much a given voxel contributed to the classification result: the blue regions indicate a low contribution while the red regions indicate a high contribution. For an example, Fig. 5 shows discriminative features marked in red colour.

3 Experimental Results

The experimental setup in Section 2.2 was evaluated for its CNN classification performance and the best performing model was used to display discriminative regions using CAM.

3.1 Classification Result of CNN

The CNN configuration was validated ten times with random data shuffling and the result is shown in Table 1. The highest test accuracy was 85% (validation accuracy: $\bar{x}$=0.69, σ=0.06) and the lowest test accuracy was 40% (validation accuracy: $\bar{x}$=0.60, σ=0.07). In contrast with established CNN architectures from vision processing, which usually have an ascending number of convolutional filters [7], the number of convolutional filters of this study was descending over the layers (20-10-5 convolutional units). Through extensive empirical testing, we found the model that has an ascending number of filters performed better than the model that has a descending number of filters (5-10-20 convolutional units). The latter yielded only 35% test accuracy. Understanding that the first convolutional layer needed some minimum amount of units in order to extract useful low-level features for good classification accuracy, the 20-10-5 architecture was used to optimally utilize the limited GPU memory. Three layers achieved the best result compared to other numbers of convolutional layers.

Fig. 4 shows the losses and accuracies from the best model, Run 7, which achieved 85% test accuracy (validation accuracy: $\bar{x}$=0.69, σ=0.06). From the twelfth epoch onwards, the training loss decreased to 0.0 while the validation loss slowly increased to 0.75 over time. From the ninth epoch onwards, the training accuracy increased to 1.0 while the validation accuracy slowly increased to 0.75 over time. The losses indicated overfitting of the model as the training loss decreased while the validation loss increased over time. There were further attempts to alter the network architecture to decrease overfitting. Nevertheless, this was the best result with the available data size. Although the validation loss increased over time, the increased validation accuracy indicated that the network did learn relevant features for the classification.

Table 1. Ten evaluations of the best configuration with random data shuffling. From left: trials, test loss, test accuracy, mean ($\bar{x}$) and standard deviation (σ) of validation loss, mean ($\bar{x}$) and standard deviation (σ) of validation accuracy. Run 7 achieved the best mean validation accuracy of 69% (σ=0.06), which led itself to have the best test accuracy of 85%.

Run	Test Loss	Test Accuracy	Val Loss ($\bar{x}$,σ)	Val Accuracy ($\bar{x}$,σ)
1	1.06	0.55	1.61 , 0.42	0.36 , 0.06
2	0.87	0.70	1.16 , 0.19	0.49 , 0.05
3	0.91	0.65	1.14 , 0.21	0.38 , 0.05
4	1.31	0.55	0.98 , 0.13	0.47 , 0.04
5	1.89	0.45	1.36 , 0.34	0.46 , 0.04
6	1.79	0.40	0.85 , 0.08	0.60 , 0.05
7	**0.68**	**0.85**	**0.68 , 0.04**	**0.69 , 0.06**
8	0.77	0.60	0.92 , 0.13	0.59 , 0.04
9	2.00	0.55	0.91 , 0.10	0.53 , 0.04
10	1.66	0.40	1.15 , 0.20	0.60 , 0.07

3.2 Visualisation Result of CAM

The best model from the CNN classification (Run 7) was further analysed with CAM to visualize relevant areas that contributed to the classification. The model with the highest test accuracy was used because the features it has learned were the most accurate to separate the data into two classes. 1,200 two-dimensional activation maps were generated from 20 test samples and each map was visually inspected and analyzed. The most common regions that appeared in all test samples are the parietal lobe and the corpus callosum, as shown in Fig. 5 (top). Although the extent of distinction (red areas) was different for each test sample, these three regions were highly discriminative in every test sample, indicating

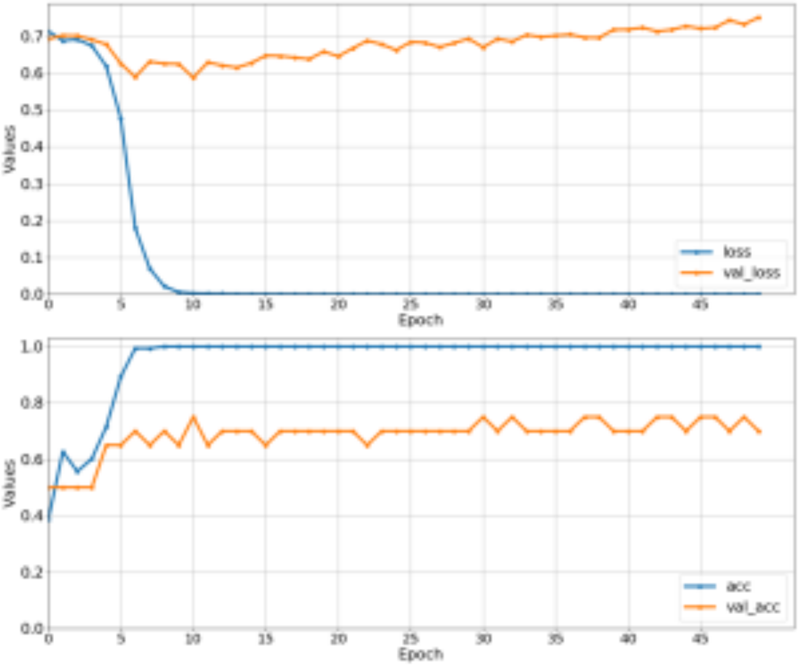

Fig. 4. (Top) The training and validation losses of Run 7: the training loss decreased while the validation loss increased over time, indicating an expected network overfitting because of the small dataset and large data dimensions. (Bottom) The training and validation accuracies of Run 7: both the training and validation accuracies increased over time. Although the validation accuracy was not as good as the training accuracy, the increased accuracies indicated that useful features were learned for the classification.

that the model regarded these areas as important in classifying migraine and non-migraine MRIs. The discriminative regions that appeared specifically in certain test samples were also visualised at the bottom of the same figure, which highlight the brain stem, the corpus callosum and the anterior cingulate cortex.

There are many medical research methods that aim to identify the differences of migraine and non-migraine brains, such as using functional, grey matter and white matter MRI data. These studies report different brain regions which were distinct in showing the differences between migraine and non-migraine brains, such as activation in the brainstem [8], decrease of grey matter in the cingulate cortex [5] and white matter abnormalities in the brain stem and other areas [1]. As we used white matter MRI data for CNN and CAM computations, the CAMs which pointed as discriminative areas at the brainstem showed that these regions were also regarded as important for CNN classification and the other regions suggested by CAMs might be worth further exploration for migraine study.

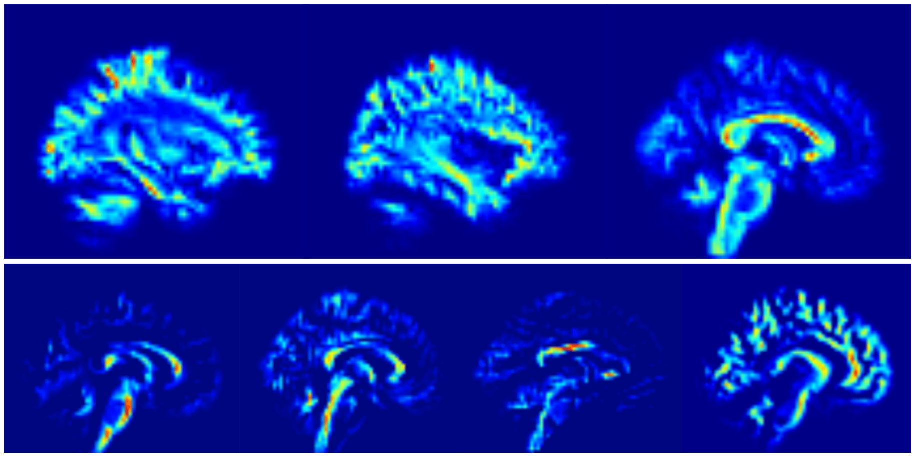

Fig. 5. (Top) Three highly discriminative regions (marked in red) appeared in every test data. From left: left parietal lobe, right parietal lobe and corpus callosum. (Bottom) Four CAMs detected from certain test samples. From left: brain stem (area 1), brain stem (area 2), corpus callosum and anterior cingulate cortex.

4 Discussion, Conclusion and Future Work

This paper described the classification of migraine MRI data using a CNN and the discriminative areas visualization using CAM. The challenge for the chosen dataset was that the preliminary statistical t-processing returned no discriminative feature. Compared to thousands or millions of images used for general CNN visual recognition tasks, we were dealing with a small dataset (198 samples) with high-dimensional data (x:121, y:145, z:121), which made the neural network prone to overfitting. Slicing dimensions and adding noise variants to increase the data size did not seem appropriate since without professional medical knowledge we might introduce errors. Nevertheless, one main contribution of the paper is the classification performance, which yielded 85% maximum test accuracy (validation accuracy: $\bar{x}$=0.69, σ=0.06), higher than the 50% baseline statistical result and the approach suggested areas that the deep learning model made a distinction for data classification. Some areas such as the brain stem was mentioned in the medical literature fro white matter abnormalities [1], while some areas such as the parietal lobe and the corpus callosum are not mentioned.

Migraine MRI data is challenging to be analysed because the differences between migraine and non-migraine classes are subtle, unlike other MRI data such as Alzheimer's disease, which displays distinguishable structural change between MRIs of patients and the control group [4]. From the medical perspective, the classification and localisation accuracies are yet to be improved, nevertheless, this is a good milestone with the limited number of samples available. Our intention is to provide a useful pipeline to assist in medical discussions as a recommender system from the point of view of a computation system - which feature the CNN used to make the classification decision. For future work, we look forward to improving results even further when more data and more powerful GPUs become available.

There are many challenges in the medical domain to which neural network approaches can potentially contribute. It is important for researchers to gain interdisciplinary knowledge and to design efficient data acquisition processes that help in the performance of neural networks. Fostering collaboration between experts from both computer science and medical domains, more medical-related problems could potentially be solved by combining expertise from both sides.

References

1. Bashir, A., Lipton, R.B., Ashina, S., Ashina, M.: Migraine and structural changes in the brain: a systematic review and meta-analysis. Neurology **81**(14), 1260–1268 (2013)
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
3. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
4. Sarraf, S., Tofighi, G.: Deep learning-based pipeline to recognize alzheimer's disease using fmri data. In: Future Technologies Conference (FTC). pp. 816–820. IEEE (2016)
5. Schmidt-Wilcke, T., Gänßbauer, S., Neuner, T., Bogdahn, U., May, A.: Subtle grey matter changes between migraine patients and healthy controls. Cephalalgia **28**(1), 1–4 (Nov 2007). https://doi.org/10.1111/j.1468-2982.2007.01428.x, https://doi.org/10.1111/j.1468-2982.2007.01428.x
6. Schulte, L.H., May, A.: The migraine generator revisited: continuous scanning of the migraine cycle over 30 days and three spontaneous attacks. Brain **139**(7), 1987–1993 (2016). https://doi.org/10.1093/brain/aww097, http://dx.doi.org/10.1093/brain/aww097
7. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
8. Sprenger, T., May, A.: Advanced neuroimaging for the study of migraine pathophysiology. Pain: Clinical Updates **20**(6) (Oct 2012), https://www.iasp-pain.org/files/Content/ContentFolders/Publications2/PainClinicalUpdates/Archives/PCU_20-6_web.pdf, [Accessed July 14, 2018]
9. Valverde, S., Cabezas, M., Roura, E., González-Villà, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Oliver, A., Lladó, X.: Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. NeuroImage **155**, 159–168 (2017)
10. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2921–2929 (2016)