Tony Blair Institute for Global Change
Trinity University, Lagos, Nigeria
Harmarth Global Educational Services
FAIR Forward – Artificial Intelligence for All
Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH
Society for Multidisciplinary & Advanced Research Techniques (SMART)

Accra Bespoke Multidisciplinary Innovations Conference (ABMIC)

Towards An Investigation on the use of Deep Learning In Image Captioning for Diagnostic Imaging

[1]Pius Kwao Gadosey & [2]Akunlibe, G.A.A
[1]Department of Computer Science, Lancaster University, Accra, Ghana
[2]Faculty of Computational Sciences & Informatics, Academic City University College, Accra, Ghana
E-mail: kwaogad@gmail.com; getrude.akunlibe@acity.edu.gh
Phone: +233595228494; +233508890169

ABSTRACT

This research sets out to exploredthe application of Deep learning models in image captioning.. Our intention is to investigated how to accurately automatically generate captions as compared to other previous studies the efficiency of the work. Deep learning uses algorithms and complex data sets in enormous datasets. We provide a brief overview of some of the most significant deep learning schemes used in computer vision problems, that is, Convolutional Neural Networks and others. Our future work will explore Chest X-rays and COVID-19 radiography datasets from Indiana State University and Qatar University respectively to implementr the research objectives

Keywords: Imaging, Deep Learning, Neural Networks, Captioning, Diagnostics, ML, Data Sets

Proceedings Citation Format
Pius Kwao Gadosey & Akunlibe, G.A.A (2022): Towards An Investigation on the use of Deep Learning In Image Captioning for Diagnostic Imaging. Proceedings of the 31st Accra Bespoke Multidisciplinary Innovations Conference. University of Ghana/Academic City University College, Ghana. 1st – 3rd June, 2022. Pp 205-214, www.isteams.net/ghanabespoke2022. dx.doi.org/10.22624/AIMS/ABMIC2022P20

1. Background of study

Machine learning has recently seen some significant advancements, piquing the interest of industry, academia, and popular culture. Artificial neural networks, which are used in deep learning, are a set of techniques and algorithms that enable computers to learn. In order to find complex patterns in enormous data sets. Increased funding is fueling the advances of Big data access, user-friendly software frameworks, and an explosion of available data computational capacity, allowing for the implementation of deeper neural networks than ever before. These architectures used to build AI Models are now the state-of-the-art approach to a wide range of computer problems such as Robotics, vision, and language modeling.

When neural networks began outperforming other methods on several high-profile image analysis benchmarks, deep learning rose to prominence in computer vision. The most well-known deep learning model (a convolutional neural network) halved the second-best error rate on the image classification task on the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)1 in 2012 (Brody, 2013).

Until recently, enabling computers to recognize things in natural photographs was thought to be a challenging problem, but convolutional neural networks have now outperformed even human performance on the ILSVRC, and have basically solved the ILSVRC classification test (i.e. with error rate close to the Bayes rate). The most recent ILSVRC competition took place in 2017, and computer vision research has moved on to increasingly demanding benchmark tasks since then. Consider the COCO (Common Objects in Context Challenge) (Shao et al., 2014).

1.1 X-rays and Ultrasound Scns

Globally, basic x-ray and ultrasound examination can resolve 70 to 80 percent of the world's diagnostic problems (Mitchell and https://www.facebook.com/pahowho, 2012). Diagnostic imaging is simply a method of diagnosing a patient by capturing and analyzing their visual content in the form of images. This method allows doctors to guess clues about a medical condition from inside your body. There are many different types of diagnostic imaging, but the most common ones used on a daily basis are X-rays, CT scans, and MRI scans. Text is usually included with images to avoid misinterpretation. Image captioning is the textual description of an image (Radhakrishnan, 2017). A single image can represent a plethora of messages. As a result, image captioning is a critical concept that is equally important in the diagnostic imaging aspect of the medical field. In recent years, hospitals have developed well-known areas of specialization. Some of these hospitals have more than one area of specialization, while others only have one, but they still provide treatment or medical services in areas other than their known area of specialization.

Regardless of the different approaches to treating various illnesses, all hospitals have a set standard that patients must follow in order to be diagnosed and treated. Among the many departments at the hospital are the Out-Patient Department, the Consulting Unit, the Laboratory, the Theater, and the Dispensary (Yonzon, n.d.). Depending on the patient's insurance or payment method, as well as their illness, they may have to visit at least three departments before receiving treatment. In addition, depending on the patient's problems, they may be required to undergo laboratory tests or scans to assist the doctor in making a proper diagnosis.

Each time a patient requires a diagnosis, it is almost certain that they will undergo a laboratory test or a scan. It has been observed that scans are normally analyzed by doctors and patients are treated. The results of those scans could be structured in such a way that if a similar case occurs in a different patient who has previously been treated, they can treat it faster for the new patient.

2. RELATED LITERATURE

2.1 Deep learning Methods

Deep learning methods are extremely effective when there are a large number of available samples during the training stage. For example, more than one million annotated images were available in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015). In most medical applications, however, there are far fewer images (i.e., 1,000). As a result, one of the most significant challenges in applying deep learning to medical images is the limited number of training samples available to build deep models without overfitting. To address this issue, research groups have devised a variety of strategies, including (a) using image patches rather than full-sized images as input (Cheng et al., 2016) to reduce input dimensionality and thus the number of model parameters; (b) expanding the data set by artificially generating samples via affine transformation (i.e., data augmentation), and then training their network from scratch with data augmentation. (c) using "off-the-shelf" feature extractors trained on a large number of natural images in computer vision, and then training the final classifier or output layer with the target-task samples (Ciompi et al., 2015); (d) initializing model parameters with those of pretrained models from nonmedical or natural images, and then fine-tuning the network parameters with task-related samples (Gupta et al., n.d.); and (e) using models trained with small-sized inputs for arbitrarily sized inference (Gupta et al., n.d.).

2.2 Related Works

Keras Image Captioning by Harshall Lamba: He used flicker 8k photos as the dataset in this experiment. He had 5 captions for each photograph, which he saved in a dictionary. He cleaned the data by making all words lowercase, removing special tokens, and removing terms with numbers (such as 'hey199', etc.). He excluded any words with a frequency of less than 10 in the entire corpus after extracting unique words so that the model might be robust to outliers. Each caption now has "startseq" and "endseq" added to it. He used the inception v3 model with Image net dataset weights to preprocess the photos. This model was fed images, and the output of the second last layer (2048 size vector).

Attention Mechanism Image Captioning | by Subham Sarkar | The Business: For image captioning, he exploited the attention mechanism. As before, the dataset used was flickr8k. The train size was 6000 photos, the validation data was 1000 images, and the test data was 1000 images. He deleted punctuation, numeric values, and single characters before preprocessing. After that, he makes a dataframe with the columns filename and captions. Before feeding each image to the ImageNet VGG-16 model, he reshaped it into 224*224*3. Only the convolutional component is included here (include top=False). He took the output from the second last layer as backbone features and eliminated the last dense layer.

Using deep convolutional networks, (Roth et al., 2015) proposed a method for organ- or body-part-specific anatomical classification of medical pictures. They used 4,298 axial 2D CT scans to train their deep network to learn five sections of the body: the neck, lungs, liver, pelvis, and legs. Their research yielded a 5.9% anatomy-specific classification error and a 0.998 average AUC (area under the receiver-operating characteristic curve). However, in real-world applications, finer separation than that used for merely five body components may be required (e.g., they may need to distinguish aortic arch from cardiac sections).

Yan et al., (2015) developed a multistate deep learning framework using a CNN to identify the body portion of a transversal slice to overcome this restriction. Because each slice may contain many organs (contained in bounding boxes), the CNN was trained in a multi-instance method (Maron and Lozano-Pérez, 1998), with the objective function changed so that the associated slice was considered right as long as one organ was properly classified. As a result, the CNN was sensitive to the discriminative boundary boxes before it was trained. To improve the representation power of the pretrained CNN, discriminative and noninformative bounding boxes were chosen based on the answers of the pretrained CNN. To apply the boosted CNN to the subject image at run time, a sliding-window technique was used. Because the CNN only had peaky responses on discriminative bounding boxes, it was able to identify body parts by concentrating on the most unique local information. This local technique proved more accurate and robust than global image context-based alternatives. The authors' body part recognition approach was evaluated on 12 body parts across 7,489 CT slices from 675 patients ranging in age from 1 to 90 years old. The data was separated into three categories: 2,413 people (225 patients) were enrolled in the training program. 656 (56 patients) for validation, and 4,043 (394 patients) for testing.

Cireşan et al., (2013) employed a deep CNN to detect mitosis in breast cancer histology images in a groundbreaking study. Their networks were trained to categorize each pixel in the photos using a patch centered on the pixel as a starting point. Their technique took first place in the Mitosis Detection Contest at the 2012 International Conference on Pattern Recognition (ICPR), 4 beating the other competitors by a large margin. Since then, several groups have applied various deep learning approaches for histology image detection. For example, (Xu et al., 2016) employed an SAE to detect cells on histological pictures of breast cancer. To boost robustness to outliers and disturbances, they used a denoising auto-encoder to train their deep model. Su et al. (Su et al., 2015) detected and segmented cells from microscopic pictures using an SAE and sparse representation. To detect and classify nuclei in histopathology images, Sirinukunwattana et al. (100) proposed a spatially constrained CNN (SC-CNN). They employed a SC-CNN to estimate the likelihood of a pixel being the nucleus's center, with pixels with high probability values spatially limited to be in the vicinity of nuclei's centers.

Moeskops et al. (NCBI, 2019) developed a multiscale CNN to improve the robustness and spatial consistency of newborn picture segmentation. To obtain multiscale information about each voxel, their network used different patch sizes and numerous convolution kernel sizes. The authors achieved encouraging segmentation results for eight tissue types using this strategy, with a Dice ratio6 ranging from 0.82 to 0.91 across five separate data sets.

On the basis of multimodal MR images, (Zhang et al., 2015) created four CNN architectures to segment newborn brain tissues. Each CNN has three input feature maps, each corresponding to 1313 voxels of T1-weighted, T2-weighted, and fractional anisotropy (FA) image patches. The scientists applied three convolutional layers and one fully connected layer to each CNN, followed by a tissue classification output layer with a softmax function. These CNNs beat rival approaches on a collection of manually segmented isointense-phase brain pictures.

Pereira et al., (2016) used CNNs in MR images to analyze brain tumor segmentation. They looked into small kernels to see whether they could have fewer parameters but deeper architectures. They trained several CNN architectures for low- and high-grade tumors and verified their method in the 2013 Brain Tumor Segmentation (BRATS) Challenge7, where their method rated first in the full, core, and enhancing regions for the challenge data set. (Brosch et al., 2016) used MR images to segment multiple sclerosis lesions using deep learning.

Brosch et al (2016) used MR images to segment multiple sclerosis lesions using deep learning. Their model was a three-dimensional CNN made up of two interconnected pathways: a convolutional pathway that learned hierarchical feature representations similar to those learned by other CNNs, and a deconvolutional pathway made up of deconvolutional and unpooling layers with shortcut connections to the corresponding convolutional layers. The deconvolutional layers were created to generate abstract segmentation features from the data provided by each convolutional layer, as well as, if relevant, the activation of the previous deconvolutional layer. This technique performed the best in terms of Dice similarity coefficient, absolution volume difference, and lesion false-positive rate when compared to five publicly accessible algorithms for multiple sclerosis lesion segmentation.

To distinguish breast ultrasound lesions from lung CT nodules, Cheng et al. (Nih.gov, 2019) utilized an SAE with a denoising approach (SDAE). The image areas of interest (ROIs) were initially scaled to 2828 pixels, with all pixels in each patch being considered as the SDAE's input. To improve their model's noise tolerance, the authors contaminated the input patches with random noise during the pre training step. They later added the resized scale factors of the two ROI dimensions, as well as the aspect ratios of the original ROIs, during the fine-tuning process to preserve the original information.

To capture different sizes of lung nodules, Shen et al. (Shen et al., 2015) constructed a hierarchical learning architecture using a multiscale CNN. Three CNNs were formed in tandem in this CNN architecture, each taking nodule patches from various scales as input. The authors adjusted the parameters of the three CNNs to be shared during training to reduce overfitting. A feature vector was created by concatenating the activations of the top hidden layer in three CNNs, one for each scale. The authors utilized a random forest and an SVM with a radial basis function kernel for classification. The random forest was trained to minimize partner objectives, which are defined as the sum of the overall hinge loss function and the total of the companion hinge loss functions. (Gönen and Alpaydın, 2011)

Suk et al. (31) employed an SAE to combine neuroimaging and biochemical variables to determine Alzheimer's disease or moderate cognitive impairment. They used MR images to extract GM volume features, PET images to extract regional mean intensity values, and CSF to extract three biological features (A42, p-tau, and t-tau). They created an augmented feature vector for each modality by concatenating the original features with the outputs of the top hidden layer of the relevant SAEs after training modality-specific SAEs. For clinical decision making, a multikernel SVM (Suk, Lee and Shen, 2013) was trained. The same researchers went on to find hierarchical feature representations by mixing diverse modalities during feature representation learning rather than during the classifier learning step (Suk, Lee and Shen, 2014).

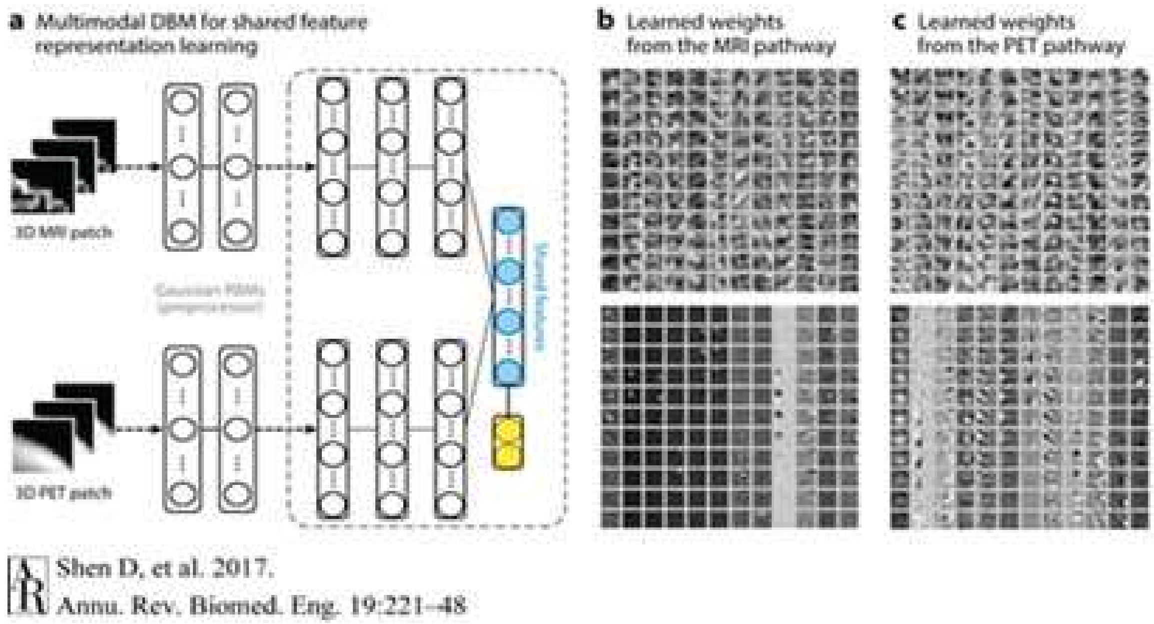

Figure 1 Shared feature learning from patches of different modalities

The shared feature learning from patches of different modalities, such as magnetic resonance imaging (MRI) and positron emission tomography (PET), with a discriminative multimodal deep Boltzmann machine (DBM) are shown in Figure 1. The yellow circles represent the input patches, and the blue circles show joint feature representation. (b,c) Visualization of the learned weights in Gaussian restricted Boltzmann machines (RBMs) (bottom) and those of the first hidden layer (top) from MRI and PET pathways in a multimodal DBM (Suk, Lee and Shen, 2014). Each column, with 11 patches in the upper block and the lower block, composes a three-dimensional patch.

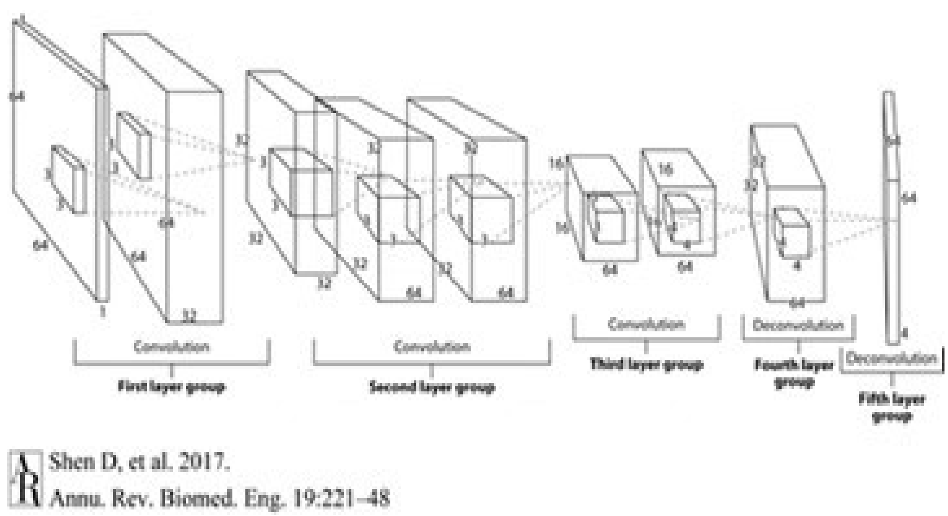

Figure 2 The architecture of the fully convolutional network used for tissue segmentation (Nie et al., 2019)

Nie et al., 2019) proposed using multiple fully convolutional networks (mFCNs) (Figure 8) to segment isointense-phase brain pictures with T1-weighted, T2-weighted, and FA modality information to segment isointense-phase brain images. They used a deep architecture to efficiently fuse high-level information from all three modalities, rather than just merging three-modality data from the original (low-level) feature maps. They thought that high-level representations from various modalities were mutually beneficial. To successfully exploit information from various modalities, the scientists first trained one network for each modality; second, they combined multiple-modality features from each network's high layer (Figure 8). In these tests, the mFCNs outperformed fully convolutional networks and other competing approaches, achieving average Dice ratios of 0.852 for CSF, 0.873 for GM, and 0.887 for WM from eight patients.

3. RESEARCH THRUST

3.1 Problem Statement

Billions of diagnostic images are generated every day, and all of these images must be analyzed by doctors or radiologists after they are generated. Unfortunately, a hospital is a space that must operate twenty-four hours a day; however, the humans who run the department, such as doctors and radiologists, are incapable of operating twenty-four hours a day productively or effectively. Any kind of break or gap in a department can have a significant impact on overall labor efficiency and human lives. Diagnostic imaging services, for example, have a significant impact on public health and can eventually reduce infant mortality rates or increase rate of detection of cancer and tumors.

3.2 Field and subject Area of Study

The field of study is in Computer Science and the subject area of study is under application of deep learning in image capturing for medical diagnosis. This concept is a subset of machine learning and artificial intelligence.

3.3 Aim- General Objective of Study

The overall aim of this study is to increase the accuracy of automatic image captioner for diagnostic image and subsequently confirm the best architecture for implementing automatic captioner that describes images into sentences by employing computer vision and natural language.

3.4 Specific Objectives

1. The model should be able to work with medical images and reports
2. The model should be able tell the best architecture with high accuracy
3. The model should be able to analyze and accurately automatically generate captions for the chest x-ray imaging.

3.5 Significances of Study

1. Increasing the accuracy of the system will save lives
2. The image captioning can speed up the diagnosis process
3. Reports can be used for other further task like

4. CONCLUDING REMARKS AND DIRECTION FOR FUTURE WORKS

In this study, we will be using chest x-ray images and reports from (Chempolil, 2021) and COVID-19 radiography. The idea is to train an AI model using the images as our dataset. This model will take the images as input, analyze and generate captions or reports. The results from these images will be compared with the known reports as a way to determine the accuracy of the model. The level of accuracy will justify deep learning use in image captioning as a way of automation to speed up the medical diagnosis process.

REFERENCES

1. Brody, H. (2013). Medical imaging. Nature, 502(7473), S81–S81. https://doi.org/10.1038/502s81a
2. Brosch, T., Tang, L. Y. W., Yoo, Y., Li, D. K. B., Traboulsee, A., & Tam, R. (2016). Deep 3D Convolutional Encoder Networks With Shortcuts for Multiscale Feature Integration Applied to Multiple Sclerosis Lesion Segmentation. IEEE Transactions on Medical Imaging, 35(5), 1229–1239. https://doi.org/10.1109/tmi.2016.2528821
3. Chempolil, A. T. (2021, February 9). Medical Image Captioning on Chest X-Rays. Medium. https://towardsdatascience.com/medical-image-captioning-on-chest-x-rays-a43561a6871d
4. Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., & Chen, C.-M. (2016). Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Scientific Reports, 6(1). https://doi.org/10.1038/srep24454
5. Ciompi, F., de Hoop, B., van Riel, S. J., Chung, K., Scholten, E. Th., Oudkerk, M., de Jong, P. A., Prokop, M., & Ginneken, B. van. (2015). Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box. Medical Image Analysis, 26(1), 195–202. https://doi.org/10.1016/j.media.2015.08.001
6. Cireşan, D. C., Giusti, A., Gambardella, L. M., & Schmidhuber, J. (2013). Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, 411–418. https://doi.org/10.1007/978-3-642-40763-5_51
7. COVID-19 Radiography Database. (n.d.). Www.kaggle.com. Retrieved May 23, 2022, from https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database?select=COVID-19_Radiography_Dataset
8. Gönen, M., & Alpaydın, E. (2011). Multiple Kernel Learning Algorithms. Journal of Machine Learning Research, 12, 2211–2268. https://jmlr.csail.mit.edu/papers/volume12/gonen11a/gonen11a.pdf
9. Gupta, A., Murat, S., Ayhan, & Maida, A. (n.d.). Natural Image Bases to Represent Neuroimaging Data. http://proceedings.mlr.press/v28/gupta13b.pdf
10. Maron, O., & Lozano-Pérez, T. (1998). A Framework for Multiple-Instance Learning. Neural Information Processing Systems; MIT Press. https://papers.nips.cc/paper/1997/hash/82965d4ed8150294d4330ace00821d77-Abstract.html
11. Martin, M. (n.d.). What is a Functional Requirement in Software Engineering? Specification, Types, Examples. Www.guru99.com. https://www.guru99.com/functional-requirement-specification-example.html#:~:text=A%20Functional%20Requirement%20(FR)%20is

12. National Center for Biotechnology Information. (2019). Nih.gov. http://www.ncbi.nlm.nih.gov/ NCBI. (2019). National Center for Biotechnology Information. Nih.gov. https://www.ncbi.nlm.nih.gov/
13. Nie, D., Wang, L., Adeli, E., Lao, C., Lin, W., & Shen, D. (2019). 3-D Fully Convolutional Networks for Multimodal Isointense Infant Brain Image Segmentation. IEEE Transactions on Cybernetics, 49(3), 1123–1136. https://doi.org/10.1109/tcyb.2018.2797905
14. Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images. IEEE Transactions on Medical Imaging, 35(5), 1240–1251. https://doi.org/10.1109/tmi.2016.2538465
15. Roth, H. R., Lee, C. T., Shin, H.-C., Seff, A., Kim, L., Yao, J., Lu, L., & Summers, R. M. (2015). Anatomy-specific classification of medical images using deep convolutional nets. 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI). https://doi.org/10.1109/isbi.2015.7163826
16. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y
17. Shao, Y., Gao, Y., Guo, Y., Shi, Y., Yang, X., & Shen, D. (2014). Hierarchical lung field segmentation with joint shape and appearance sparse learning. IEEE Transactions on Medical Imaging, 33(9), 1761–1780. https://doi.org/10.1109/TMI.2014.2305691
18. Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. Annual Review of Biomedical Engineering, 19(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442
19. Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015). Multi-scale Convolutional Neural Networks for Lung Nodule Classification. Lecture Notes in Computer Science, 588–599. https://doi.org/10.1007/978-3-319-19992-4_46
20. Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., & Summers, R. M. (2016). Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. ArXiv:1603.08486 [Cs]. https://arxiv.org/abs/1603.08486
21. Su, H., Xing, F., Kong, X., Xie, Y., Zhang, S., & Yang, L. (2015). Robust Cell Detection and Segmentation in Histopathological Images Using Sparse Reconstruction and Stacked Denoising Autoencoders. Medical Image Computing and Computer-Assisted Intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention, 9351, 383–390. https://doi.org/10.1007/978-3-319-24574-4_46
22. Suk, H.-I., Lee, S.-W., & Shen, D. (2013). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Structure and Function, 220(2), 841–859. https://doi.org/10.1007/s00429-013-0687-3
23. Suk, H.-I., Lee, S.-W., & Shen, D. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage, 101, 569–582. https://doi.org/10.1016/j.neuroimage.2014.06.077
24. Team, K. (n.d.). Keras documentation: Keras Applications. Keras.io. https://keras.io/api/applications/
25. Understanding the VGG19 Architecture. (2020, February 26). OpenGenus IQ: Computing Expertise & Legacy. https://iq.opengenus.org/vgg19-architecture/

26. Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., & Madabhushi, A. (2016). Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images. IEEE Transactions on Medical Imaging, 35(1), 119–130. https://doi.org/10.1109/tmi.2015.2458702
27. Yan, Z., Zhan, Y., Peng, Z., Liao, S., Shinagawa, Y., Metaxas, D. N., & Zhou, X. S. (2015). Bodypart Recognition Using Multi-stage Deep Learning. Information Processing in Medical Imaging: Proceedings of the ... Conference, 24, 449–461. https://doi.org/10.1007/978-3-319-19992-4_35
28. Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., & Shen, D. (2015). Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage, 108, 214–224. https://doi.org/10.1016/j.neuroimage.2014.12.061