Bryant University

HONORS THESIS

Feature Detection in Medical Images Using Deep Learning

BY Anthony Pasquarelli

ADVISOR • Brian Blais

Submitted in partial fulfillment of the requirements for graduation
with honors in the Bryant University Honors Program
April 2018

Table of Contents

Feature Detection in Medical Images using Deep Learning
Senior Capstone Project for Anthony Pasquarelli

ABSTRACT

This project explores the use of deep learning to predict age based on pediatric hand X-Rays. Data from the Radiological Society of North America's pediatric bone age challenge were used to train and evaluate a convolutional neural network. The project used InceptionV3, a CNN developed by Google, that was pre-trained on ImageNet, a popular online image dataset. Our fine-tuned version of InceptionV3 yielded an average error of less than 10 months between predicted and actual age. This project shows the effectiveness of deep learning in analyzing medical images and the potential for even greater improvements in the future. In addition to the technological and potential clinical benefits of these methods, this project will serve as a useful pedagogical tool for introducing the challenges and applications of deep learning to the Bryant community.

Feature Detection in Medical Images using Deep Learning
Senior Capstone Project for Anthony Pasquarelli

INTRODUCTION

Deep learning is a new subset of machine learning that attempts to distinguish patterns in sounds, images, and other data by mimicking the layers of neurons that exist in the human brain. Deep learning allows Teslas to recognize pedestrians and allows Spotify to craft customized playlists. In the medical field, Computer Aided Diagnosis and medical image analysis have encountered some of their largest breakthroughs in recent years through deep learning. However, the emergence of deep learning was built up decades of progress in machine learning and artificial intelligence research.

Machine Learning initially emerged in the 1960's as the computer revolution was beginning. As soon as there were computers and large data sets to analyze, scientists were finding new ways to analyze them and predict outcomes. Kononenko (2001) describes the three initial branches of machine learning as (1) classical work in symbolic learning described by Hunt et al. in their book *Experiments in Induction* in 1996. (2) statistical methods laid out in *Learning Machines: Foundations of Trainable Pattern-Classifying Systems* by Nils Nilsson in 1965, and (3) neural networks as shown in *Principle of Neurodynamics. Perceptrons and the Theory of Brain Mechanisms* by Rosenblatt in 1962. According to Kononenko, these were the 3 seminal books that established the field of machine learning, although it was not recognized at the time.

With the continued development as machine learning as a field, applications began to emerge in medical image analysis. One technique that developed was the naive Bayesian Classifier. This was a statistical technique that was initially developed by Good in 1964, and then applied to medical images by Kononenko et al. (1993) where he found that in 5 out of 8 medical applications, Bayesian Classification outperformed all other algorithms. Symbolic learning, which was pioneered by Hunt in 1966, was being applied to medical images from the very beginning by creating decision trees and rules for diagnosis and prognosis. Research in decision trees and rules became increasingly active after Quinlan (1979) created the Iterative Dichotomizer 3 algorithm (ID3). Shortly after, other researchers began applying this algorithm to the medical domain such as Kononenko (1993) where decision trees were used to try to diagnose thyroid diseases, rheumatology, and breast cancer recurrence.

However, during this time there was little progress being made in the domain of neural networks until a seminal 1986 paper by Rumelhart, Hinton, and Williams called *Learning Representations by back—propagating errors*. While the idea of back-propagation had been around on paper since the 60's and implemented on a computer in 1970 (Linnainmaa), it had not been applied to neural networks until 1974 (Werbos) and still did not become widely known until that 1986 paper that so clearly explained the technique and its utility. Back-propagation re-invigorated research in neural networks and very quickly started delivering results, such as the 1989 paper by LeCun et al out of Bell Labs that used back-prop neural networks to read handwritten zip codes.

Back propagation neural networks in the 1990's were the direct predecessors of modern deep learning techniques and more sophisticated networks continued to be developed until there was another large breakthrough in the field in 2012 with the publication of *ImageNet Classification with Deep Convolutional Networks,* by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. In this paper, the group explained the architecture of AlexNet: 5 convolutional layers, 3 fully connected layers, and a 1000-way softmax function, which kickstarted the field of deep learning. Once the field began to be established, just like with machine learning, deep learning algorithms quickly began to be applied to the medical domain. There has been a substantial explosion in recent years, with 240 of 300 papers involving the application of deep learning to medical image analysis being published in 2016 or 2017 (Litjens et al., 2017). The application domain is incredibly diverse, including everything from the diagnosis of Alzheimer's disease (Suk & Shen, 2013) to the detection of diabetic retinopathy (Gulshan et al., 2016).

This paper focuses on applying these technologies to predict pediatric bone age based on x-ray images. This paper provides a brief overview of the existing literature using deep learning for medical images and will subsequently describe our approach for achieving optimal accuracy for the target problem.

RELATED WORK

Because many of the first major strides in the field of deep learning were with large sets of 2D images like ImageNet, it was a natural progression to apply these existing networks to similar problems in the medical field. Instead of classifying cats or dogs or flowers, the networks are now classifying chest x-rays as lateral or frontal for example (Rajkomar et al. 2016). Image classification simply takes an input, in this case one or multiple images that constitute an exam and produces a single output or diagnosis. According to Litjens et al., early on there were 2 major strategies. (1) using a pretrained network that was developed by someone else or (2) manually fine-tuning a pre-trained network on medical data. Both strategies have been used widely because of their relative ease of use. They simply needed to run new data sets through existing networks and hope it would provide good results. While the pediatric bone age problem is a regression problem not a classification problem, much of the principles and processes still apply, with the only changes being the final layers of the network.

A 2017 paper by Spampinato et al. corroborated the findings of Menegola et al. that specifically designed and trained networks will have better results than pre-trained ones. Spampinato et al. used several different networks to assess bone age, including GoogLeNet, OxfordNet and OverFeat. They found that fine-tuning the existing networks to more closely apply to the problem domain resulted in a 30% boost in performance; however, they also created a CNN architecture from scratch called BoNet, which in the end yielded the best results. They found that while many "off the shelf" networks are effective, custom CNN's can be the most effective with sufficient data and technical expertise.

A paper by Lakhani and Sundaram (2017) took a similar approach, using existing networks to a different domain. They looked at detecting tuberculosis in chest radiographs also using existing networks such as AlexNet and GoogLeNet, two famous and powerful CNN's. They used both trained and untrained versions of the network. The trained versions had previously been trained on the ImageNet dataset while the others had not; however, all the networks were still trained on a subset of the chest radiographs. Even so, the pre-trained networks had better accuracy, as shown in the figure below. This paper lends credence to similar findings that including large amount of non-medical images can also be helpful in getting accurate results (Bar et al. 2015).

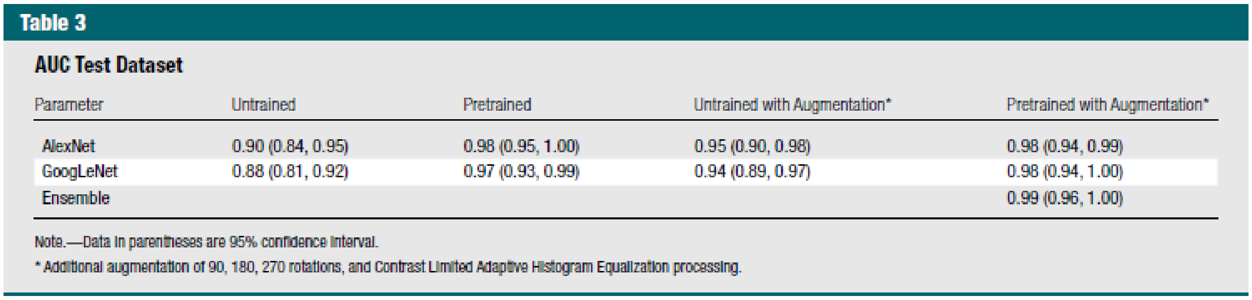

Table 3

AUC Test Dataset

Parameter	Untrained	Pretrained	Untrained with Augmentation*	Pretrained with Augmentation*
AlexNet	0.90 (0.84, 0.95)	0.98 (0.95, 1.00)	0.95 (0.90, 0.98)	0.98 (0.94, 0.99)
GoogLeNet	0.88 (0.81, 0.92)	0.97 (0.93, 0.99)	0.94 (0.89, 0.97)	0.98 (0.94, 1.00)
Ensemble				0.99 (0.96, 1.00)

Note.—Data in parentheses are 95% confidence interval.
* Additional augmentation of 90, 180, 270 rotations, and Contrast Limited Adaptive Histogram Equalization processing.

Figure 1 – AUC Test Dataset (Lakhani and Sundaram 2017)

In the examples above, the existing networks were trained on at least some medical data; however, there have been some papers that showed promising results even when the network was *only* trained on non-medical data. This was shown in Bar et al's paper (2015) where they looked at pathology in chest radiographs using the Decaf pre-trained CNN, which is a slightly modified version of a CNN developed by Krizhevsky et al. in 2012. They trained this CNN on the famous ImageNet dataset and then applied the network to 433 chest x-rays that were examined and labeled by radiologists. The network achieved around 90% accuracy when attempting to detect various types of pathology such as right pleural effusions and cardiomegaly, an impressive accuracy for networks not trained on medical data.

There has been success in many different medical domains using many different techniques. If there is sufficient data, resources, and technical expertise, then creating custom network architectures is likely to yield the most accurate results. However, in the absence of those things using an existing network can also show promise. Even here there is freedom to use networks that were pre-trained on non-medical images, or those that were not.

The current literature on applying deep learning to medical image analysis explains the multitude of techniques and application of this technology. Deep learning algorithms have already been applied to many different body parts and data types. It has become clear that convolutional neural networks are the technology of choice for most classification problems, and regression problems which are similar. It has also become clear that through the recent years, large strides have been made towards optimizing well known CNN architectures, and that using existing and well tested networks can provide sufficiently accurate results.

OUR APPROACH

In this section, we will discuss the various steps, methods, and approaches used to gain an understanding of convolutional neural networks and how to apply them to pediatric hand x-rays as well as techniques to achieve optimal accuracy.

After conducting our literature review, it became clear that the most feasible approach was to leverage an existing CNN that was already proven to be effective at image recognition. We also planned to leverage the additional accuracy and reduced training costs of transfer learning. As a result, we chose a pre-trained version of InceptionV3, a CNN developed by Google that placed highly in ImageNet competitions at the time it was developed.

Dataset

The training dataset consisted of 12,611 pediatric hand x-rays ranging from 1 month to 228 months old. The set consisted of both genders and a mixture of left and right hands, although the majority were left handed. All the images were provided by the Radiological Society of North America, acquired from Stanford Children's Hospital and Colorado Children's Hospital and were labelled with corresponding skeletal ages. A test set of 200 labelled images was also provided by RSNA.

Pre-processing

The provided images were of varying dimensions and needed to be resized to 299x299, the input size of InceptionV3. To preserve the aspect ratio of the original images, either horizontal or vertical padding was added to all images. In addition, the input images needed to be represented as colored images due to expectations of the InceptionV3 architecture, despite the original images being greyscale.

During various training cycles, data augmentation was also used in an attempt to prevent overfitting and improve generalization. 90-degree rotation, vertical and horizontal flipping, and height and width shifts were all employed with varying degrees of success. Overall, data augmentation did not prove effective in increasing accuracy. Instead, loss plateaued after only 15 epochs, perhaps reaching but not crossing a local minimum. In addition, accuracy was substantially worse at similar levels of training without augmentation.

Evaluation

All the results reported were tested against the provided test set of 200 images that was completely independent from the training set. All versions of the model were trained against the entirety of the training set.

The effectiveness of the model was determined by the mean absolute error of the predictions compared to the actual labels. The difference between actual and predicted was measured and averaged across all inputs in the test set.

Implementation Details

This project was done using Python, Tensorflow, and Keras using either custom code, pre-trained models, or other public libraries. All training and testing was done with a quad core 3GHZ CPU, 16 GB of RAM and an Nvidia GTX 1080 GPU with 8GB of VRAM.

RESULTS

Performance of the InceptionV3 network was recorded after scoring the test set with various parameters. Overall, the lowest average error of 9.71 months was achieved with 120 epochs of training data, no augmentation, and using the adam optimizer as displayed in Table 1. During testing, we found that after more than 120 epochs, loss was continuing to minimize; however, average error began to rise, likely indicating overfitting on the training set.

Also shown in Table 1 was the ineffectiveness of data augmentation, resulting in a more than 2x increase in average error after 5 epochs of training and a 3x increase in error after 30 epochs. Even after nearly 30 epochs of training, loss remained plateaued at around 32.00.

Additionally, we achieved the best results using the adam optimizer. It is not clear exactly why adam yielded better results than rmsprop, however, both were very effective due to their adaptive learning rate.

The results shown in Table 1 also display the ineffectiveness of transfer learning for hand x-rays. Theoretically, transfer learning allows for a reduction in training time and better defense against over-training by exposing the network to a variety of images before fine-tuning it on the desired target set. By using a version of InceptionV3 that was pre-trained on ImageNet, we were expecting to reap those benefits; however, after testing, the results showed that there was no noticeable difference in average error from the model between weights from training on ImageNet and randomized weights.

Weights	Training (epochs)	Augmentation	Optimizer	Results
ImageNet	30	0	rmsprop	11.69
ImageNet	90	0	rmsprop	10.99
ImageNet	120	0	rmsprop	10.12
ImageNet	120	0	adam	9.71
ImageNet	150	0	adam	10.1
Random	0	0	rmsprop	132.35
ImageNet	0	0	rmsprop	126.95
Random	5	0	rmsprop	16.04
ImageNet	5	0	rmsprop	16.3
Random	5	1	rmsprop	35.55
ImageNet	5	1	rmsprop	38.2
ImageNet	30	1	rmsprop	32.09

Table 1 – Performance Results from InceptionV3 based on Mean Absolute Error

DISCUSSION

In this paper we used deep learning and convolutional neural networks to effectively predict the bone age of pediatric hand x-rays. We trained InceptionV3 on over 12,000 images provided by the Radiological Society of North America and received a respectable average error of 9.71 months. Throughout this process we also discovered some additional insights about applying CNNs to medical imaging problems.

- *Transfer learning using ImageNet and InceptionV3 does not appear to be effective in increasing accuracy*. It is possible that this could be the result of some underlying piece of InceptionV3's architecture or a problem with the ImageNet dataset specifically when being applied to x-rays. Our lack of success here does not definitively mean transfer learning is ineffective. It is possible that ImageNet was simply too foreign from the target dataset to learn anything transferable, and if instead Inception was pre-trained on unrelated x-ray images then transfer learning could have been effective. However, this result does appear to contradict a substantial portion of the literature that has shown the effectiveness on non-medical transfer learning being applied to 2D medical images (Bar et al. 2015).

- *Data augmentation did not have the desired effects of increased accuracy and generalization*. In our limited tests we found that data augmentation increased training times and decreased accuracy. After 5 epochs data augmentation yielded an average error twice and large and after 30 epochs an error nearly 3 times as large with a loss that had plateaued. It is possible that we reached a local minimum and could have eventually improved given enough training time, but we were unable to test this hypothesis.
- *CNN's display incredible potential for medical imaging analysis*. Our simple approach of using existing architectures and making few changes throughout the pipeline already yielded excellent results. With time and more sophisticated techniques deep learning will prove to be an exceptionally powerful tool for radiologists.

In addition to the above findings, this paper will serve as a pedagogical tool for the Bryant community to learn more about an incredibly powerful emerging technology.

REFERENCES

Arimura, Hidetaka, Shigehiko Katsuragawa, Kenji Suzuki, Feng Li, Junji Shiraishi, Shusuke Sone, and Kunio Doi. "Computerized Scheme for Automated Detection of Lung Nodules in Low-Dose Computed Tomography Images for Lung Cancer screening." *Academic Radiology* 11, no. 6 (June 1, 2004): 617–29.

Brosch, T., Tam, R., Initiative, A. D. N., & others. (2013). Manifold learning of brain MRIs by deep learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 633–640). Springer.

Chen, Yen-Chen, Wan-Chi Ke, and Hung-Wen Chiu. "Risk Classification of Cancer Survival Using ANN with Gene Expression Data from Multiple Laboratories." *Computers in Biology and Medicine* 48 (May 2014): 1–7.

Chen, H., Dou, Q., Yu, L., & Heng, P.-A. (2016). VoxResNet: Deep Voxelwise Residual Networks for Volumetric Brain Segmentation. *arXiv:1608.05895 [Cs]*.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *arXiv:1606.06650 [Cs]*.

Grt123 2017. *The Solution of Team "grt123" in DSB2017*. Python. https://github.com/lfz/DSB2017.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., … Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, *316*(22), 2402.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. "Deep Residual Learning for Image Recognition." *arXiv:1512.03385 [Cs]*, December.

Hilario, Melanie, Alexandros Kalousis, Markus Müller, and Christian Pellegrini. "Machine Learning Approaches to Lung Cancer Prediction from Mass Spectra." *PROTEOMICS* 3, no. 9 (September 1, 2003): 1716–19.

Huang, Gao, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. 2016. "Deep Networks with Stochastic Depth." *arXiv:1603.09382 [Cs]*, March.

Ji, S., Xu, W., Yang, M., and Yu, K., "3D Convolutional Neural Networks for Human Action Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, Jan. 2013.

Kononenko, I. (1993). Inductive and Bayesian Learning in Medical Diagnosis. *Applied Artificial Intelligence*, *7*(4), 317–337.

Kononenko, I., Bratko, I., & Kukar, M. (1997). Application of machine learning to medical diagnosis. Machine Learning and Data Mining: Methods and Applications, 389, 408.

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, *23*(1), 89–109.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kuan, Kingsley, Mathieu Ravaut, Gaurav Manek, Huiling Chen, Jie Lin, Babar Nazir, Cen Chen, Tse Chiang Howe, Zeng Zeng, and Vijay Chandrasekhar. 2017. “Deep Learning for Lung Cancer Detection: Tackling the Kaggle Data Science Bowl 2017 Challenge.” *arXiv:1705.09435 [Cs]*, May.

Lakhani, Paras, and Baskaran Sundaram. “Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks.” *Radiology* 284, no. 2 (2017): 574–82.

LeCun, Y; Boser, B; Denker, J; Henderson, D; Howard, R; Hubbard, W; Jackel, L, “Backpropagation Applied to Handwritten Zip Code Recognition,” in Neural Computation , vol.1, no.4, pp.541-551, Dec. 1989 89

Lim, S. E., Xing, Y., Chen, Y., Leow, W. K., Howe, T. S., & Png, M. A. (2004). Detection of femur and radius fractures in x-ray images. In *Proc. 2nd Int. Conf. on Advances in Medical Signal and Info. Proc* (Vol. 65).

Linnainmaa, S. (1970). The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master’s thesis, Univ. Helsinki.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017). A Survey on Deep Learning in Medical Image Analysis. *arXiv:1702.05747 [Cs]*.

Menegola, A., Fornaciali, M., Pires, R., Avila, S., & Valle, E. (2016). Towards Automated Melanoma Screening: Exploring Transfer Learning Schemes. *arXiv:1609.01228 [Cs]*.

Olczak, Jakub, Niklas Fahlberg, Atsuto Maki, Ali Sharif Razavian, Anthony Jilert, André Stark, Olof Sköldenberg, and Max Gordon. “Artificial Intelligence for Analyzing Orthopedic Trauma Radiographs.” *Acta Orthopaedica* 0, no. 0 (July 6, 2017): 1–6.

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., & Calhoun, V. D. (2013). Deep learning for neuroimaging: a validation study. *arXiv Preprint arXiv:1312.5847*.

Rajkomar, A., Lingam, S., Taylor, A. G., Blum, M., & Mongan, J. (2017). High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. *Journal of Digital Imaging*, *30*(1), 95–101.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [Cs]*.

RSNA Pediatric Boneage Challenge. (2017). http://rsnachallenges.cloudapp.net/competitions/4. [Data set]: https://www.dropbox.com/s/1lys03k6n7uyim9/boneage-training-dataset.zip?dl=0

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323, 533–536

Shin, Hoo-Chang, Holger R. Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M. Summers. “Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning.”, February 10, 2016.

Spampinato, C., S. Palazzo, D. Giordano, M. Aldinucci, and R. Leonardi. “Deep Learning for Automated Skeletal Bone Age Assessment in X-Ray Images.” *Medical Image Analysis* 36 (February 2017): 41–51.

Suk, H. I., & Shen, D. (2013). Deep learning-based feature representation for AD/MCI classification. *Unknown Journal*, *16*(Pt 2), 583–590.

Werbos, P. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University, Cambridge, MA, 1974.

Y. Bar, I. Diamant, L. Wolf, S. Lieberman, E. Konen and H. Greenspan, "Chest pathology detection using deep learning with non-medical training," *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, New York, NY, 2015, pp. 294-297.

Zhou, Zhi-Hua, Yuan Jiang, Yu-Bin Yang, and Shi-Fu Chen. “Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles.” *Artificial Intelligence in Medicine* 24, no. 1 (January 1, 2002): 25–36. doi:10.1016/S0933-3657(01)00094-X.