PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ
ESCUELA DE POSGRADO

"AUTOMATIC LYMPHOCYTE DETECTION ON GASTRIC CANCER IHC IMAGES USING DEEP LEARNING"

Artículo para optar el grado de Magíster en Informática con Mención en Ciencias de la Computación

AUTOR
EMILIO RAFAEL GARCÍA RÍOS

ASESOR
CÉSAR ARMANDO BELTRÁN CASTAÑON

JURADO
IVAN ANSELMO SIPIRAN MENDOZA
CÉSAR AUGUSTO OLIVARES POGGI

LIMA - PERU
2017

Automatic Lymphocyte Detection on Gastric Cancer IHC Images using Deep Learning

Emilio Garcia, Renato Hermoza, Cesar Beltran Castanon
GRPIAA - Departamento de Ingeniería
Pontificia Universidad Católica del Perú
San Miguel, Lima, Peru

Luis Cano, Miluska Castillo, Carlos Castañeda
Departamento de Investigación
Instituto Nacional de Enfermedades Neoplásicas
San Borja, Lima, Peru

***Abstract*—Tumor-infiltrating lymphocytes (TILs) have received considerable attention in recent years, as evidence suggests they are related to cancer prognosis. Distribution and localization of these and other types of immune cells are of special interest for pathologists, and frequently involve manual examination on Immunohistochemistry (IHC) Images. We present a model based on Deep Convolutional Neural Networks for Automatic lymphocyte detection on IHC images of gastric cancer. The dataset created as part of this work is publicly available for future research.**

***Keywords*-cell detection; deep learning; immunohistochemistry; gastric cancer**

I. Introduction

Gastric Cancer is one of the five most common types of cancer among men and women according to the World Health Organization. It represents 7% of all cases and 9% of all deaths among cancer patients.[1]. With 12 months of disease-specific survival and 90% of all cases dying within the first five years, it is one of the most aggressive and deadliest types of cancer. Therefore it is of interest for medical professionals to accurately estimate patients' prognoses.

Current evidence suggests that TILs may have an association with the prognoses and clinical features of cancer patients. In some scenarios, high concentrations of Cytotoxic T Cells and Memory T cells (T-Mem) are associated with favorable results, whereas elevated levels of Regulatory T Cells (T-Reg) may contribute to the pathogenesis[2].

However, there are other scenarios in which this relation changes drastically. For example, in the specific case of gastric cancer of the cardia, high levels of tumor-infiltrating macrophages are favorable for carcinogenesis, whereas elevated levels of T-Reg lead to better clinical results.

Moreover, not only the type of immune cells but also their location is of interest for pathologists, as the concentration of tumor-infiltrating T-Reg is only relevant to the prognosis if found in the tumor stroma[3].

*Supplementary material is available at https://github.com/grpiaa-pucp/prisma-public/tree/master

Over the last years, increasing computer power has been enabling considerable improvements in different areas of machine learning. Multilayered Neural Networks are making impressive progress in computer vision tasks[3]. There are even cases where convolutional neural networks have achieved better performance than humans[4, 5].

In this paper, we present an approach to automatically detect and count TILs on IHC images of gastric cancer using Deep Convolutional Neural Networks. We also describe an innovative approach to label images for the training and testing sets using a piece of software built in-house. One of the contributions of this work is its validated methodology and results, as no previous work has been found related to gastric cancer. Another significant contribution is the lymphocyte images dataset which can be downloaded and used for further research and improvements in the area. The proposed software has a great potential to be used as a tool for helping medical professionals and researchers in further cancer studies.

We organized the rest of this paper as follows: we introduce Deep Learning and Convolutional Neural Networks along with related work in Section 2, the Methodology applied is described in Section 3, and Experiments are presented in Section 4. Finally, we present discussion, conclusions and future work in Section 5.

II. Previous and Related Work

The idea of replacing hand-engineered features with multilayer networks existed since a long time in pattern recognition[6, 7]. However, a working solution for training multilayered architectures with simple stochastic gradient descent was not developed until the 1970s and 1980s[8, 9, 10]. Later, it was about in 2006 that the field regained attention as Deep Learning and became an important area in Machine Learning.

Deep Learning techniques allow machines to identify patterns and recognize images and voices. Its goal is for machines to be able to learn without needing prior preparation of training data (manual feature extraction)[11]. In the last few years, it has changed the way of working in signal processing considerably[12].

Advances in signal processing research, big data, and the drastic increment of computing power in CPU and GPU have made popular the use of Deep Learning. Consequently, there are also substantial improvements in the applications of Deep Learning for computer vision and object recognition. Examples of successful applications include supervised and non-supervised feature extraction and classification tasks[13].

In 2011, *Cirean et al.* achieved better performance than humans for the first time with an object recognition task[4]. Later, in February 2012, he reported a new error rate of 0.23% in the MNIST handwritten digit recognition problem[13]. The same year *Krizhevsky et al.* won the ImageNet competition by combining convolutional neural networks and max-pooling with graphic processing units[14].

Later in 2015, *Delahunt et al.* built a device named Autoscope that enables users to diagnose malaria using a digital microscope and a combination of algorithms for computer vision and image classification[15].

Recently, in March 2016, *Abdel-Zaher and Eldeib* achieved 99.68% of accuracy in breast cancer classification. These results were very promising compared to previous work[16]. Similarly, other works that have used Deep Learning techniques outperformed previous studies in the areas of prediction and detection in medicine[17, 18, 19].

In biological and medical fields, *Pan et al.* presented an average accuracy of 78.6% for detection of lung cancer cells, whereas *Chen and Chefd'hotel* reported to achieve a coefficient correlation between manual and automated counting of up to 99.49% using image decomposition and convolutional neural networks[20, 21]. Several other achievements in medical and microscopy image analysis made possible by deep learning have also been reported in literature, showing that CNNs are the most popular among other types of models in this type of applications[22].

III. Methodology

A. Overview

An overview of our model is presented in 1. Gastric cancer biopsy tissue was scanned at 40x magnification. The acquired images were later used to extract smaller patches for single cells. Each of the single-cell samples received a positive/negative label and was used to create test and training sets.

The training data was feed into a Deep Convolutional Neural Network to learn classifying single images as positive or negative whether they are lymphocyte cells or not.

After training the network, we evaluated individual patches from test images using a sliding window algorithm. Next, we calculated the final output by applying non-maximum suppression to all single classifications in the previous step.

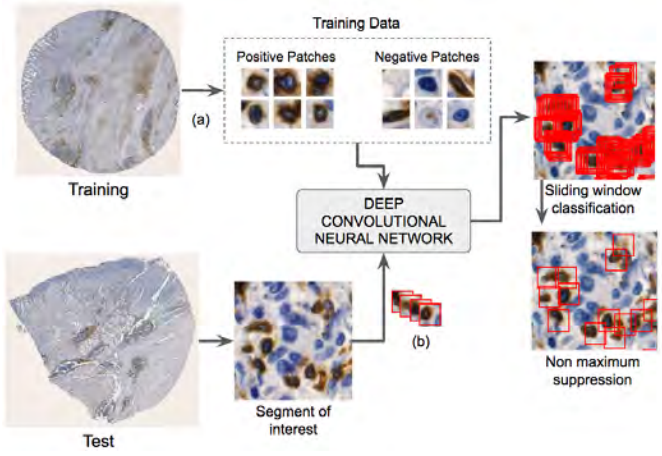


Figure 1. Overview of the proposed model.

B. Labeling tool

Due to the difficult on extracting and labeling cell patches manually, we provided a web application that allows for image visualization with a frame-like cropping tool and asked experts to label the patches by themselves. The process was as simple as identifying a cell and double-clicking on it. The labeling tool in reference is shown in Figure 2.

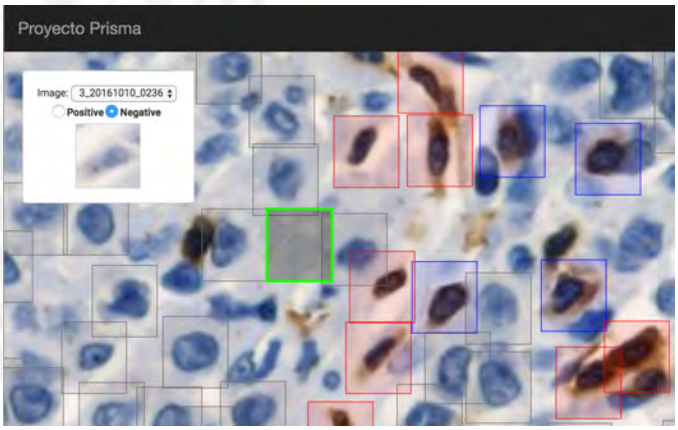


Figure 2. Labeling tool.

C. Deep Convolutional Neural Network

Our Deep Convolutional Neural Network (DCNN) has nine layers, excluding input, which are arranged in three convolutional layers (C1, C2, and C3), three max-pooling layers (MP1, MP2, and MP3), and two fully connected layers (FC1 and FC2) with dropouts. The output layer is a fully-connected layer using the softmax function as the activation function; it outputs two classes (lymphocyte or non-lymphocyte). We show the complete network architecture in Figure 3.

The convolutional layers C1, C2, and C3 use 3x3 kernels with strides = [1, 1, 1, 1] and their outputs are passed to

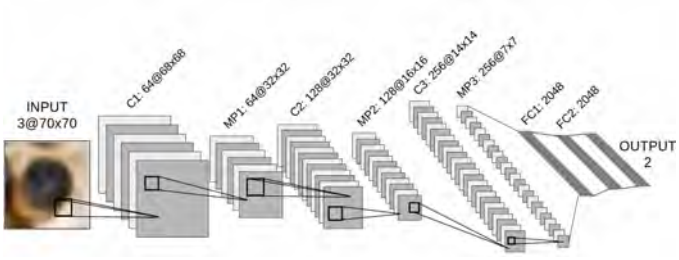


Figure 3. Deep Convolutional Neural Network Architecture.

the max-pooling layers MP1, MP2, and MP3 which have kernel size and strides = [1, 2, 2, 1]. The number of features in each pair of convolutional and max-pooling layers are 64, 128, and 256 respectively. We later reshape the output of the 3rd max-pooling layer to a vector. Finally, in both fully connected layers, FC1 and FC2, we have 2048 feature vectors.

We used rectified linear units (ReLU), dropouts and batch normalization as it is known that they are useful for overfitting reduction and performance improvement with unseen data[23, 24]. For optimization, we used the ADAM algorithm, which was originally proposed by *Kingma and Ba* in [25].

IV. Experiments

In this section, we describe the dataset and experiments setting. We also validate the proposed algorithm and compare it to medical professional output.

A. Data Set

A clinical dataset containing gastric cancer tissue samples was used to test the proposed model. All IHC images were acquired using an Olympus BX63 Microscope and CD3 stains.

We extracted the data from 10 biopsy micrographs of the cancerous gastric tissue scanned at 40x magnification using a software utility developed as part of this work. Another 35 additional 600x500 pixel images were used to compare against the results provided by human pathologists.

For training, pathologists extracted and labeled 70x70 pixel patches containing individual cells from IHC images. All cells stained with CD3 that were identified as lympohcytes by the human pathologist were labeled as Positive whereas all other cells including contrast stained cells were labeled as Negative. Then, we trained a convolutional deep neural network to make binary classifications for inputs of the same size.

After extracting individual cell patches, the database consisted of 3,257 images (70 x 70 pixel each). Then, we used data augmentation techniques such as rotation and reflection to increase the dataset size and got a total of 10868 individual cell patches. The distribution of the dataset for training and testing is shown in Table I.

Table I
Description of the Data Set

Set	Quantity
Before data augmentation	3257
After data augmentation	10868
Positive Train Samples	4437
Negative Train Samples	4257
Positive Test Samples	1143
Negative Test Samples	1031

B. Experiment Setting

Before experiments, the training set was divided into subsets (train and validation) in order to execute cross-validation for each of the parameter configurations. Elements for each subset were chosen randomly.

We used a random search approach to test several parameter configurations. The rules for each of the parameters are described in Table II below:

Table II
Random Search Criteria for Parameters

Parameter	Value Range
Training Steps	500 - 18,000
Learning Rate	$10^-3.5 - 10^-1.5$
Regularization Strength	$10^-3.5 - 10^-2$
Early Stoping Rounds	200-300

The experiments goal was to maximize precision and we used an early stopping rule to stop training when the model didn't show improvement over the last 200-300 steps. We also used Softmax Cross Entropy as the cost function. For validation, we used three specific metrics: Accuracy, Precision and Recall.

C. Performance on single-cell images

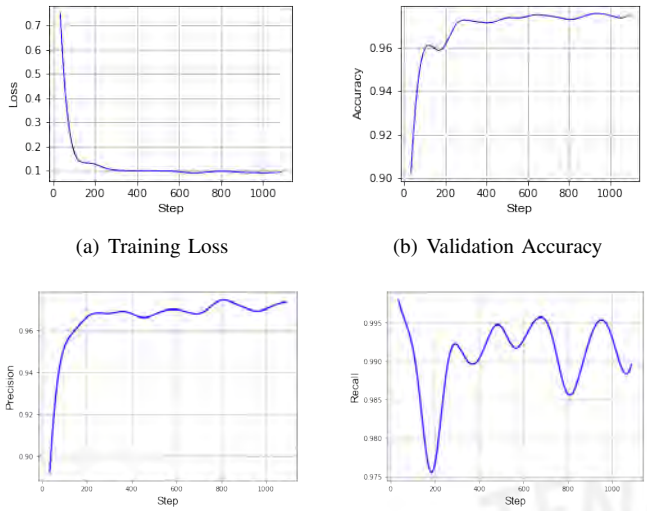
In training and validation, our best model achieved near 98.2% accuracy after 1,000 iterations with no significant improvements until iteration 2,000. Figure 4 shows the Training Loss (a), Validation Accuracy (b), Validation Precision (c), and Validation Recall (d) for our best model.

In the test set, the model achieved an overall 94% Precision, 95.83% F-score, and 92% Recall. These values are shown in Table III.

Table III
Performance in test set

Accuracy	Precision	F-score	Recall
94.0	95.83	93.88	92.0

Of the actual positive samples in the test set, 92% were correctly classified while 96% of the total classifications were correct for negative samples.



(a) Training Loss (b) Validation Accuracy

(c) Validation Precision (d) Validation Recall

Figure 4. Training and Validation Metrics.

Figures 5 and 6 show the normalized confusion matrix and Receiver operating characteristic curve respectively.

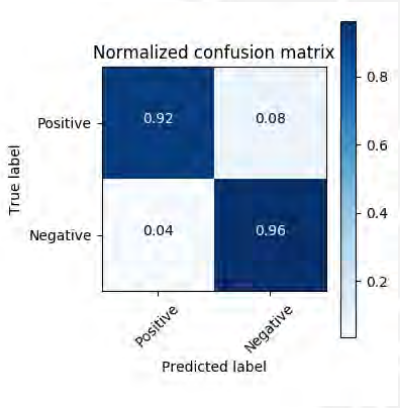


Figure 5. Confusion matrix.

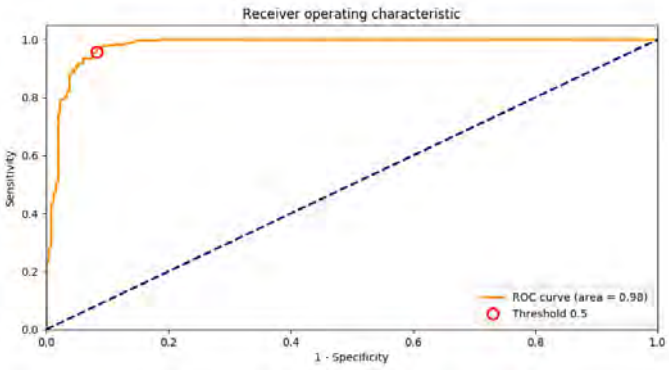


Figure 6. ROC curve.

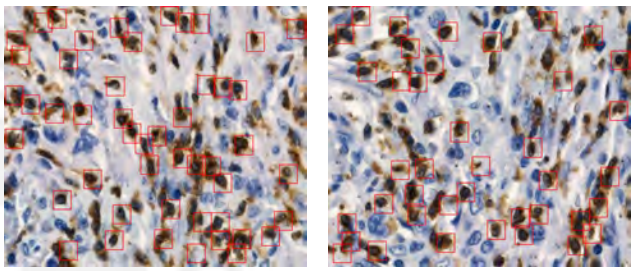


Figure 7. Classification on 600 x 500 pixel images

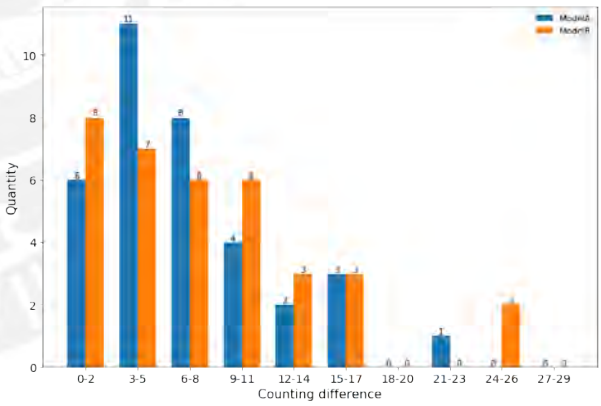


Figure 8. Behavior in counting differences between augmented and non-augmented datasets

D. Comparison against human experts

We used 35 images, 600 x 500 pixel each, to compare the model results against human pathologists. These validation images were randomly selected them from new full-sized IHC micrographs. Samples of classification outputs are shown in Figure 7.

When analyzing results using data augmentation, 17 images out of 35 had an output difference in terms of detected cells equal to or lower than 5, whereas a total of 29 had a count difference equal to or lower than 11. Therefore, only six images had a count difference higher than 11.

If we compare our best model using data augmentation (Model A) against our best model without data augmentation (Model B), better results are seen in Model A.

However, both models present the same behavior: as the difference between numbers of human and algorithm outputs increase, the number of images exhibiting those differences decrease. Figure 8 explains this relationship in a more visual fashion.

V. Discussion and Conclusions

We present an approach for automatic TILs detection and counting on IHC images of gastric cancer using Deep Convolutional Neural Networks. Our experiments produced an acceptable 94% Accuracy and 95.83% Recall. It is also

valuable that our work received exhaustive validation with the pathologist. A similar work was found in the literature [21], but we enhanced the segmentation and labeling process. While *Chen and Chefd'hotel* applied a semi-automatic segmentation of cells, we implemented a fully automated segmentation and labeling process. We also developed an innovative approach to collect images for the training set using a piece of software built in-house. Another significant difference is the number of instances for the training set, while *Chen and Chefd'hotel* extracted manually 491 positives and 539 negative samples, we extracted 1395 positive and 1322 negative samples, this new dataset is a significant contribution to the scientific community and is available as supplementary material. Other works and methods can also use the new dataset created as part of this work to continue improving the accuracy in lymphocyte detection. Finally, the provided software, which made the patches extraction easier, can also be used to create datasets for other types of cells or images. While there is room for improvement, the proposed method has potential to be used as a resource for helping medical professionals and researchers in further cancer studies.

Acknowledgment

The authors would like to thank Joselyn Sánchez and Luis Bernabé for their helpful contribution while creating the lymphocytes dataset. We are also thankful to advisors Ivan Sipiran and César Olivares for their excellent reviews and suggestions, Corrado Daly for his contribution getting our computing infrastructure ready, and to all of our friends and colleagues for their valuable comments and discussions that helped to better shape this work.

References

[1] S. W and W. P, *World Cancer Report 2014 (PDF)*, ser. IARC Nonserial Publication. International Agency for Research on Cancer (I A R C) (UN), 2014.

[2] A. Amedei, C. D. Bella, E. Silvestri, D. Prisco, and M. M. D'Elios, "T Cells in Gastric Cancer: Friends or Foes." *Clinical & Developmental Immunology*, pp. 1–10, Jan. 2012.

[3] C. Edwards, "Growing Pains for Deep Learning." *Communications of the ACM*, vol. 58, no. 7, pp. 14–16, Jul. 2015.

[4] D. Cirean, U. Meier, J. Masci, and J. Schmidhuber, "Multi-column deep neural network for traffic sign classification," *Neural Networks*, vol. 32, pp. 333 – 338, 2012, selected Papers from {IJCNN} 2011.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[6] O. G. Selfridge, "Pandemonium: a paradigm for learning in mechanisation of thought processes," 1958.

[7] F. Rosenblatt, *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957.

[8] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive modeling*, vol. 5, no. 3, p. 1, 1988.

[9] Y. LeCun, "Une procédure d'apprentissage pour réseau a seuil asymmetrique (a learning scheme for asymmetric threshold networks)," in *Proceedings of Cognitiva 85, Paris, France*, 1985.

[10] D. Parker, "Learning logic (report no. 47). cambridge, massachusetts institute of technology," *Center for Computational Research in Economics and Management Science*, 1985.

[11] "How Deep Learning Will Change Our World." *Trends Magazine*, no. 156, pp. 27–30, 2016.

[12] L. Deng, D. Yu *et al.*, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[13] D. C. Ciresan, U. Meier, and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," *CoRR*, vol. abs/1202.2745, 2012.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[15] C. B. Delahunt, C. Mehanian, L. Hu, S. K. McGuire, C. R. Champlin, M. P. Horning, B. K. Wilson, and C. M. Thompon, "Automated microscopy and machine learning for expert-level malaria field diagnosis," in *Global Humanitarian Technology Conference (GHTC), 2015 IEEE*. IEEE, 2015, pp. 393–399.

[16] A. M. Abdel-Zaher and A. M. Eldeib, "Breast cancer classification using deep belief networks," *Expert Systems With Applications*, vol. 46, pp. 139–144, Mar. 2016.

[17] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen, "Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in us images and pulmonary nodules in ct scans," *Scientific reports*, vol. 6, p. 24454, 2016.

[18] Guo-Ping Liu, Jian-Jun Yan, Yi-Qin Wang, Wu Zheng, Tao Zhong, Xiong Lu, and Peng Qian, "Deep Learning Based Syndrome Diagnosis of Chronic Gastritis." *Computational & Mathematical Methods in Medicine*, pp. 1–8, Jan. 2014.

[19] P.-P. Ypsilantis, M. Siddique, H.-M. Sohn, A. Davies, G. Cook, V. Goh, and G. Montana, "Predicting Response to Neoadjuvant Chemotherapy with PET Imag-

ing Using Convolutional Neural Networks." *PLoS ONE*, vol. 10, no. 9, pp. 1–18, Sep. 2015.

[20] H. Pan, Z. Xu, and J. Huang, "An Effective Approach for Robust Lung Cancer Cell Detection." *Patch-Based Techniques in Medical Imaging*, p. 87, Jan. 2015.

[21] T. Chen and C. Chefd'hotel, "Deep Learning Based Automatic Immune Cell Detection for Immunohistochemistry Images." *Machine Learning in Medical Imaging: 5th International Workshop, MLMI 2014, Held in Conjunction with MICCAI 2014, Boston, MA, USA, September 14, 2014. Proceedings*, p. 17, Jan. 2014.

[22] G. Carneiro, Y. Zheng, F. Xing, and L. Yang, "Review of deep learning methods in mammography, cardiovascular, and microscopy image analysis," in *Deep Learning and Convolutional Neural Networks for Medical Image Computing*. Springer, 2017, pp. 11–32.

[23] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.