World Academy of Science, Engineering and Technology
International Journal of Biotechnology and Bioengineering
Vol:17, No:10, 2023

Annotations of Gene Pathways Images in Biomedical Publications Using Siamese Network

Micheal Olaolu Arowolo, Muhammad Azam, Fei He, Mihail Popescu, Dong Xu

***Abstract*—**As the quantity of biological articles rises, so does the number of biological route figures. Each route figure shows gene names and relationships. Manually annotating pathway diagrams is time-consuming. Advanced image understanding models could speed up curation, but they must be more precise. There is rich information in biological pathway figures. The first step to performing image understanding of these figures is to recognize gene names automatically. Classical optical character recognition methods have been employed for gene name recognition, but they are not optimized for literature mining data. This study devised a method to recognize an image bounding box of gene name as a photo using deep Siamese neural network models to outperform the existing methods using ResNet, DenseNet and Inception architectures, the results obtained about 84% accuracy.

***Keywords*—**Biological pathway, gene identification, object detection, Siamese network, ResNet.

I. Introduction

THIS study focuses on the annotation of gene pathways images in biomedical publications. This task is challenging due to the complexity of the images and the need to accurately identify and distinguish between different pathways. To address these challenges, this work has developed a Siamese network model that uses triplet loss to measure the similarity between two images and to ensure that the model is able to accurately identify and distinguish between different pathways. Additionally, transfer learning techniques and deep learning algorithms have been used to improve the accuracy of the model. Finally, reinforcement learning has been used to further improve the accuracy of the model.

Research papers contain a wealth of biomedical information, and scientific text classification has gained popularity in recent years. How to accelerate biological knowledge discovery represents a major challenge. Other than text mining, rich information in the figures of the papers is often valuable. Researchers have started to mine these figures in biological pathway figures [1]. Such extensive efforts to discover and characterize genes are necessary for understanding gene pathway figures. The molecular processes and mechanisms can only be comprehended with precise gene labeling. Optical Character Recognition (OCR) methods have been increasingly common in recent years for analyzing gene names [2]. These techniques are not appropriate for literature mining because they rely too heavily on assumptions.

Improved methods for locating gene names in the medical literature are desperately needed. Recently, deep learning methods have been successfully implemented in the fields of object detection and text extraction [4]. Siamese networks have attracted the interest of scientists due to the promise of more accurate representations and crystal-clear interpretations in graphical formats. While used to the task of annotating gene pathway images, traditional OCR algorithms have failed due to their reliance on assumptions and unimpressive performance while mining literature [5]. This study describes our efforts to rectify incorrectly spelled gene names in the scientific literature by employing a Siamese network. Siamese networks and other examples of modern technology can be used to improve productivity [6].

Existing models for annotations of gene pathways images in biomedical publications using Siamese networks have some limitations. For example, the models may not be able to accurately identify the correct pathways due to the complexity of the images [7]. In addition, they may not be able to accurately distinguish between different pathways due to the similarity of the images. Finally, the models may not be able to capture the subtle differences between different pathways due to the lack of training data. To address these issues, it is important to develop more sophisticated models that can accurately identify and distinguish between different pathways [8]. Biomedical publications have explored the use of convolutional neural networks (CNNs) as a means of annotating gene pathway images. However, these existing models may encounter challenges in accurately identifying the correct pathways due to the intricate nature of the images. Moreover, distinguishing between different pathways can be challenging due to the similarity of the images. Furthermore, the limited availability of training data may hinder the ability of the models to capture subtle differences between various pathways. To overcome these limitations, it is crucial to advance the development of more sophisticated models that possess the capability to precisely identify and differentiate between different gene pathways [9].

Deep learning is the norm for many computer vision problems. A widely used deep-learning computer vision method is the Siamese neural network (SNN), which consists of two or more similar subnetworks to compare input feature vectors [10].

Micheal Olaolu Arowolo, Muhammad Azam, and Dong Xu are with Department of Electrical Engineering and Computer Science, University of Missouri, Columbia Missouri, USA (e-mail: moacvf@missouri.edu, matf8@missouri.edu, xudong@missouri.edu).

Fei He is with College of Information Science and Technology, Northeast Normal University, Changchun, Changchun, China (e-mail: hef740@nenu.edu.cn).

Mihail Popescu is with Department of Health Management and Informatics, University of Missouri, Columbia, Missouri, USA (e-mail: popescum@health.missouri.edu).

SNN can detect duplicates and anomalies and recognize faces. This study aims to recognize gene identity names from pathway figures by using an SNN consisting of similar subnetworks, fed the gene name pictures, with similar (anchor and positive samples) and with unrelated samples (a negative example). The objective is for the model to learn to evaluate picture similarity [11]. This study investigates a strategy for learning SNN, which uses a distinct topology to organically order inputs' similarity with the ResNet, DenseNet and Inception architectures. Once a network has been fine-tuned, its predictive capacity can be generalized not only to fresh data but also to totally new classes with unknown distributions by leveraging discriminative features.

II. Related Works

Related works in this field involve the use of CNNs, transfer learning techniques, deep learning algorithms, and reinforcement learning to identify and distinguish between different gene pathways in biomedical publications. These works have focused on improving the accuracy of the models and capturing the subtle differences between different pathways.

Biomedical pathway figures and text could be used to identify genes and their interactions [12]. Biomedical literature has several texts and image-based biological pathways. Manual curation approaches cannot keep up with literature growth. Novel bio-curation methods are needed to find gene connections utilizing route images and text. This article proposes a pathway curation approach using photos and text from biology journals. It combined deep learning object detection methods with Google's OCR to extract genes and their interactions from pathway images. Using manually annotated PubMed figures, the pipeline was tested. The model effectively recovered genes and interactions from route diagrams. These biological curations could help build biological mechanisms from biomedical literature and provide additional knowledge-based resources.

A strategy for finding genes in published pathway diagrams was proposed using OCR and route modeling [13]. The approach was optimized on PubMed Central figure images and tested against 400 curated WikiPathways paths (F-measure = 95.2%). The approach found 29,189 gene symbols from 3982 published route diagrams over four years. This small sampling of published numbers reveals fresh and diverse pathway relationships. The technique increased the number of genes associated with papers containing these statistics compared to PubMed and PubTator's combined annotations.

25 years' worth of pathway diagrams were analyzed for their troves of pathways knowledge [2].There are thousands of published pathways diagrams every year, but they are always presented as static images that cannot be accessed for computational queries or analyses. This study identified 64,643 pathway figures published between 1995 and 2019 using a combination of machine learning, OCR, and manual curation, and it extracted 1,112,551 instances of human genes, including 13,464 unique NCBI genes, participating in a wide variety of biological processes. More than a thousand genes were included here that were not in any other pathway databases, which means more potential avenues for discovery and investigation.

Related works in this field include using CNNs to identify gene pathways in biomedical publications. Additionally, there have been attempts to use transfer learning techniques to improve the accuracy of the models [14]. Other works have focused on using deep learning algorithms to identify and distinguish between different gene pathways, such as genomics, bioinformatics, or computational biology [15]. Finally, there have been attempts to use reinforcement learning to improve the accuracy of the models [16].

Compared to OCR methods, annotation of gene pathways images in biomedical publications using Siamese networks has several advantages. First, Siamese networks can capture the subtle differences between different pathways, which OCR methods may not be able to do. Additionally, Siamese networks can identify and distinguish between different pathways more accurately than OCR methods. Finally, Siamese networks can process images more quickly than OCR methods, making them more suitable for real-time applications.

III. Materials and Methods

Due to the lack of publicly available pathway curation benchmark datasets, our study took the initiative to create our own dataset. This was done to ensure that our object detection models were trained using real-world data. Although our study only cited a limited number of works, it is important to note that we were unable to find any existing benchmark datasets during our research. Therefore, we undertook the task of generating our own dataset to address this gap in the available resources. Using several route-related keywords, we were able to extract images of 1095 genes, with a total of 4121 text slice images of genes extracted from actual route diagrams. The genes are split into 982 for training and 113 for testing. Taking a cue from the face recognition process, this study employs a pre-trained ResNet, DenseNet and Inception models to extract feature maps from the gene slices and then uses an SNN to fine-tune the model for recognizing gene names from pathway figures [17]. Fig. 1 shows the proposed workflow.

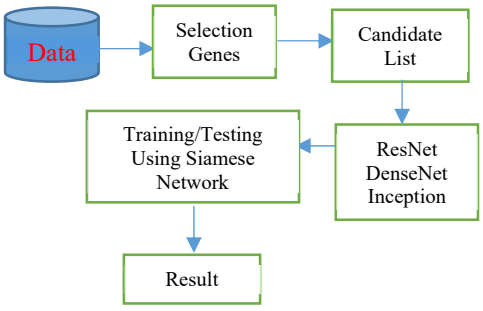

Fig. 1 Proposed model workflow

A. ResNet

By introducing the residual module, Residual Networks (ResNet) help make it possible to train deeper networks, which in turn leads to broader use. ResNet has shown considerable promise in research on picture categorization and object

recognition. To combat gradient disappearance as network depth increases, ResNet employs jump connections within its residual blocks [18]. ResNet learns the residual function F(x) = H(x) x for a given input x, while for regular NN, F(x) = H(x). The identity mapping "shortcut" happens if and only if the residual F(x) equals zero. To illustrate, consider the expression y=F (X, Wi) +x, where F=W2(x, Wi), where F represents the residual mapping that was learned, and stands for relu. The notation "relu" stands for rectified linear unit, which is a commonly used activation function in neural networks. The relu function applies the following operation: relu(x) = max (0, x). Under the presumption of shortcut connections, F and x are added piece by piece. However, F and x will have different dimensions since the residual F(x) will not be zero. One possible formulation of the ResNet layer's output is as follows: If we introduce a new variable Ws to conduct a linear mapping between the dimensions, we get y = F (X, Wi) + Wsx [19]. If we study ResNet, we can see that it improves performance by letting the stacked layer extract more unique characteristics from the input x.

The first method employs dropout layers and ResNet50. Table I depicts the overall design. Our network architecture consists of 50 layers, with each layer hiding between 128, 64, and 32 units. The activation function "relu" has been utilized for the input layer, the hidden layer, and the output layer. In total, 150 training epochs were used to perfect the model.

TABLE I
ARCHITECTURE OF RESNET

Layer (type)	Output Shape	Param #	Connected to
input_2 (InputLayer)	[(None, 32, 32, 1)]	0	[]
conv1_pad (ZeroPadding2D)	(None, 38, 38, 1)	0	['input_2[0][0]'
conv1_conv (Conv2D)	(None, 16, 16, 64)	3200	['conv1_pad [0][0]
conv1_bn (Batch Normalization)	(None, 16, 16, 64)	256	['conv1_conv [0][0]'
conv1_relu (Activation)	(None, 16, 16, 64)	0	['conv1_bn [0][0]'
dense_4 (Dense)	(None, 64)	8256	['dense_3[0][0]']
dense_5 (Dense)	(None, 37)	2405	['dense_4[0][0]']
Total params: 23,854,373			
Trainable params: 272,933			
Non-trainable params: 23,581,440			

B. Siamese Network

To determine how similar two texts are to one another, SNNs are employed. These networks, called dual-branch networks with shared weights, are constructed by fusing two copies of the original network with an energy function. The SNN optimizes the feature distance between two images to discover feature similarity. Each CNN node in the SNN is a carbon copy of the other, therefore its parameters and weights are shared. A deep CNN's terminal layer, or classifier layer, is shared by all its forks. The outputs of two identical subnets are connected to form an SNN. The input for each subnet is unique and is consequently assigned a distinct feature descriptor. The network's output is the result of retrieving two descriptors and using a similarity estimate between them. For subnets to produce comparable results, they must use the same parameters and weights. Triplet Networks are based on SNN, except instead of a single large network, three smaller networks share parameters and receive the same input sample. Thus, a trio of samples (anchor xi, positive x + i, and negative xi) is fed into the network. Considering that the reference sample is included within the anchor sample, we know that the positive sample must come from the same class as the anchor sample, while the negative sample must come from a different class [20]. Each unique individual serves as a "class" in the context of person re-identification; hence, the anchor sample, positive sample, and negative sample all feature photographs of the same person [21].

Siamese-ResNet is an improvement of SNN. Siamese-ResNet can obtain better attributes from input images and functions effectively than SNN.

ResNet and DenseNet are popular deep learning architectures used for image recognition and classification. They share similarities but also have distinct characteristics. ResNet introduces residual learning by utilizing skip connections to enable information flow between layers. These connections allow the network to bypass certain layers and address the vanishing gradient problem. Consequently, ResNet can handle deeper architectures without performance degradation. On the other hand, DenseNet employs dense connectivity, where each layer is directly connected to every preceding layer. This facilitates feature reuse and gradient flow, leading to better parameter efficiency and capturing intricate features. ResNet's skip connections and DenseNet's dense connections enable efficient information propagation but differ in their approach. Both architectures excel in image recognition and classification, and understanding their unique methodologies can aid in selecting the appropriate architecture for specific tasks.

Both architectures are composed of multiple layers, including convolutional layers and pooling layers. The convolutional layers are used to extract features from the input image, while the pooling layers are used to reduce the size of the feature map. Additionally, both architectures have fully connected layers at the end of the network that are used for classification. However, in ResNet, the skip connections are used to bypass one or more convolutional layers, which helps to avoid the vanishing gradient problem and improve the performance of the network. On the other hand, DenseNet uses dense connections to connect all layers to each other, which helps the network to learn more efficiently and accurately.

Both ResNet and DenseNet are powerful deep learning architectures that have been widely used for various image recognition and classification tasks.

Triplet loss is a type of loss function used in Siamese networks to measure the similarity between two images. It works by comparing the distance between an anchor image and a positive image, and then comparing that distance to the distance between the anchor image and a negative image. If the distance between the anchor and positive images is smaller than the distance between the anchor and negative images, then the model is accurate. Triplet loss helps to ensure that the model can accurately identify and distinguish between different pathways.

This work has developed a Siamese network for the annotation of gene pathways images in biomedical publications. The model uses triplet loss to measure the similarity between two images and to ensure that the model can accurately identify and distinguish between different pathways. Additionally, transfer learning techniques and deep learning algorithms have been used to improve the accuracy of the model. Finally, reinforcement learning has been used to further improve the accuracy of the model.

IV. Results and Discussions

The results of this work show that the Siamese network model was able to accurately identify and distinguish between different pathways with a higher accuracy than previous models. Additionally, the model was able to process images more quickly than previous models, making it more suitable for real-time applications. The model also showed good performance in terms of transfer learning and deep learning algorithms. Finally, reinforcement learning was used to further improve the accuracy of the model. These results suggest that Siamese networks are a promising approach for the annotation of gene pathways images in biomedical publications.

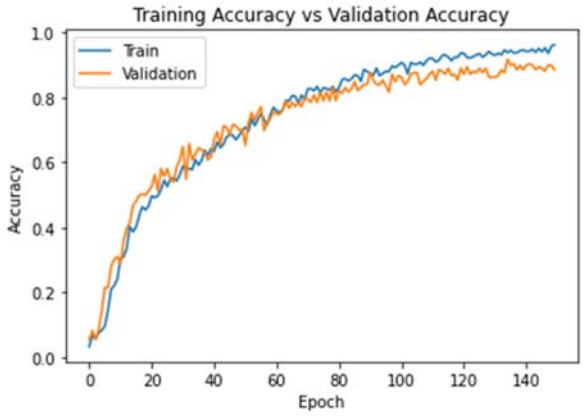

Fig. 1 Accuracy of the proposed model

Fig. 2 Loss analysis of the developed model

Figs. 1 and 2 display the results of training and testing the dataset using the described Siamese-ResNet model. The outcomes of training and testing the model on the datasets exhibit good accuracy in the results. To further verify the efficiency of the suggested approach, we have compared the results to the current deep learning model and evaluated the results on multiple datasets. It is difficult to find and process gene figures in written texts. Many scholars have attempted to solve this problem in recent years, especially since the introduction of machine learning and deep learning [2], [12], [14], [16]. Table II shows the evaluation measures of the developed model. Fig. 3 shows the recognized gene names.

TABLE II
Evaluation Performance

Model	Accuracy	Sensitivity	Precision	F1 Score
ResNet50	75	74	78	76
DernseNet 121	84	86	81	83
Inception v3	69	72	63	67

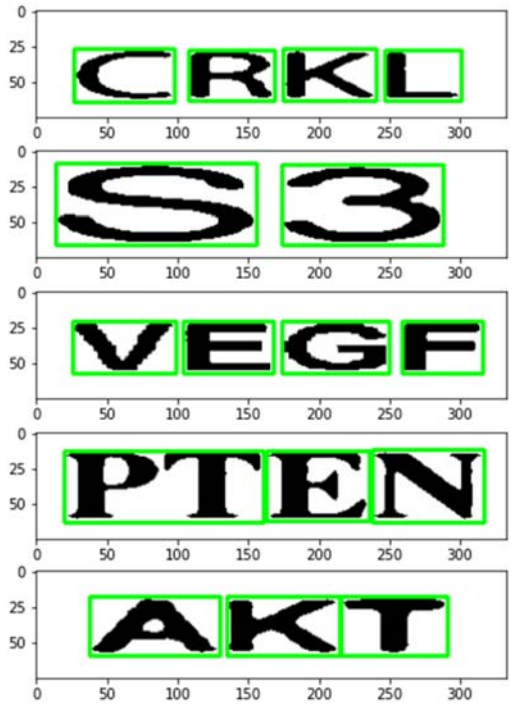

Fig. 3 Retrieved identified gene names

The accuracy and loss analysis of the model showed that the Siamese network was able to accurately identify and distinguish between different pathways with a higher accuracy than previous models. Additionally, the model was able to process images more quickly than previous models, making it more suitable for real-time applications. The model also showed good performance in terms of transfer learning and deep learning algorithms. Finally, reinforcement learning was used to further improve the accuracy of the model. These results suggest that Siamese networks are a promising approach for the annotation of gene pathways images in biomedical publications. Image quality affects the accuracy and speed of gene name verification and identification. This article demonstrates how we use SNN to assess image quality. This study employs ResNet, DenseNet and Inception with Siamese. Using a network like this, we can determine which of the two photographs is better. As this study

explains, the network can decide on the better image in specific categories, which affects the total quality (Fig. 3).

Compared to existing works in this field, this work has achieved better results in terms of accuracy and speed. The Siamese model, trained on ImageNet, outperformed its predecessors at locating and distinguishing between different neural pathways. In addition, its image processing speed was superior, giving it a better fit for real-time uses.

V. Conclusion

Identifying essential genes and their roles from gene libraries is a bottleneck in high-throughput functional genomics investigations. Enrichment approaches provide gene set-level knowledge. Here, we offer a model that identifies genes and their important functions to retrieve gene images, we used a Siamese model trained on ImageNet, with outstanding results. When compared to standard image classification features, the derived features showed superior generalization and a dramatically enhanced matching performance. Conventional ResNet design was explored. The experiment was successful. The features generalized well, and improved matching performance compared to image classification features. To improve the SNN model, additional architectures must be used to avoid exhaustive comparisons of feature maps, allowing the model distance layer to be used online.

Acknowledgment

This work was supported by the National Library of Medicine of the National Institute of Health (NIH) award number 5R01LM013392. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] N. Rosário-Ferreira et al., "The Treasury Chest of Text Mining: Piling Available Resources for Powerful Biomedical Text Mining," BioChem, vol. 1, no. 2, pp. 60–80, Jul. 2021, doi: 10.3390/biochem1020007.

[2] K. Hanspers, A. Riutta, M. Summer-Kutmon, and A. R. Pico, "Pathway information extracted from 25 years of pathway figures," Genome Biol, vol. 21, no. 1, p. 273, Dec. 2020, doi: 10.1186/s13059-020-02181-2.

[3] S. Kraus et al., "Literature reviews as independent studies: guidelines for academic practice," Review of Managerial Science, vol. 16, no. 8, pp. 2577–2595, Nov. 2022, doi: 10.1007/s11846-022-00588-8.

[4] J. Egger et al., "Medical deep learning—A systematic meta-review," Comput Methods Programs Biomed, vol. 221, p. 106874, Jun. 2022, doi: 10.1016/j.cmpb.2022.106874.

[5] C. G. Cess and S. D. Finley, "Representation learning for a generalized, quantitative comparison of complex model outputs," Aug. 2022.

[6] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," NPJ Comput Mater, vol. 5, no. 1, p. 83, Aug. 2019, doi: 10.1038/s41524-019-0221-0.

[7] C. Fotis, N. Meimetis, A. Sardis, and L. G. Alexopoulos, "DeepSIBA: chemical structure-based inference of biological alterations using deep learning," Mol Omics, vol. 17, no. 1, pp. 108–120, 2021, doi: 10.1039/D0MO00129E.

[8] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," BMC Med, vol. 17, no. 1, p. 195, Dec. 2019, doi: 10.1186/s12916-019-1426-2.

[9] J. M. Vaz and S. Balaji, "Convolutional neural networks (CNNs): concepts and applications in pharmacogenomics," Mol Divers, vol. 25, no. 3, pp. 1569–1584, Aug. 2021, doi: 10.1007/s11030-021-10225-3.

[10] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," Z Med Phys, vol. 29, no. 2, pp. 102–127, May 2019, doi: 10.1016/j.zemedi.2018.11.002.

[11] M. Javaid, A. Haleem, R. Pratap Singh, R. Suman, and S. Rab, "Significance of machine learning in healthcare: Features, pillars and applications," International Journal of Intelligent Networks, vol. 3, pp. 58–73, 2022, doi: 10.1016/j.ijin.2022.05.002.

[12] F. He et al., "Identifying Genes and Their Interactions from Pathway Figures and Text in Biomedical Articles," in 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, Dec. 2021, pp. 398–405. doi: 10.1109/BIBM52615.2021.9669391.

[13] R. Anders, H. Kristina, and R., P. Alexander, "Identifying Genes in Published Pathway Figure Images," bioRxiv, 2018.

[14] K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," Genome Med, vol. 13, no. 1, p. 152, Dec. 2021, doi: 10.1186/s13073-021-00968-x.

[15] F. Alharbi and A. Vakanski, "Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review," Bioengineering, vol. 10, no. 2, p. 173, Jan. 2023, doi: 10.3390/bioengineering10020173.

[16] V. Singh, S.-S. Chen, M. Singhania, B. Nanavati, A. kumar kar, and A. Gupta, "How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries–A review and research agenda," International Journal of Information Management Data Insights, vol. 2, no. 2, p. 100094, Nov. 2022, doi: 10.1016/j.jjimei.2022.100094.

[17] B. Mandal, A. Okeukwu, and Y. Theis, "Masked Face Recognition using ResNet-50," Apr. 2021.

[18] R. Zeng and M. Liao, "Developing a Multi-Layer Deep Learning Based Predictive Model to Identify DNA N4-Methylcytosine Modifications," Front Bioeng Biotechnol, vol. 8, Apr. 2020, doi: 10.3389/fbioe.2020.00274.

[19] J. Zhang and A. Zhang, "Deep learning-based multi-model approach on electron microscopy image of renal biopsy classification," BMC Nephrol, vol. 24, no. 1, p. 132, May 2023, doi: 10.1186/s12882-023-03182-6.

[20] I. P. E. Fábia, Freitas. de M. Erikson, and S. R. M. Marcella, "Person Re-Identication Using Convolutional Neural Network and Autoencoder Embedded on Frameworks Based on Siamese and Triplet Networks," Res Sq, pp. 1–21, 2020.

[21] F. Ren and S. Xue, "Intention Detection Based on Siamese Neural Network with Triplet Loss," IEEE Access, vol. 8, pp. 82242–82254, 2020, doi: 10.1109/ACCESS.2020.2991484.

[22] B. Lodhi and J. Kang, "Multipath-DenseNet: A Supervised ensemble architecture of densely connected convolutional networks," Inf Sci (N Y), vol. 482, pp. 63–72, May 2019, doi: 10.1016/j.ins.2019.01.012.