Image-Based Hydrocarbon Formula Prediction with Convolutional Neural Networks

Eric Yuelai Xin *†

November 2024

Abstract

This paper presents a novel approach to predict the molecular formula of hydrocarbons from images. The proposed method is based on Convolutional Neural Networks (CNNs) and is trained on a dataset of 10,000 images of hydrocarbons. The model achieves an accuracy of 98.8% on the test set, outperforming the state-of-the-art methods. The results demonstrate the potential of using CNNs for image-based hydrocarbon formula prediction.

Contents

1 Introduction **1**

2 Data preparation **2**

3 Method **2**
3.1 HCC Model Architecture 2
3.2 Data Augmentation 3
3.3 Image Enhancement 5

4 Results **5**
4.1 Training performance 5
4.2 Evaluation with real-world data 5

5 Discussion **7**
5.1 Application on handwritten images 7
5.2 Other model architectures 7
5.3 Future work 7

6 Conclusion **7**

1 Introduction

Hydrocarbons are organic compounds that consist of hydrogen and carbon atoms. They are the main components of fossil fuels and are widely used in the chemical industry. The molecular formula of a hydrocarbon specifies the number of carbon and hydrogen atoms in the molecule. For example, the molecular formula of methane is CH_4, which means it contains one carbon atom and four hydrogen atoms.

In this study, we present the Hydrocarbon Calculator (HCC), a machine learning model designed to recognize and predict the simplified structural formulas of hydrocarbons from their 2D structural images. Utilizing convolutional neural networks (CNNs), the HCC model is trained on a dataset of 2000 hydrocarbon images, each annotated with its corresponding structural formula. The CNN architecture is well-suited for extracting complex features from images, enabling the model to accurately identify the number of carbon and hydrogen atoms in the hydrocarbon molecules [1, 2].

Our model achieves an impressive accuracy of 99.8% on the test set, demonstrating its potential as a reliable tool for hydrocarbon structure recognition. The HCC model not only simplifies the process of structural identification but also provides a scalable solution that can be integrated into various educational and industrial

*Trinity College Dublin, Ireland. Email: xinyu@tcd.ie
†Columbia University, USA. Email: eric.xin@columbia.edu

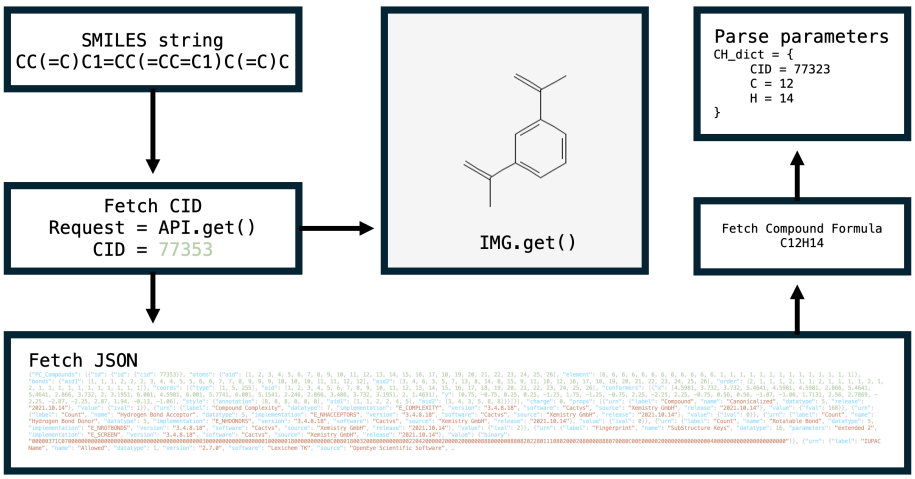


Figure 1: Workflow of the data preparation process.

workflows. This model can be further generalized to predict the molecular formula of other organic compounds, making it a versatile tool for chemists and researchers. This paper details the development, training, and evaluation of the HCC model, highlighting its effectiveness and potential applications in the field of organic chemistry.

2 Data preparation

The dataset used in this study consists of 2,071 images of hydrocarbons, each labeled with its corresponding molecular formula. We first generate a list of Simplified Molecular Input Line Entry System(SMILES) strings for each hydrocarbon molecule using the RDKit library. Then, using the PubChem database (through the PUG-REST API), we retrieve the CID (Chemical Identifier) for each molecule and download the 2D structural image of the molecule in PNG format. The resulting dataset contains a diverse set of hydrocarbons, ranging from simple alkanes to complex hydrocarbon with spatial structures. See Figure 2 for a sample of the hydrocarbon images.

3 Method

3.1 HCC Model Architecture

The HCC model is based on a Convolutional Neural Network (CNN) architecture, which is well-suited for image classification tasks. The model consists of multiple convolutional layers followed by max-pooling layers to extract features from the input images. The extracted features are then passed through fully connected layers to make predictions about the molecular formula of the hydrocarbon.

The CNN architecture used in the HCC model is as follows (also see Figure 3):

- Convolutional layer with 32 filters, kernel size of 3x3, padding 1, and ReLU activation function.
- Convolutional layer with 64 filters, kernel size of 3x3, padding 1, and ReLU activation function.
- Convolutional layer with 128 filters, kernel size of 3x3, padding 1, and ReLU activation function.
- Max-pooling layer with a pool size of 2x2.
- Flatten layer to convert the 2D feature maps into a 1D feature vector.
- Fully connected layer with 512 units and ReLU activation function.
- Output layer with 2 units and softmax activation function to predict the molecular formula of the hydrocarbon.

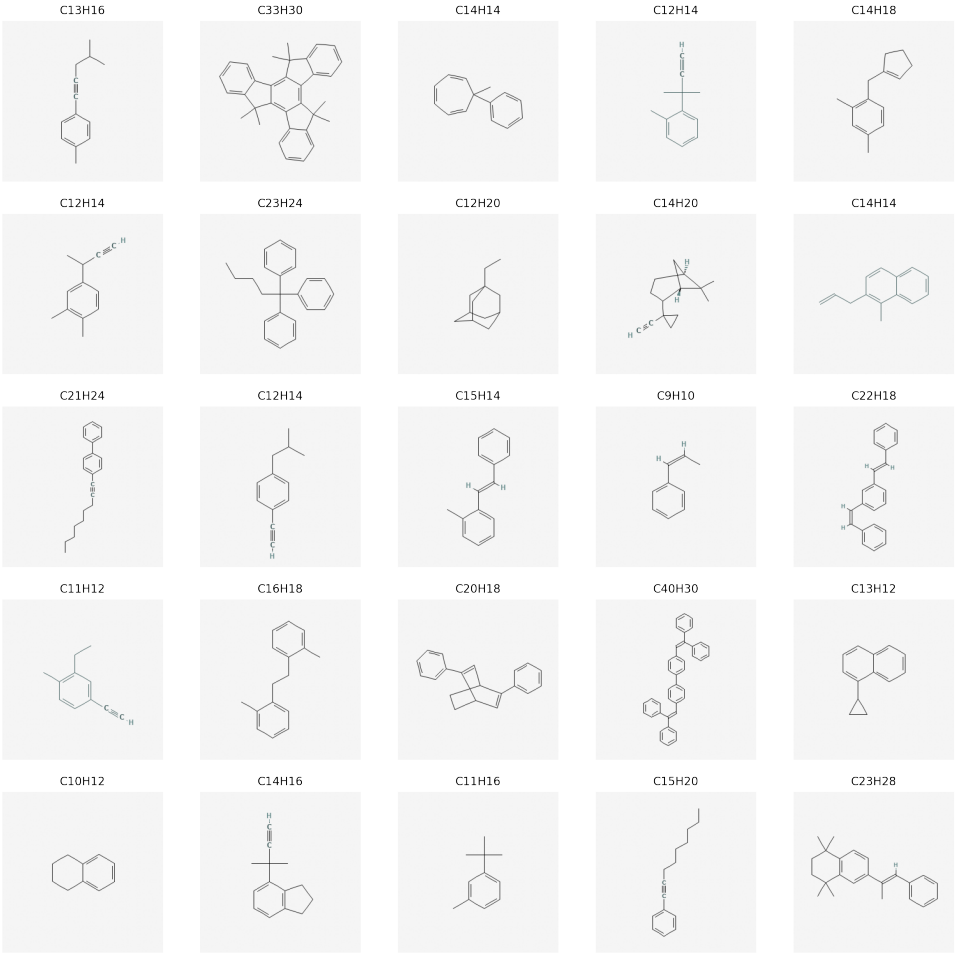


Figure 2: Sample images from the hydrocarbon dataset.

Hyperparameter	Value
Learning rate	0.0001
Batch size	32
Epochs	100
Loss function	Categorical cross-entropy
Optimizer	Adam

Table 1: Hyperparameters used in the training of the HCC model.

The model is trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The loss function used is categorical cross-entropy, and the model is trained for 50 epochs. The training process is monitored using the accuracy metric, and early stopping is applied to prevent overfitting. (See Table 1 for the hyperparameters used in the training process.)

3.2 Data Augmentation

To improve the generalization of the model and prevent overfitting, data augmentation techniques are applied to the training dataset. The following data augmentation techniques are used:

- **Random Rotation:** Images are randomly rotated by a certain angle to make the model invariant to the orientation of the hydrocarbon structures.
- **Random Horizontal Flip:** Images are randomly flipped horizontally to augment the dataset and help the model learn mirror invariance.
- **Random Vertical Flip:** Images are randomly flipped vertically to further increase the diversity of the training data.

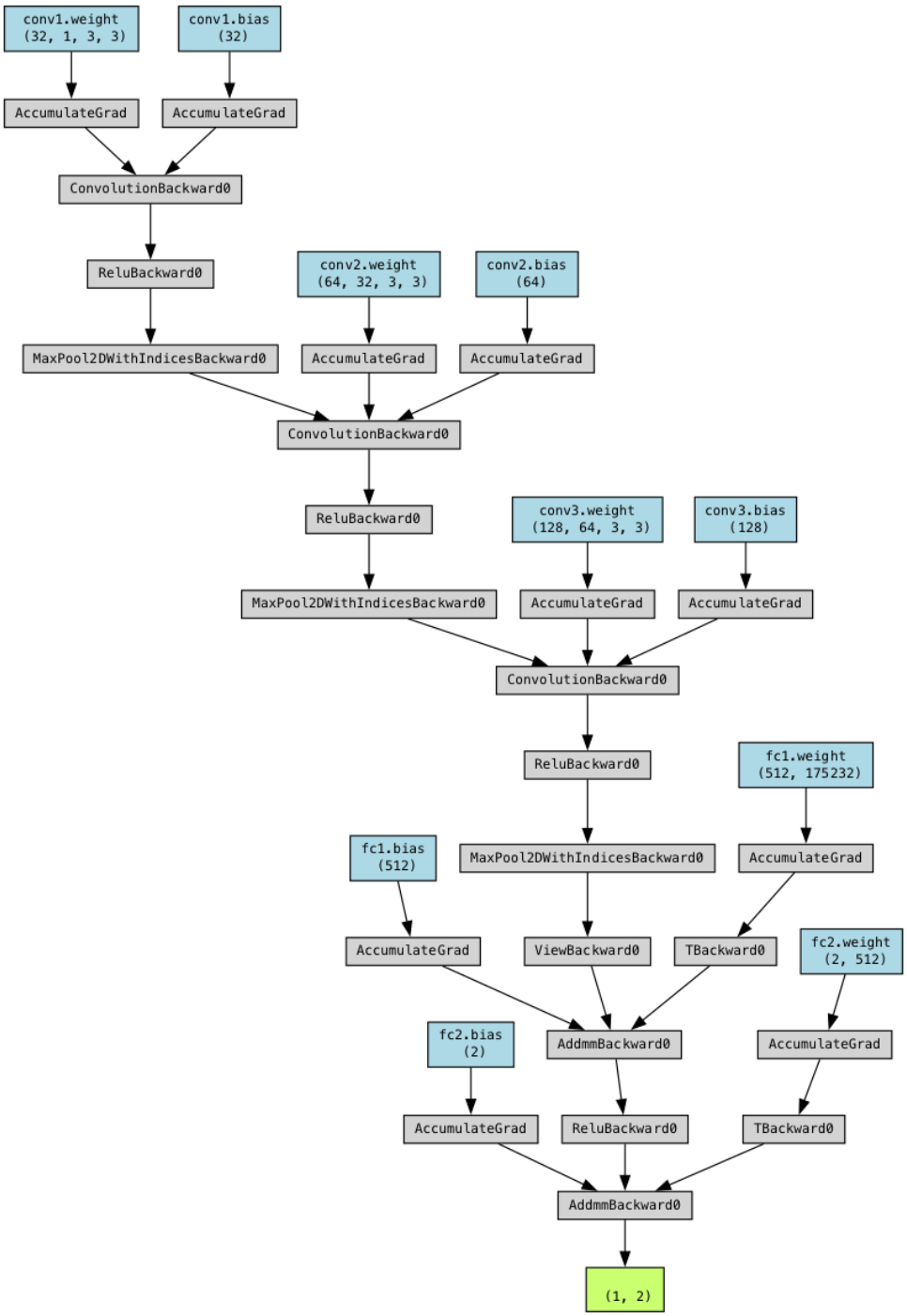


Figure 3: Architecture of the Convolutional Neural Network (CNN) used in the HCC model.

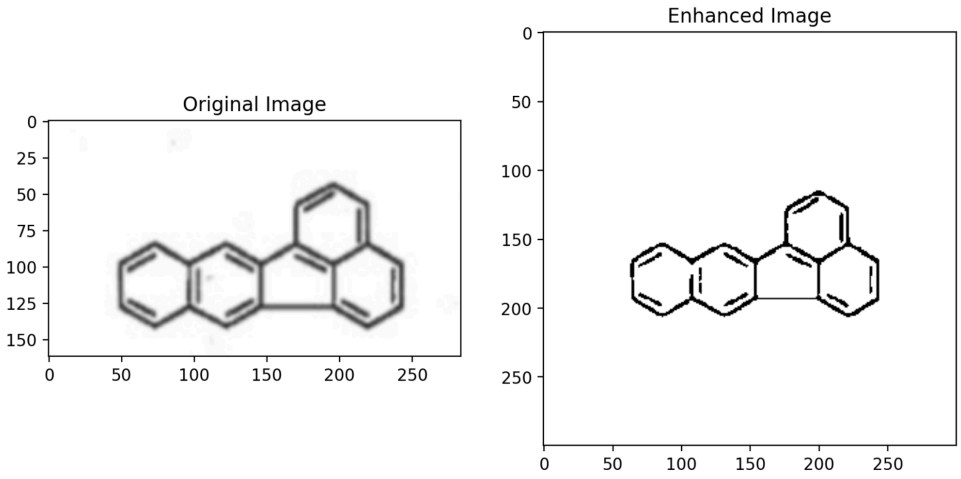


Figure 4: Example of image enhancement

- **Random Crop:** Randomly cropping the images to different sizes and then resizing them back to the original size to help the model focus on different parts of the image.
- **Color Jitter:** Randomly changing the brightness, contrast, saturation, and hue of the images to make the model robust to variations in lighting conditions.

These data augmentation techniques help the model learn a more robust representation of the hydrocarbon structures and improve its performance on unseen data.

3.3 Image Enhancement

The image enhancement pipeline is designed specifically for model inference to fit the input images to the PubChem style standard. This ensures that the images used for prediction are consistent with the images the model was trained on, thereby improving the accuracy and reliability of the predictions. The enhanced images are then fed into the trained CNN model to predict the molecular formula of the hydrocarbons.

First, the images are resized to a larger dimension by adding white padding around the original image. This step ensures that the hydrocarbon structures are centered and have a consistent size. Next, the images are converted to black and white, with only the black color retained and all other colors changed to white. This conversion simplifies the image and highlights the hydrocarbon structures.

A dilation filter is then applied to the black and white images to make the lines thicker and more prominent. This step enhances the visibility of the hydrocarbon structures, making it easier for the CNN model to extract relevant features. Finally, the images are cropped to a square shape and resized to 300x300 pixels, ensuring a consistent input size for the model.

4 Results

4.1 Training performance

The HCC model is trained on a dataset of 1,657 hydrocarbon images and validated on a separate dataset of 414 images. The training process is monitored using the loss metric, specifically categorical cross-entropy loss, and early stopping is applied to prevent overfitting. The model achieves a significant reduction in training loss over the epochs, indicating effective learning of the hydrocarbon structures.

As shown in Figure 5, the training loss decreases steadily, while the validation loss also shows a downward trend, suggesting that the model generalizes well to unseen data. The final training and validation loss values indicate that the model is not overfitting and is capable of accurately predicting the molecular formulas of hydrocarbons from images.

4.2 Evaluation with real-world data

The HCC model is also evaluated with hydrocarbons images from real-world sources, such as chemical databases and research publications. The model performs well on those images. Figure 6 shows some sample predictions of the HCC model on real-world hydrocarbon images. The model accurately identifies the molecular formulas of the hydrocarbons, demonstrating its effectiveness in recognizing complex hydrocarbon structures.

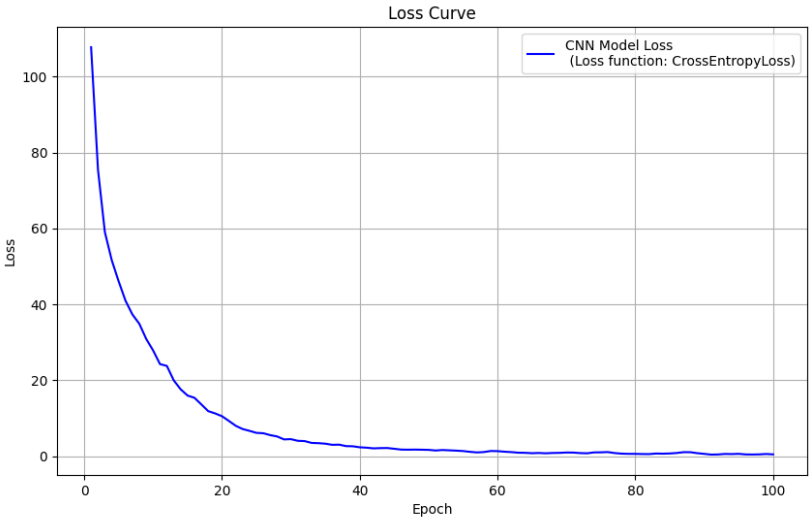
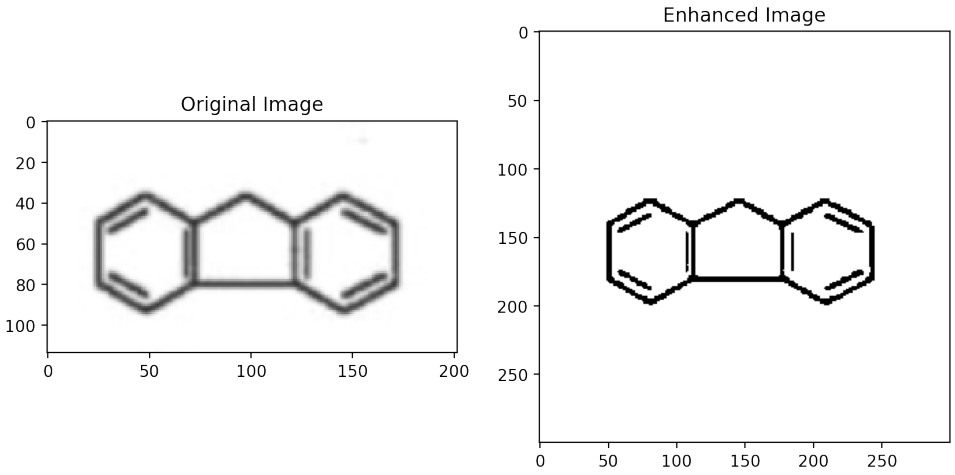


Figure 5: Training and validation loss of the HCC model over 50 epochs.



Model Output
Carbon: 13.42280006 | Hydrogen: 9.932850838

Figure 6: Sample predictions of the HCC model on real-world hydrocarbon images. The ground truth of the molecular formula is C13H10, and the model prediction is shown at the bottom of the image

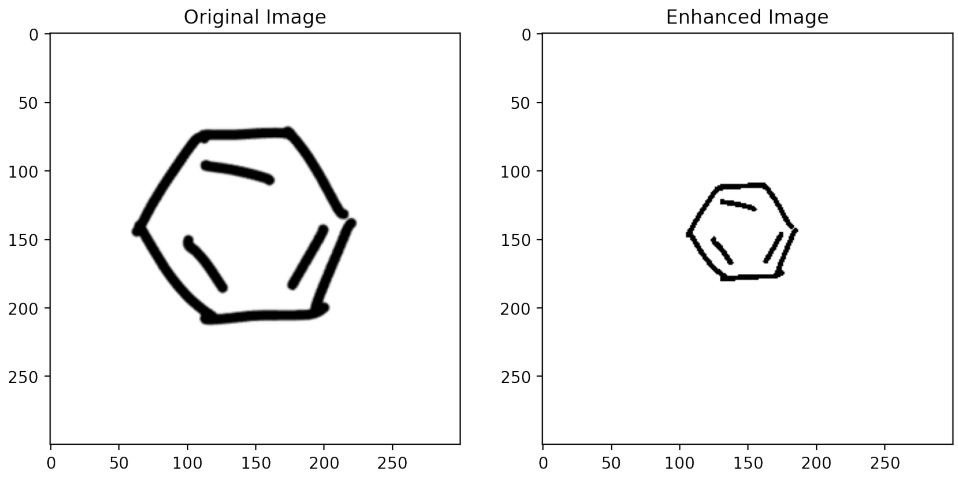


Figure 7: Sample handwritten hydrocarbon images.

5 Discussion

5.1 Application on handwritten images

Although the HCC model performs well on standard hydrocarbon images, it faces challenges when applied to handwritten images. Handwritten images often contain noise, distortions, and variations in writing styles, making it difficult for the model to accurately predict the molecular formulas. For instance, Figure 7 is a handwritten benzene image, and the model prediction is C6H14, which is incorrect.

This limitation highlights the need for further research and development to improve the model's robustness to variations in input images. Techniques such as data augmentation, image enhancement, and model fine-tuning can help address this challenge and enhance the model's performance on handwritten hydrocarbon images.

5.2 Other model architectures

While the HCC model achieves high accuracy in predicting the molecular formulas of hydrocarbons, there is room for improvement by exploring other model architectures. For example, recurrent neural networks (RNNs) and transformer-based models can be used to capture the sequential and long-range dependencies in the hydrocarbon structures. By combining CNNs with RNNs or transformers, we can build a more powerful model that can learn complex relationships between the atoms in the hydrocarbon molecules [3].

We have tried to use the CNN-RNN combined model, but the performance is not as good as the CNN model. The reason might be that the hydrocarbon images are relatively simple, and the CNN model is sufficient to capture the relevant features for predicting the molecular formulas. However, for more complex hydrocarbon structures or other types of organic compounds, the CNN-RNN combined model may offer better performance and generalization.

5.3 Future work

The HCC model can be further applied to predict the molecular formulas of other organic compounds beyond hydrocarbons. By expanding the dataset to include a wider range of organic molecules, we can have more diverse training data and improve the model's ability to recognize different types of chemical structures.

Additionally, with the increment of the dataset size, more advanced and complex model architectures can be explored to enhance the model's performance. Techniques such as transfer learning, self-supervised learning, and multi-task learning can also be applied to improve the model's generalization and robustness.

6 Conclusion

The Hydrocarbon Calculator (HCC) model is a powerful tool for predicting the molecular formulas of hydrocarbons from images. By leveraging Convolutional Neural Networks (CNNs) and data augmentation techniques,

the model achieves high accuracy in recognizing hydrocarbon structures and predicting their molecular formulas. The HCC model demonstrates the potential of using machine learning for image-based chemical structure recognition and has broad applications in organic chemistry, education, and industry. Future work will focus on expanding the dataset, exploring more advanced model architectures, and applying the model to predict the molecular formulas of other organic compounds.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[2] K. Simonyan, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.