Project 2 - Dog Breeds Classification

Thai Boonchai (6322790138) Preravitch Siripanich (6322773761)

Arnuparp Cheammarerng (6322770346) Phanuwich Thepnok (6322672730)

Department of Information, Computer and Communication Technology (ICT), Sirindhorn International Institute of Technology (SIIT), Thammasat University, Khlong Luang, Pathum Thani, Thailand

1. Team and Contribution

Our group comprises four dedicated individuals. Which are Thai Boonchai, Preravitch Siripanich, Arnuparp Cheammarerng, and Phanuwich Thepnok. each bringing their unique skills and contributions to the table. Together, we form a dynamic team that excels in various aspects of our project. By the first person is Thai Boonchai, Thai is responsible for researching an information and working on project report. His ability to summarize information and write a report in term of an academic tone helps us to make this project smoothly. For Preravitch and Arnuparp, is responsible for building a Convolutional Neural Networks (CNN) model. Their ability to analyze and build a model helps to make this project successful. And lastly, Phanuwich is responsible for design decisions about the user experience for your app and summarizing the report. His skill makes our project to be done in the most perfect way.

2. Introduction and data analysis

[1] Dogs, often referred to as man's best friend, hold a special place in the hearts of people all over the world. They are more than just pets because of their unwavering loyalty to their owners; they are cherished members of the family. However, a disheartening trend has emerged in today's world. Many people fall victim to unscrupulous people who sell dogs of one breed as dogs of another, often at exorbitant prices. The intricate relationships between subordinated breeds, as well as the significant variations found within each breed, necessitate fine-grained dog breed classification. Furthermore, dog owners and veterinarians frequently require a diverse collection of dog images for a variety of reasons, ranging from assisting in the search for lost dogs to meeting medical requirements.

[2] There is a promising solution to these challenges in the age of Artificial Intelligence and Machine Learning. [3] – [5] These days, machine learning is widely used. Convolutional neural networks (CNN) are one of the most significant achievements of machine learning and deep learning [6]. CNN is made up of various neurons with varying learning biases and weights. In this, each neuron receives some inputs, performs some dot product or mathematical manipulation, and then outputs some desired outputs. CNN architecture is distinct from that of other neural networks. CNN's architecture aids in the reduction of features and the creation of a defined network to improve accuracy. In this project, we will use our algorithm to identify an estimate of the canine's breed with a dog's detector and human's detector. Using Convolutional Neural Networks (CNN) models for classification.

Dataset

The dataset that we got is the dog and human dataset. According to the dog dataset we can see that there are more than 133 dog breeds. Which include Training Set, Validation Set, and Test Set. All images that we got are all in a JPG file (.jpg). And for human dataset, the whole store in a folder that has a name on it and in the folder there is 1 picture that contains a human picture of that person. There are 5479 human images and 8,351 dog images used in the experiment.

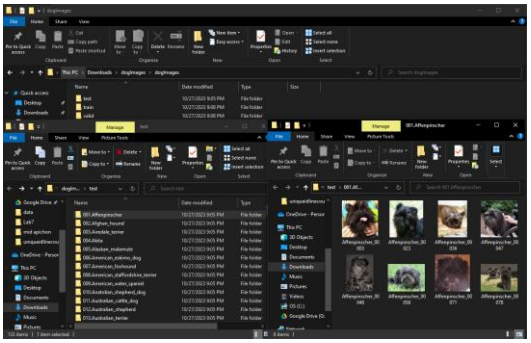

Figure 1 Dog Dataset

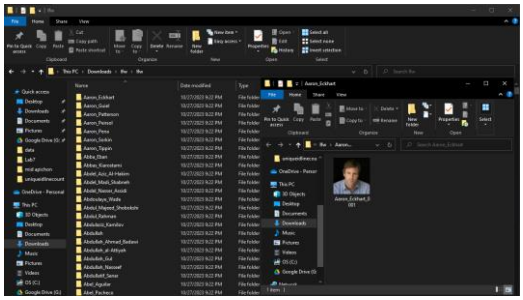

Figure 2 Human Dataset

However, some dog breeds are crossbred. This may make classification difficult. This is due to the fact that facial traits, such as the eyes, ears, and nose, may be comparable to those of the purebred breed.

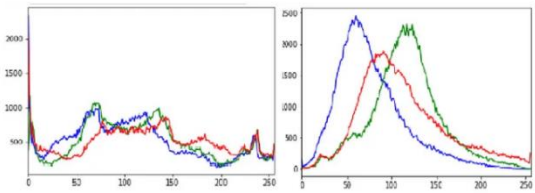

Figure 3 The graph above also shows that the training data for each breed are slightly imbalance.

3. Algorithm design

In this section, we present a detailed algorithm design for identify an estimate of the canine's breed with a dog's detector and human's detector using Convolutional Neural Networks (CNN).This algorithm aims to provide an accurate estimate of canine's breed and human name base on the given dataset.

```
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_12 (Conv2D)           (None, 222, 222, 32)      896

max_pooling2d_12 (MaxPooli   (None, 74, 74, 32)        0
ng2D)

conv2d_13 (Conv2D)           (None, 72, 72, 64)        18496

max_pooling2d_13 (MaxPooli   (None, 24, 24, 64)        0
ng2D)

conv2d_14 (Conv2D)           (None, 22, 22, 128)       73856

max_pooling2d_14 (MaxPooli   (None, 7, 7, 128)         0
ng2D)

conv2d_15 (Conv2D)           (None, 5, 5, 128)         147584

max_pooling2d_15 (MaxPooli   (None, 1, 1, 128)         0
ng2D)

flatten_3 (Flatten)          (None, 128)               0
...
Total params: 375109 (1.43 MB)
Trainable params: 375109 (1.43 MB)
Non-trainable params: 0 (0.00 Byte)
```

Figure 4 Layers

A. Convolutional Layers

The convolution layer is mandatory since it is the core building block of the CNN algorithm. Convolutional layers perform the extraction of features from input images by enforcing a local connectivity pattern between neurons in adjacent layers.

Number of Convolutional Layers

We decided to use four convolutional layers due to the complexity of the task. Dog breed classification is inherently complex, as breeds often share similar features, and there can be samples or dogs that contain mixed breeds. Multiple convolutional layers allow the model to learn convoluted and hierarchical features.

Depth of Convolutional Layers

The layers increase in depth (16, 32, 64, 128) to enable the network to capture more abstract and complex features.

B. Pooling Layers

[7] [8] Pooling layers are used between two convolutional layers to reduce the spatial volume of the input image after convolution (down sampling).

Number of Pooling Layers

Four max pooling layers were used to downsample the feature maps and reduce computational complexity. The

spatial dimensions are reduced after each convolutional layer, making the model more computationally efficient.

Global Average Pooling Layer

The global average pooling layer was introduced after the convolutional layers to reduce the spatial dimensions and extract global information. This simplifies the network architecture, reduces the number of parameters, and helps control overfitting.

C. FC (Fully-Connected) Layer

The flattening layer allows the transition from convolutional layers to fully connected layers. This is done by taking the output of the last pooling layer and converting it into a flat vector, which is the input for the fully connected layers. FC layers are basically dense layers after the flattening layer.

Number of Dense Layers

Three fully connected dense layers were added after global average pooling. This choice follows the common architecture pattern used in CNNs for image classification. The final dense layer outputs probabilities for each dog breed class.

The first and second dense layer has 512 units with a ReLU activation function. A 512-unit layer is deep enough to capture complex patterns but not too deep to risk overfitting.

Dropout Layer

However, in cases of overfitting, two dropout layers at a rate of 0.5 were added before the next dense layer and before global average pooling to mitigate overfitting.

Output Layer

The output layer has 133 units, corresponding to the number of dog breed classes in the dataset. The activation function used was SoftMax since it is suitable for the classification method of various multi-classes.

Rationale for Parameter Selection:

The selection of the convolutional and max pooling layers is influenced by the task's complexity and the need to capture intricate features.

The global average pooling layer simplifies the network, which is particularly beneficial when dealing with limited data and to control overfitting.

The choice of two fully connected dense layers is based on common practices for image classification tasks, striking a balance between model complexity and computational efficiency.

The dropout layer is introduced to reduce overfitting by preventing the network from relying too heavily on specific neurons during training.

The number of units in the output layer aligns with the number of dog breed classes in the dataset.

Justification:

The selected architecture aligns with well-established practices in deep learning for image classification tasks.

The model parameters and architecture have been chosen to strike a balance between capturing complex patterns, preventing overfitting, and being computationally efficient.

The selection of these parameters and design decisions is based on the characteristics of the dog breed dataset, which requires the model to learn intricate features.

The design offers a good trade-off between model complexity and performance, and it is supported by empirical success in similar image classification tasks in the literature.

4. Evaluation

Dog breeds are highly complex, and it is a challenging task for classification. When looking into Dog breeds classification it requires a very deep understanding of CNN algorithms, as well as a deep architecture built based on CNN model. The deeper the architecture, the greater the accuracy of the classification for each breed.

CNN model	Accuracy
Our model	11%
VGG16	71%
ResNet50	81%
Xception	85%

Our algorithm completely lacks the depth for training the model compared to other CNN benchmarks models (VGG16, ResNet50, Xception) , even if we perform data augmentation. Our model was outperformed in accuracy by the other benchmarks due to it having more depths of the CNN network architecture/more convolutional layers (from 16 to 71 layers). Aside from the high number of weighted-layers, there are also benchmarks which use global average pooling and depth-wise separable convolutional layers.

Our CNN

Strengths:

Simplicity: The architecture is relatively simple, making it easy to understand and implement, especially for educational or prototype purposes.

Convolutional Layers: The use of convolutional layers allows the network to capture hierarchical features in images, which is crucial for image classification tasks.

Max Pooling: Max-pooling layers help reduce the spatial dimensions of the data, decreasing the computational burden and helping prevent overfitting.

Dropout: The addition of dropout layers helps reduce overfitting by randomly deactivating a fraction of neurons during training.

Global Average Pooling: This layer reduces the spatial dimensions of the data to a single vector, which can be used for classification. It also reduces the number of parameters in the model.

Weaknesses:

Shallow Architecture: The network is relatively shallow, which may limit its ability to capture complex features in images compared to deeper architectures like ResNet or Inception.

Limited Capacity: The model may struggle to handle complex image classification tasks that require a high degree of feature abstraction and may not achieve state-of-the-art performance.

Parameter Efficiency: While simplicity is an advantage, the model may not be parameter-efficient. It lacks advanced architectural features like residual connections (as in ResNet) or depth-wise separable convolutions (as in Xception) to improve parameter efficiency.

Overfitting: Although dropout layers are included to mitigate overfitting, the network may still be susceptible to overfitting on smaller datasets without further regularization techniques.

Architecture May Not Be Optimized: The architecture used is a relatively standard design and may not be fine-tuned for a specific task. Fine-tuning or using more specialized architectures might yield better results for specific applications.

Criticism & Improvement

We could use more data since the data for dog breed classification is slightly imbalanced. It's possible to enhance the model with transfer learning but that may cost way too much computational cost. The model of our algorithm would improve in accuracy if we would add more layers/create a deeper CNN for classification. The model could also use depth-wise convolutional layers with different level extraction.

Benchmarks - Transfer Learning

VGG16

[10] The simplest CNN algorithm is the VGG16, it is a very deep convolutional network used for large scale image recognition. As the number in its name implies, VGG16 consists of 16 weighted-layers with a single size of 3x3. Reducing volume size is handled by max pooling and two fully-connected layers, each with 4,096 nodes are then followed by a SoftMax classifier.

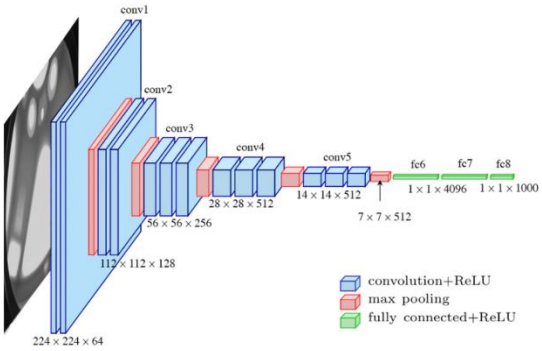

Figure 5 VGG16

VGG16 also uses 'pre-training' which trains a small version of the network and later converges those small networks into a deeper network.

Strengths:

We still use VGG16 despite it being created in 2014, for a smaller network architecture.

Weaknesses:

There are two major drawbacks to using VGG16:

1. It is painfully slow to train.
2. The network architecture weights themselves are quite large (in terms of disk/bandwidth - 533 MB).

ResNet50

ResNet50 is referred to as CNN Architecture with micro-architectures or 'sets of building blocks' which leads to Macro-architecture (The end or final network). The model uses a normal convolutional layer, max pooling layer and global average pooling layer.

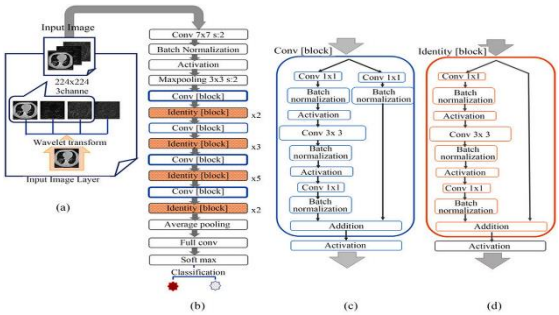

Figure 6 ResNet50

ResNet50 is much deeper compared to VGG16 since its implementation contains 50 weighted-layers, but the model is substantially smaller than VGG16 due to the use of global average pooling (reducing the sizes to about 102 MB). ResNet50 uses residual modules with standard SGD for loss which can train extremely deep networks while avoiding the vanishing gradient problem.

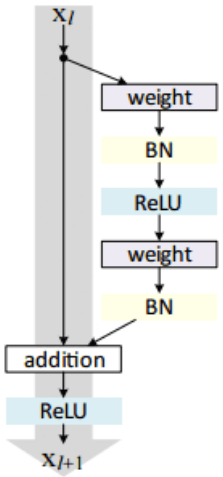

Figure 7 residual building blocks

Strengths

Deeper Architecture

ResNet-50 is significantly deeper than VGG16, with 50 layers compared to VGG16's 16 layers. We have found out that this depth allows ResNet-50 to capture more complex and fine-grained features in images, making it well-suited for tasks that require a high level of detail and abstraction.

Residual Connections

The use of residual connections in ResNet-50 helps mitigate the vanishing gradient problem, making it easier to train very deep networks.

Weaknesses

Computational Complexity

The depth of ResNet-50 can make it computationally expensive and memory-intensive, especially during training. This can be a limitation on hardware with limited resources.

Overfitting

If the dataset is uniform and not large enough, ResNet50 could be prone to overfitting since the model is very deep. Regularization techniques can often solve this issue.

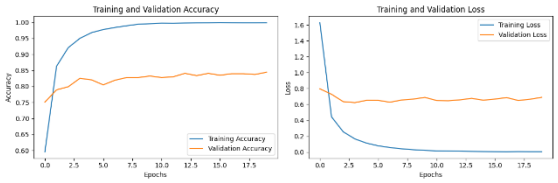

Figure 8 Overfitting Data

Xception

Xception is an architecture extension of Inception, where both architectures used 'multi-level extraction' by computing 1x1, 3x3 and 5x5 with the same module of the network. The output of these filters is then concatenated before being sent to the next layer of the network.

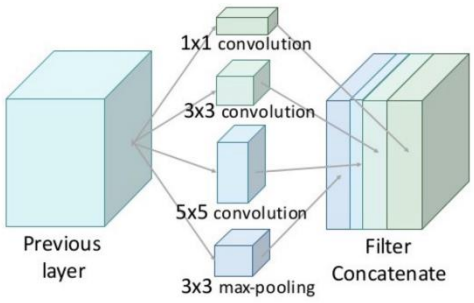

Figure 9 Xception

The distinct addition of Xception from Inception is that the model would use the depth-wise separable convolution layer, which can reduce the number of computations and parameters.

Strengths

Extremely Deep Architecture

Xception consists of 71 weighted-layers which is 21 more than ResNet50. While the architecture is deeper than ResNet50 it is still computationally effective and efficient due to using both depth-wise separable convolutional layers (which reduce the number of computations and parameters) and spatial convolutional layers.

Weaknesses

Complexity of implementation

Implementing the model for Xception requires greater knowledge of Inception and convolutional layers, since the model uses depth-wise separable convolutional layers which may be complex when compared to other benchmarks (VGG16) and our algorithms.

Criticism & Improvement for most Benchmarks

It would be best for each benchmark to have additional Data. Using transfer learning, we could create a custom CNN classifier on top of each benchmark for greater accuracy and at the same time add dropout to prevent overfitting.

Further improvement

Data Augmentation

- We perform data augmentation for a variety of transformations and distortions to be applied to the original image. The purpose of image augmentation is to increase the diversity of the training dataset and improve the accuracy of classification.

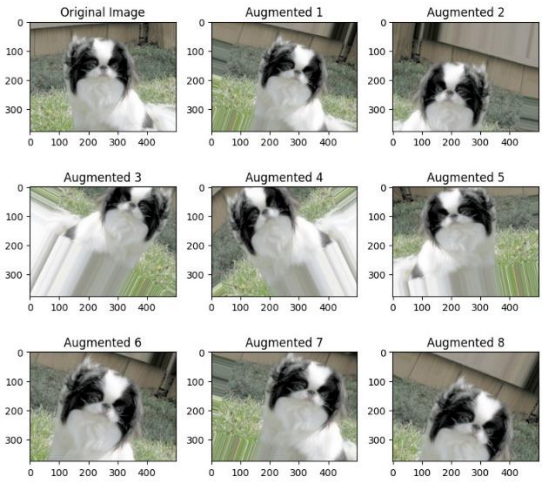

Figure 10 Augmented Images example

- Augmented Settings Used
- Rotation range = 40,
- Width shift range=0.2,
- Height shift range=0.2,
- Shear range=0.2,
- Zoom range=0.2
- We have found an increase of 3% when using these settings for augmentation.

OpenCV DNN Module

By using another face detection method, OpenCV DNN Module with 'deploy.prototxt' and 'res10_300x300_ssd_iter_140000.caffemodel' gave off a super accurate human detection rate when compared to just detecting face cascade. The reason for this is the OpenCV DNN module was used with a pre-trained face detection model for better human detection accuracy.

Web application

Transform algorithm into web application using react as frontend and flask as backend and the results are following:

Figure 11 Result 1

Figure 12 Result 2

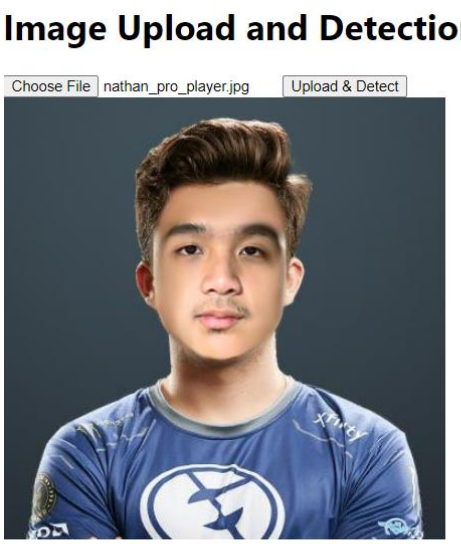

Figure 13 Result 3

5. Conclusion

In this project, our team has embarked on a challenging endeavor within the realm of image classification, specifically focusing on fine-grained dog breed classification and human identification through Convolutional Neural Networks (CNNs). This project required the development of a sophisticated algorithm design incorporating many levels and components of the CNN architecture.

The algorithm developed exhibits several strengths that make it suitable for educational and prototyping purposes. Its simplicity and use of important aspects, such as convolutional layers, max pooling, dropout layers, and global average pooling, allow it to capture hierarchical information while reducing overfitting.

Nevertheless, certain weaknesses should be acknowledged. The architecture's relative shallowness may limit its effectiveness in handling highly complex image classification tasks, and it may not be as parameter-efficient as more advanced architectures such as ResNet or Xception. Overfitting is still a possibility, especially with smaller datasets.

In our evaluation, we have compared our custom CNN architecture to benchmark models, including VGG16, ResNet50, and Xception. Each of these benchmarks has its own set of advantages and disadvantages. VGG16, for example, while simple, suffers from delayed training and large disk space needs. ResNet50's depth enables it to capture fine-grained characteristics, albeit at the expense of computational intensity. Xception maintains a balance between complexity and efficacy, but with a more elaborate implementation, thanks to its deep architecture and efficient design.

Moving forward, there are various ways to improve. Fine-tuning the bespoke architecture, experimenting with other regularization approaches, and investigating alternative cutting-edge architectures are all possible possibilities. Furthermore, assessing the model's performance on a broader and more diverse data set would offer useful insights into its capabilities and limits.

In summary, this study demonstrates a noteworthy attempt to employ CNNs to solve the difficult challenge of dog breed categorization and human identification. It provides a full insight into our bespoke architecture's strengths and shortcomings in comparison to recognized benchmark models. This knowledge establishes the framework for future CNN model upgrades and applications.

6. References

[1] Kristina Kledzik, “Why Are Dogs Called ‘Man’s Best Friend’?”, https://www.rover.com/blog/dogs-called-mans-best-friend/

[2] B. Kumar Shah, A. Kumar, A. Kumar, "Face Recognition Based Dog Breed Classification Using Coarse-to-Fine Concept and PCA," in Proceedings of the 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. [page numbers] | Publisher: IEEE.

[3] S. Thapa, P. Singh, D. K. Jain, N. Bharill, A. Gupta and M. Prasad, "Data-Driven Approach based on Feature Selection Technique for Early Diagnosis of Alzheimer's

Disease", 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2020.

[4] S. Adhikari, S. Thapa and B. K. Shah, "Oversampling based Classifiers for Categorization of Radar Returns from the Ionosphere", 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 975-978, 2020.

[5] S. Thapa, S. Adhikari, A. Ghimire and A. Aditya, "10.1109/ICESC48915.2020.9155833", 2020 8th R10 Humanitarian Technology Conference (R10-HTC), 2020.

[6] A. Ghimire, S. Thapa, A. K. Jha, A. Kumar, A. Kumar, and S. Adhikari, "AI and IoT Solutions for Tackling COVID-19 Pandemic", 2020 International Conference on Electronics Communication and Aerospace Technology, 2020.

[7] C. Marín Navas, F. J. Navas González, V. Castillo López, L. Payeras Capellà, M. Gómez Fernández and J. V. Delgado Bermejo, "Impact of breeding for coat and spotting patterns on the population structure and genetic diversity of an islander endangered dog breed", Res. Vet. Sci, vol. 131, no. April, pp. 117-130, 2020.

[8] R. Kumar, M. Sharma, K. Dhawale and G. Singal, "Identification of Dog Breeds Using Deep Learning", Proc. 2019 IEEE 9th Int. Conf. Adv. Comput. IACC 2019, pp. 193-198, 2019.

[9] M. V. S. Rishita and T. A. Harris, "Dog breed classifier using convolutional neural networks", 2018 Int. Conf. Networking Embed. Wirel. Syst. ICNEWS 2018 - Proc, 2018.

[10] Tao, J., Gu, Y., Sun, J., Bie, Y., & Wang, H. (2021). Research on VGG16 Convolutional Neural Network Feature Classification Algorithm Based on Transfer Learning. In 2021 2nd China International SAR Symposium (CISS) (pp. TBD). IEEE.

[11] Sharma, A., Zehra, A., Das, A., Rastogi, K., Agarwal, M., Mascarenhas, S., J, J., M, V., & Deepa, S. (2023). Brain Tumor Classification: A Comparison Study CNN, VGG 16 and ResNet50 Model. In Proceedings of the 2023 International Conference on Data Science and Network Security (ICDSNS). IEEE.