Capstone project - Dog breed classifier

Riccardo Mattia Galli

Udacity data scientist nanodegree

1 Definition

In this section we outline the project domain background and the problem which is posit in. Then we explain how we developed the proposed solution and which metrics will be used to evaluate it.

1.1 Project overview

This project aims to solve a canonical supervised machine learning classification task: Image classification. In a supervised learning context, image classification algorithms extrapolate information from a labelled dataset in order to learn how to appropriately classify new input data. A machine learning algorithm usually contains the following components:

1. A loss function, which measures how close the algorithm performance is from its objective. In a classification setting the loss function, for example, can describe the number of misclassified observations. The goal of the algorithm is indeed to adjust its parameters in order minimize loss after each iteration.
2. A model architecture, which defines the behaviour of the algorithm, its parameters and its structure.
3. An optimizer, which determines how a given model should update its parameters to minimize the loss function and therefore learn to perform better.[10].

In recent years, huge improvements in image classification tasks has been achieved by artificial neural networks (ANN). These models represent a mathematical model of the biological neural networks. Each node of this network (a neuron) is responsible to apply computations to parts of the input data. An ANN is divided in layers of nodes or neurons, where each layer applies specific operations to the received input and then passes the output as input for the next layer. There are three different types of layer categories:

1. Input layer, which is responsible of processing the input data.
2. Hidden layer(s), which represent the layer(s) between input and output layer. The hidden layer(s) are devoted to extracting and increase the feature space from the input layer, thus generating new parameters and retrieving ulterior useful information for the classifier task.
3. Output layer, which generates the final output (scalar or probability) based on the previous (hidden) layers computations.

Deep learning is a machine learning subfield which works with ANNs formed by multiple hidden layers. in this field, convolutional neural networks (CNNs) represent the state of the art in image classification tasks [4]. Classical ANN architectures, called multi-layers perceptrons (MLPs), have usually only dense layers, where each node in one layer is fully connected with all the nodes of the next layer, as show by figure 1. Instead, CNNs analyze the input with convolu-

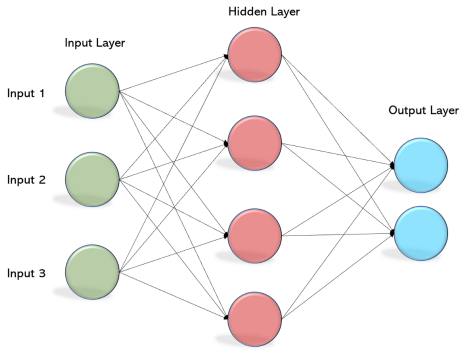

Fig. 1. Multilayer perceptron example, image taken from this Medium article.

tional kernels mimicking the biological visual receptive field [6]. This technique has the advantage of considering the spatial relationship in the input as information. In image classification, for example, pixels that are close probably represent the same object and this information is lost in the MLP architecture due to considering each input part as independent. A second advantage is that CNNs, despite having many layers, produce less parameters to train, since they do not have fully connected layers. This results in CNNs training faster than MLPs when considering the same classification task. Overall, these are some of the reasons that make CNNs the top performing model architectures in image classification task, such as on the ImageNet database [3]. A representation of a CNN architecture applied to images classification is shown by figure 2.

1.2 Problem statement

The goal of the project is to build a machine learning model which can:

1. Effectively predict dog breeds from dog images.
2. When a human image is provided, predict the most resembling dog breed.
3. Ignore pictures of other classes.

1.3 Metrics

The model is evaluated following accuracy. In a classification task, accuracy represent the percentage of correctly classified images over all the set images

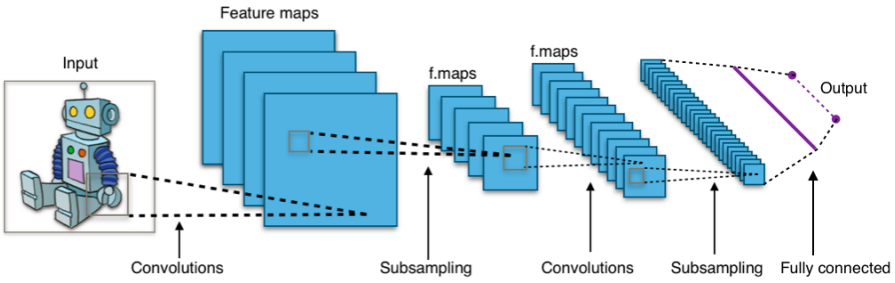

Fig. 2. Convolutional neural network example, image taken from Wikimedia.

(see equation 1). We decided to rely on accuracy since the output classes of dog breeds are decently balanced and therefore this metric should work fine (mean N of images per class: 50, min: 26, max: 77). At the same time, there is no huge difference between the human faces and dogs dataset sizes.

$$\frac{\text{N correctly classified items}}{\text{N total classified items}} \tag{1}$$

In order to provide an unbiased evaluation metric, the model accuracy is computed on the model performance over the test set. The test set contains images which have not been used to train or validate the model.

2 Analysis

In this section, we describe the dataset structure and the techniques used to solve the image classification task.

2.1 Data exploration

Two main datasets, provided by Udacity, are used to complete this project. Both of these datasets are a collection image files:

1. The first dataset corresponds to 13233 RGB input files of human faces images. The images are standardized and have both height and width of 250 pixels.
2. The second dataset corresponds to 8351 RGB input files of dog images. The images width and height span from 113 to 4003 and from 112 to 4278 pixels respectively.

A third custom dataset is used to test the final algorithm. The custom dataset contains 6 images, 2 representing human faces, 2 representing dogs and 2 representing objects of different sizes with RGB channels.

2.2 Exploratory visualization

Visualizing examples from the two main datasets can help understanding the type of images used to train and evaluate the machine learning algorithms. Figure 3 shows three images of the human faces dataset, while three images of the dog dataset can be observed in figure 4.

Fig. 3. Images from the human faces dataset.

Fig. 4. Images from the dog dataset.

2.3 Algorithms and techniques

The project is a combination of four different machine learning models, in order:

1. Face detector, which is an OpenCV implementation of a Haar feature-based cascade classifier [1]. It detects whether a human face is present in a given picture and therefore, whether a human is present in an image.
2. ResNet50 classifier, This is a pre-trained CNN that classifies input images into a 1000 different classes, using weights obtained from the Imagenet recognition challenge [7]. 118 of the 1000 classes describe different dog breeds. The final algorithm relies on this model to predict whether a dog is present in the image.

3. Custom CNN, which is a CNN developed from scratch with Keras [5] for classifying different dog breeds. This model is trained on the dog dataset described in subsection 2.1, which is divided into training, validation and test set that respectively serve to learn the model parameters, extract the best model and test the model's accuracy.
4. Transfer learning model: This is a more accurate model than the custom one in the same task. It exploits the pre-trained architecture of the ResNet50 [7] to predict dog breeds by adapting it for out dataset.

2.4 Benchmark

Benchmarks have been established by Udacity regarding the models accuracy score, metric discussed in subsection 1.3. The custom CNN model should achieve an accuracy greater than 1% on the test set. The transfer learning model should achieve, on the test set, an accuracy greater than 60%.

3 Methodology

In this section we explain the steps performed to prepare the input data and implement the algorithms described in subsection 2.3. Keras is the main library used to implement all the models and prepare the input data [5].

3.1 Data Preprocessing

Data preprocessing is applied to the input data in order to increase the quality and stability of the prediction during training. The face detector requires input images to be converted to gray scale images. ResNet50 input images require preprocessing as well. Input images are therefore resized, converted from 3D to 4D (n samples, rows, columns, channels), reordered in the color channel from RGB to BGR and lastly normalized according to fixed mean vector, as described by equation 2.

$$normalization = \Big\{ mean = [103.393, 116.779, 123.68] \tag{2}$$

The custom CNN and transfer learning model use the same data preprocessing pipeline for the ResNet50 [7] to transform validation and test dataset.

3.2 Implementation

The implementation mostly concerns the custom CNN, the ResNet-50 transfer learning model and the final algorithm.The custom CNN implementation takes inspiration from a simple CNN described from Udacity in the notebook. It consists of multiple convolutional layers (with ReLU activation) that extract more feature maps as the network goes deeper, each paired with a maxpooling layer to

reduce maps sizes. The final two layers are a general average pooling layer (GAP) to flatten the input and a dense layer with softmax activation that transforms the feature maps of the last convolutional/maxpooling layers to probabilities distributed for the 133 dog breeds outputs. As anticipated in subsection 2.1, the input images consists of 3 feature maps (RGB channels) for each pixel, which are passed to the hidden layers by the input layer. The 3 hidden convolutional layers outputs, respectively, 16, 32 and 64 feature maps. The hidden layers network is implemented with a kernel size of 2, a striding of 1 and no padding. Now it follows a more detailed description of the network components:

1. 2D convolution: performs a convolution operation on given input feature maps. The kernel size, stride, padding and output feature maps number define the behaviour and the number of learnable parameters for this operation.
2. ReLU: the rectified linear unit (ReLU) activaton follows the convolution operation in order to capture the non-linear features contained in the images.
3. Maxpooling: it performs a down sampling of the input. It works by outputting the maximum element from each subpart in the input space. The behaviour of the maxpooling operation is defined by the kernel size and the stride parameters. This operation serves to both diminish the number of learnable parameters, hence the training complexity and to avoid overfitting.

The output layer takes as input the output of the hidden layers and computes the following operations to transform the input data to class probabilities:

1. General average pooling: it performs flattening on the feature maps generated from the previous hidden layers. It prepares the hidden layers output to be connected to the last layer.
2. Dense layer: fully connected layer which collects the previously generated flattened data and outputs a class probability following softmax activation function.

The transfer learning model is an improvement over the custom CNN one. It solves the same dog breeds classification, but by exploiting a more powerful pre-trained architecture. We selected the ResNet-50 [7] model to perform a more advanced image classification on the dog dataset. The ResNet-50 is a convolutional neural network with 50 hidden layers. This state of the art model has shown optimal performance on the ImageNet [3] dataset and won, in 2015, the ILSVRC competition [7]. These properties make the ResNet-50 architecture an optimal choice to serve for transfer learning with the dog dataset. Here, the idea is to take advantage of the pretrained hidden layers and to substitute its output layer with our custom CNN 133 nodes output layer, matching the dog breeds classification task. Only the weights and biases of the last layer are trained with the dog dataset. Figure 5 visually summarizes how transfer learning works.

The last implementation step concerns the final algorithm which combines the previously described models together: The ResNet-50 [7] dog detector and OpenCV [1] human face detector components check whether the input picture contains a dog or a human, if this is the case, the ResNet-50 transfer model algorithm predicts the resembling or actual dog breed. However, if neither a human or a dog is detected, the algorithm returns an error.

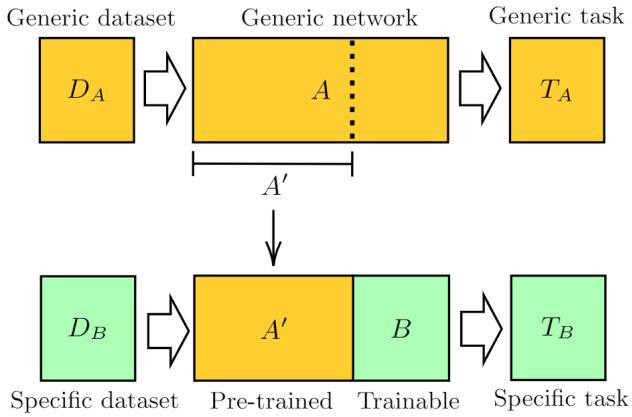

Fig. 5. Image representing how transfer learning works, from this website.

3.3 Refinement

The models and algorithms implemented in this project have been refined to improve the overall accuracy and meet the benchmarks stated in subsection 2.4. Mainly:

1. Batch size: Both the custom CNN and the transfer learning model are trained following mini-batch gradient descent. Due to the high volume of training sets in deep learning, it is common practice to optimize the model's parameters with subset of the training data. Higher batch size results in higher inference accuracy but also in higher computational complexity [8]. We discovered that a batch size of 20 was a good trade-off between performance and accuracy.
2. Optimizer choice: RMSprop [9] was found to be a good optimizer for this task.
3. Loss function: Categorical cross-entropy was found to well describe the loss during the training of the neural network.
4. Detecting dogs before humans: The order of the detectors has been optimized to first use the model that produces less false positives. Indeed, the dog detector has a smaller miss-classification rate on human images (0%) than the human detector on dog images (11%).

4 Results

In this section we evaluate the trained models and discuss the results of the algorithm.

4.1 Model evaluation and validation

Both the custom CNN and the transfer learning model have been trained on the training set only. The number of epochs have been selected empirically and

resulted in 20 epochs for both. At each epoch, the model was tested on validation set to see how well the model generalizes on unseen data. The model that performed best on the validation, with its learned parameters, was saved and then later retrieved to be tested for accuracy on the test set. Saving the model with the best validation score ensures that the final model is the one that better generalizes on unseen data. The best custom CNN model scored an accuracy of 5.02% on the test set. The ResNet-50, as expected, scored an higher accuracy of 82.6% on the same test set. The last step was testing the final algorithm on the custom dataset, described in 2.1. The results confirm that the algorithm is working as expected and that is able to distinguish between the three type of images following the accuracy metrics of its components. A visualization of the application output can be seen in figure 6

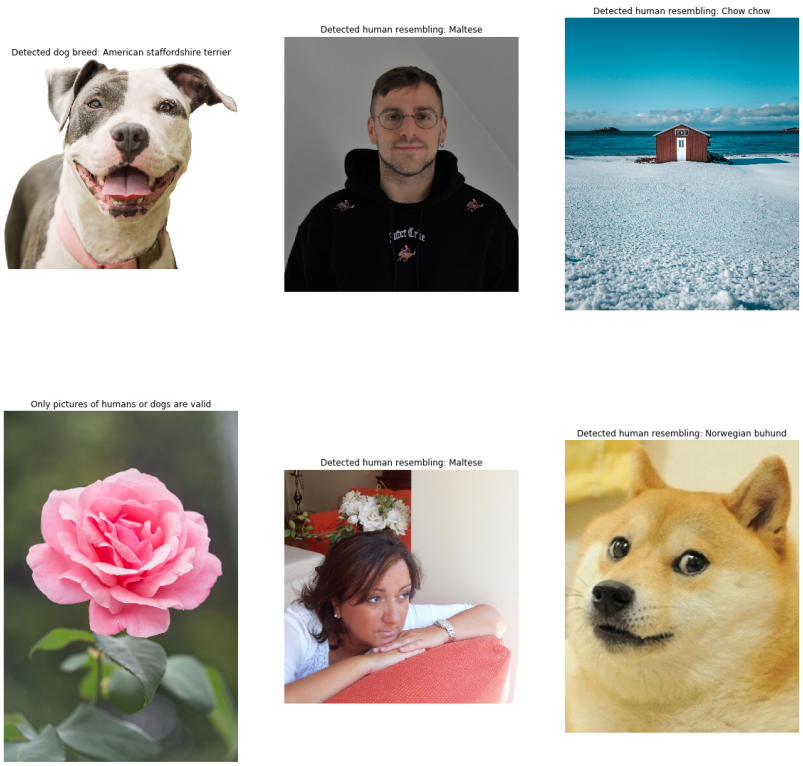

Fig. 6. Visualization of the final algorithm applied to the custom dataset.

4.2 Justification

Both the trained models have an higher accuracy of the benchmarks described in subsection 2.4.

5 Conclusion

In this section we briefly summarize the project once again and provide possible future improvements.

5.1 Reflection

In this project, we have analyzed and solved an image classification task. First, we introduced the problem to be solved and outlined main background concepts like supervised learning, artificial neural networks and convolutional neural networks (see 1). We then described, in section 2, the input datasets, the models used to exploit these datasets and the relative benchmarks. We then explained, in section 3, the pre-processing pipeline for the input data, how we built a CNN and a transfer learning model. Lastly, we showed, in section 4, that both the CNN and the transfer learning model obtained positive results and have an higher classification accuracy than their benchmarks.

5.2 Improvement

The implemented algorithm solves the original problem, with good accuracy in classifying actual or resembling dog breeds. There is, though, a margin of optimization to improve the resulting algorithm, as can be seen from the dog picture and house picture labeled as humans. Here we expose some ideas:

1. Increasing the dog dataset size: Having more examples to train an algorithm with helps increasing the overall model accuracy and reducing the occurrence of over - or underfitting. In this regard, data augmentation could be exploited to feed more synthetic data to our model during training.
2. Testing other models for transfer learning: Using different models for transfer learning could result in a better classification algorithm for this task.
3. Testing different optimization algorithms and loss functions: Despite the good accuracy of our choices, there might be better ones than the Rmspropr [9] optimizer and the categorical cross entropy loss function.
4. Increasing the batch size. As described in subsection 3.3, having a bigger batch size directly impacts the model accuracy. Training a model with a more powerful hardware could make possible to efficiently use bigger batch sizes.
5. Images could be further preprocessed by, for example, eliminating the background from foreground, thus allowing only dog or human features to be exploited from the model.

6. Another evaluation metrics could be used in describing and assessing the models performance. Classes were decently balanced and therefore accuracy was found to be reliable. However, one could think about using precision, in order to also reduce false positives in the human faces detector.

References

1. Lienhart, R., and Maydt, J., 2002. An extended set of Haar-like features for rapid object detection. Proceedings. International Conference on Image Processing, 1, I-I.
2. Caruana, R. and Niculescu-Mizil, A., 2006, June. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168).
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.
4. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
5. Chollet, F., and others, 2015. Keras. GitHub. Retrieved from https://github.com/fchollet/keras.
6. Luo, W., Li, Y., Urtasun, R. and Zemel, R., 2016. Understanding the effective receptive field in deep convolutional neural networks. In Advances in neural information processing systems (pp. 4898-4906).
7. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
8. Radiuk, P.M., 2017. Impact of training set batch size on the performance of convolutional neural networks for diverse datasets. Information Technology and Management Science, 20(1), pp.20-24.
9. Zaheer M., Reddi S. J., Sachan D., Kale S. and Kumar S., 2018. Adaptive methods for nonconvex optimization. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18) (pp. 9815–9825).
10. Sun, S., Cao, Z., Zhu, H., and Zhao, J., 2019. A survey of optimization methods from a machine learning perspective. IEEE transactions on cybernetics, 50(8), 3668-3681.