Image Classification Using Multiple Pre-trained Deep Learning Models

HONG Yuxiang

yhongbb@connect.ust.hk

Introduction

This project is to evaluate six basic pre-trained deep learning models (AlexNet, VGG16, ResNet50, Inception V3, DenseNet121, and MobileNetV2) in image classification tasks. Through the standard ImageNet pre-processing process, I made model predictions on 10 images and computed Top-1 and Top-5 accuracies for each model. The experimental results show that DenseNet121 achieves 100% accuracy on both, which performs best. VGG16 and InceptionV3 also achieve 100% accuracy on Top-5, demonstrating their excellent image classification capabilities. Under the comparison, ResNet50 and MobileNetV2 have relatively lower accuracies.

Methods

In this project, I conducted an image classification task using six widely used pre-trained deep learning models. A total of 10 images in PNG or JPEG format were provided. The images in the folder *images* were used for image classification; the images in *images_label* have classified labels, which were used to compare with outputs generated by the models.

These six models in the experiment are AlexNet, VGG16, ResNet50, Inception V3, DenseNet121, and MobileNetV2. All these models are obtained from the torchvision library in PyTorch and were pre-trained on the ImageNet dataset. Each model was loaded and set to evaluation mode (*.eval()*) to ensure consistent inference behavior.

Before feeding the images into the models, all images are uniformly scaled to 224 × 224, converted to a Tensor, and normalized using the ImageNet standard normalization parameters: Mean: [0.485, 0.456, 0.406], standard deviation (std): [0.229, 0.224, 0.225].

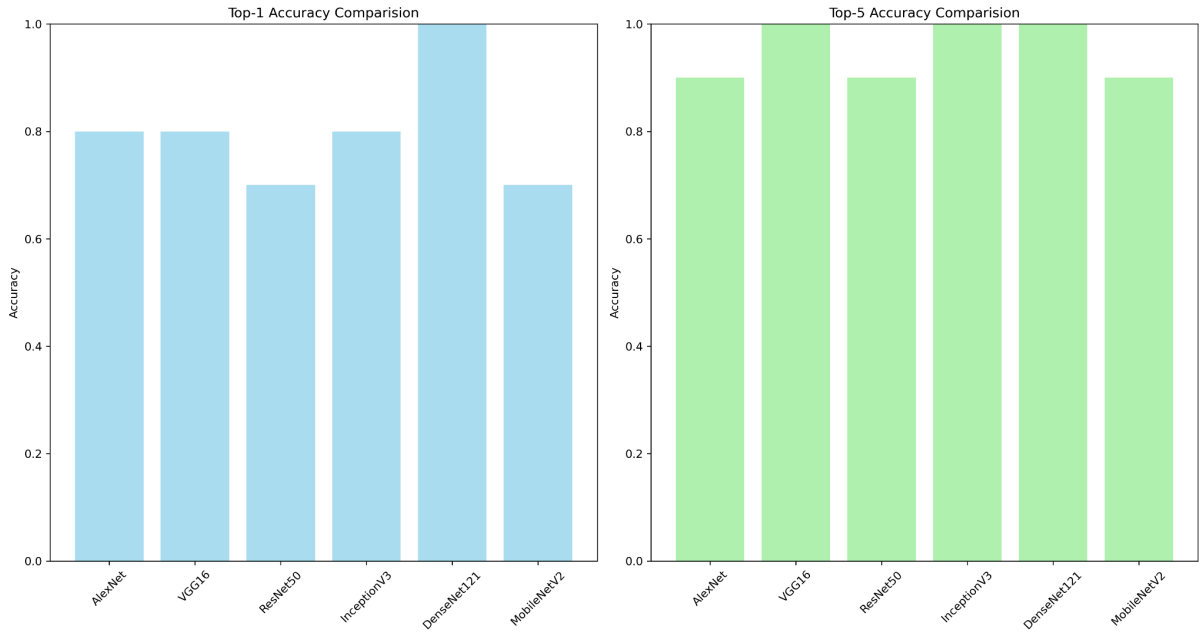
For evaluation, I calculated two standard metrics for each model:

- Top-1 Accuracy: whether the model's most confident prediction (top-1) matches the ground truth label.
- Top-5 Accuracy: whether the ground truth label is included in the model's top 5 predicted classes.

These metrics were used to compare the performance of the different models on the same set of images.

Results

This experiment evaluated the image classification capabilities of six pre-trained models on 10 test images, and the main comparison indicators are Top-1 and Top-5 accuracies.
Below is the bar chart comparing the Top-1 and Top-5 accuracies and the table of models' evaluation summary.



	AlexNet	VGG16	ResNet50	InceptionV3	DenseNet121	MobileNetV2
Top-1 accuracy	0.800	0.800	0.700	0.800	1.000	0.700
Top-5 accuracy	0.900	1.000	0.900	1.000	1.000	0.900

As shown in the figure and table, we can see that DenseNet121 performed best on both of these two indicators, achieving 100% accuracy. VGG16 and InceptionV3 also achive 100% Top-5 accuracy, but their Top-1 accuracies are slightly lower. AlexNet, RestNet50, and MobileNetV2 have relatively lower accuracies on both Top-1 and Top-5 accuracies.

Below is the detailed prediction output table, which contains the predictions made by the six models, input images, and probabilities of the prediction.

image_id	AlexNet (class, probability)	VGG16 (class, probability)	ResNet50 (class, probability)	InceptionV3 (class, probability)	DenseNet121 (class, probability)	MobileNetV2 (class, probability)
1	(golfcart, 0.87)	(golfcart, 0.99)	(golfcart, 0.68)	(golfcart, 1.00)	(golfcart, 0.63)	(golfcart, 0.93)
2	(pop_bottle, 0.09)	(vacuum, 0.53)	(syringe, 0.54)	(syringe, 0.70)	(water_bottle, 0.33)	(syringe, 0.36)
3	(library, 0.32)	(library, 0.95)	(bookshop, 0.51)	(library, 0.94)	(library, 0.95)	(bookshop, 0.50)
4	(car_mirror, 1.00)	(car_mirror, 1.00)	(car_mirror, 1.00)	(car_mirror, 1.00)	(car_mirror, 0.97)	(car_mirror, 0.99)
5	(monitor, 0.12)	(laptop, 0.42)	(laptop, 0.23)	(laptop, 0.77)	(desktop_computer, 0.23)	(desk, 0.30)
6	(zebra, 1.00)	(zebra, 1.00)	(zebra, 1.00)	(zebra, 1.00)	(zebra, 1.00)	(zebra, 1.00)
7	(school_bus, 1.00)	(school_bus, 1.00)	(school_bus, 1.00)	(school_bus, 1.00)	(school_bus, 1.00)	(school_bus, 1.00)
8	(pillow, 1.00)	(pillow, 1.00)	(pillow, 1.00)	(pillow, 1.00)	(pillow, 1.00)	(pillow, 1.00)
9	(fireboat, 1.00)	(fireboat, 1.00)	(fireboat, 1.00)	(fireboat, 1.00)	(fireboat, 1.00)	(fireboat, 1.00)
10	(carousel, 1.00)	(carousel, 1.00)	(carousel, 1.00)	(carousel, 1.00)	(carousel, 1.00)	(carousel, 1.00)

From this table, we can observe that all six models performed well on classifying Image 4, and the probabilities of predictions on Images 1, 2, 3, and 5 are relatively lower or not uniform.

There is also one crucial and interesting experimental phenomenon: when the contents of images are the zebra (Image 6), the school bus (Image 7), the pillow (Image 8), the fireboat (Image 9), and the carousel (Image 10), all six models gave 1.000 probabilities of predictions (that is, 100% confidence coefficient). Probabilities of predictions on these contents of images by six models are highlighted to distinguish.

Discussion

I will discuss the differences in predictions by models and what might have caused the differences. The reasons why image 4 is well classified by all models might be due to the proportion of the target subject in the image and whether there were many interfering elements.

Image 4_n02965783_car_mirror

In Image 4, the car mirror is centered and shown at an easily identifiable perspective. And there are not many distracting elements around, such as objects that cover the car mirror or objects that occupy a larger proportion of the image. Therefore, all six models have satisfied performances in classifying Image 4.

Next, based on the analysis of Image 4, let's examine whether the images that all six models predicted with 1.000 probability (Images 6, 7, 8, 9, and 10) share similar causes.

Image 6_n02391049_zebra

Image 7_n04146614_school_bus

Image 8_n03938244_pillow.

Image 9_n03344393_fireboat

Image 10_n02966193_carousel

From these four images, we can infer that, except for the occupied proportion in the image and the distracting elements, the physical attributes (such as the appearance, color, and geometric structure) and special features also play important roles in the model's image classification tasks. To be more specific, the stripes on the zebra's body in Image 6, the yellow color and design of the school bus in Image 7, the square shape of the pillow in Image 8, the special water spraying device and the red color of the fireboat in Image 9, and the "horses" of the carousel in Image 10, all of these attributes can be faster captured by models, then, are utilized for identifying which class the image is belonged for.

However, for classifying the rest of the images (Image 1, 2, 3, and 5), the performances of all six models were affected by the factors we discussed when analyzing Image 4.

Image 1_n03445924_golfcart

Image 2_n04557648_water_bottle

Image 3_n03661043_library

Image 5_n03180011_desktop_computer

We can clearly observe that there are more distracting elements in these images, which occupy a large proportion of the images or occupy the center of the images. This complex image constitution makes it difficult for models to capture features and identify the main subject of these images, causing a relatively lower prediction probability.

The most obvious finding in the results is that DenseNet121 performed the best in both Top-1 and Top-5 accuracies. DenseNet was introduced and constructed from the concept that shorter connections between layers close to the input and output can make the convolutional networks deeper and more accurate, and it connects each layer to every other layer in a feed-forward fashion. This architecture makes it have the advantages of alleviating the vanishing-gradient problem and strengthening feature

propagation [1]. Compared with other models, DenseNet121 can better capture multi-level features, thereby enhancing its classification confidence and accuracy.

Conclusion

This experiment conducted an image classification task on a set of 10 custom images by using six classic pre-trained deep learning models (AlexNet, VGG16, ResNet50, Inception V3, DenseNet121, and MobileNetV2), and calculated the Top-1 and Top-5 accuracies of each model as evaluation metrics.

The experimental results show that DenseNet121 overperforms the rest of the models, and its Top-1 and Top-5 both achieve 100%, which indicates that this model can accurately identify all image classes of this image set and has a high prediction confidence coefficient. VGG16 and InceptionV3 also achieve 100% on Top-5 accuracy, but their Top-1 accuracies are both 80%. AlexNet, ResNet50, and MobileNetV2 all have relatively lower Top-1 accuracies, which are correspondingly 80%, 70%, and 70%, but their Top-5 accuracies are maintained at 90%, which states that they have a certain ability for image classification.

Something notable is that a few images (like the zebra, the school bus, the pillow, the fireboat, and the carousel) are correctly classified by all six models with 100% prediction probabilities, which means that these images' physical attributes and their special features can be easily captured by these models, assisting them in classifying image classes.

In addition, DenseNet121's great performance has a close relation with its special architecture, which achieves efficient feature reusage, making it have a stronger image classification capability.

This experiment provides a reference basis for the selection of different pre-trained models in practical image classification tasks. At the same time, it demonstrates that DenseNet121 is a great option to pursue a high accuracy in this kind of task.

Reference

[1] Huang, G., Liu, Z., Laurens, V. D. M., & Weinberger, K. Q. (2016, August 25). *Densely connected convolutional networks*. arXiv.org. https://arxiv.org/abs/1608.06993