Image classification with distribution shift AI@2023UNICT Challenge

Manuel Scionti SCIONTI.MANUEL@HOTMAIL.IT

1. Model Description

For this object detection task, I chose to use the **Faster R-CNN** model, as described in the paper by Ren et al. (2016) (Ren et al., 2016). This deep convolutional neural network is mostly used for object detection and can be thought of as an advancement over prior methods such as R-CNN and Fast-R CNN. This model's capacity to build a region proposal network, which creates region proposals that are then transferred to the detection model (Faster R-CNN) for object testing, is a critical feature. In addition to Faster R-CNN, I used various models to serve as a type of 'backbone' for the object identification model, basically acting as a feature extractor. The goal is to compare their performance and identify which model produces the greatest outcomes when employed as a backbone. I've specifically chosen these three models as backbone:

1. ResNet50 FPN (He et al., 2016)
2. ResNet50 FPN v2 (He et al., 2016)
3. MobileNet (Howard et al., 2017)

ResNet50 FPN and ResNet50 FPN v2 are a deep neural networks that combines elements of **Residual Networks** (ResNet) and **Feature Pyramid Networks** (FPN).

1.1. The Feature Pyramid Network

The Feature Pyramid Network (FPN) (Lin et al., 2017) is a customized feature extractor built with accuracy and speed in mind. It is meant to improve object detection performance in models such as Faster R-CNN. When compared to ordinary feature pyramids, FPN substitutes the standard feature extractor and generates numerous feature map layers known as multi-scale feature maps, which provide higher-quality information for object detection. FPN has two pathways: one from the bottom up and one from the top down. The bottom-up pathway is a normal feature extraction convolutional network, however as you progress up the pathway, the spatial resolution drops but the semantic value of each layer grows. However, not all layers are suitable for object detection; the lower layers have high resolution but lack the necessary semantic information for effective detection, leading to a significant slowdown in speed. Therefore, in object detection tasks, only the upper layers are utilized, which is why SSD (Single Shot MultiBox Detector) performs less effectively for small objects. What sets FPN apart is its top-down pathway, which reconstructs higher-resolution layers from a semantically rich layer. These reconstructed layers are strong in terms of semantic understanding but may lack precision in object localization due to the downsampling and upsampling processes. To address this, FPN introduces lateral connections that link the reconstructed layers with the corresponding feature maps. These connections help improve object localization and also simplify training, akin to the skip connections used in ResNet architecture.

1.2. The backbones

1.2.1. Faster R-CNN with ResNet50 FPN

When combined with the ResNet50 Feature Pyramid Network (FPN) backbone, faster R-CNN demonstrates strong and accurate object identification capabilities. The ResNet50 FPN design is built on the Residual Network (ResNet) architecture, which excels at feature representation learning. The deep residual connections in ResNet alleviate the vanishing gradient problem, allowing for the training of extremely deep neural networks. When paired with the Feature Pyramid Network, it enables the model to handle objects of varied sizes and complexity effectively. This combination of ResNet50 and FPN results in a feature-rich backbone that improves the Faster R-CNN's object identification performance, especially in settings with changing sizes and overlapping instances.

1.2.2. Faster R-CNN with ResNet50 FPN v2

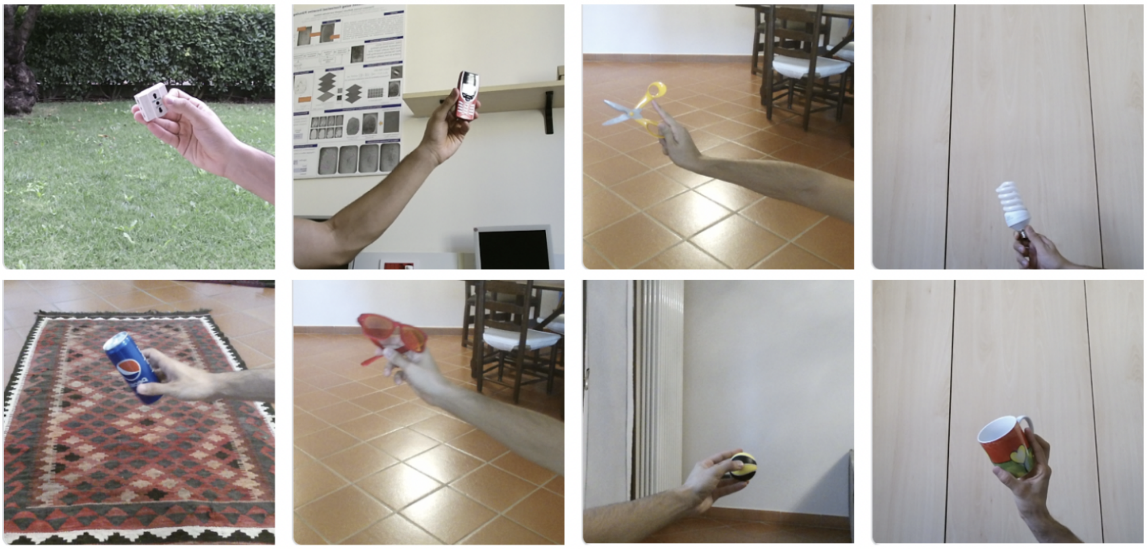
The incorporation of the ResNet50 FPN v2 backbone into the Faster R-CNN framework results in a strong mix of architectural advances. ResNet50 FPN v2 expands on ResNet's basic capabilities by strengthening feature representation with skip connections, batch normalization, and efficient weight initialization. The "v2" variation adds modifications such as enhanced residual connections, lowering computing cost while maintaining feature richness. This backbone gathers contextual information while retaining computational efficiency when used within Faster R-CNN. The use of FPN simplifies the processing of object instances at many sizes, guaranteeing strong detection performance in difficult settings. Faster R-CNN with ResNet50 FPN v2 emerges as an appealing option for object identification jobs requiring a balance of accuracy and processing economy.

1.2.3. Faster R-CNN with MobileNet

When MobileNet is used as the backbone in Faster R-CNN, it adds a new degree of efficiency and agility. MobileNet is well-known for its lightweight architecture, which makes it ideal for resource-constrained situations. MobileNet's depth-wise separable convolutions minimize the amount of parameters and processing needs while retaining the capacity to extract relevant features. Faster R-CNN can now efficiently perform object identification tasks on mobile and edge devices. The Faster R-CNN's region proposal network enhances MobileNet's feature extraction capabilities, resulting in a simplified yet efficient design. While MobileNet may be less accurate than bulkier backbones, it thrives in cases requiring real-time or on-device object identification, illustrating the synergy between model efficiency and practical deployment.

2. Dataset

The dataset provided by the professor for this assignment is part of a Kaggle competition (ConcettoSpampinato, 2023). It consists of a collection of hand-held object images and is characterized by a notable domain shift between the training and test data. In the training set, there are 1,600 images, with 200 images dedicated to each of the 8 classes. Conversely, the test set comprises 800 images, all originally sized at 350x350 pixels. These 8 classes encompass a diverse range of objects, including plug adapters, mobile phones, scissors, light bulbs, cans, sunglasses, balls, and cups, as illustrated in Figure 1.



.

Figure 1: Dataset's classes

A distinctive characteristic of this dataset is that Class 0, which corresponds to plug adapters, has the same backdrop as all of the test set photos, resulting in a phenomenon known as **domain shift.** This domain change has the potential to significantly alter the model's performance. Bounding boxes have been strategically supplied as supplementary information for this picture classification challenge to reduce this impact. In a.csv file, these bounding boxes are properly structured and include information such as picture id, x1, y1, x2, y2, and class. These boxes allow us to limit the model's attention to the picture regions we specify. To improve the model's performance, I made strategic use of the **Albumentations** library, a well-known picture improving tool. During the training phase, I applied a set of transformations, including horizontal flipping with a 50% probability and grid distortion with a 30% probability.

3. Training procedure

The training process can be summarised as follows:

1. **Region Proposal Network (RPN) Training**
   - Positive and negative class labels are assigned to anchors based on their Intersection over Union (IoU) overlap with ground-truth boxes.
   - Non-maximum suppression (NMS) is used to reduce redundancy among RPN proposals.

2. **Loss Functions**
   - The model uses a composite loss function that combines classification and bounding box regression losses.

3. **Optimizer**
   - Stochastic Gradient Descent (SGD) is chosen as the optimizer with specific settings, including a learning rate of 0.005, momentum of 0.9, and weight decay of 0.0005.

4. **Learning Rate Scheduling**
   - A learning rate scheduler is implemented to gradually reduce the learning rate by a factor of gamma at defined intervals (step_size epochs).

5. **Training Duration**
   - The model is trained for a total of 5 epochs, utilizing a NVIDIA T4 GPU provided by Kaggle for computation.

4. Experimental Results

Backbone Model	Accuracy
ResNet50 FPN v2	0.995
ResNet50 FPN	0.842
MobileNet	0.615

Table 1: Accuracy of Different Backbone Models

In the experiment involving Faster R-CNN with three different backbone models, ResNet50 FPN, ResNet50 FPN v2, and MobileNet, it was observed that ResNet50 FPN v2 outperformed the others, achieving an impressive accuracy of 99%. In contrast, MobileNet exhibited the lowest performance, with an accuracy of 60%. Several hypotheses can be considered to explain the poorer performance of MobileNet in this object detection task. MobileNet, known for its efficiency and lightweight architecture, may struggle with the complexity and diversity of objects present in the dataset. Its shallower and more streamlined design could result in a limited ability to capture intricate object features and nuances, particularly in cases where objects exhibit significant variations in size, shape, or context. Additionally, the reduced model size of MobileNet may lead to a decreased capacity to represent and differentiate objects effectively, impacting its detection accuracy. These factors collectively highlight the importance of selecting a backbone model tailored to the specific characteristics and challenges of the object detection task at hand.

References

Giovanni Bellitto Simone Palazzo ConcettoSpampinato, Federica Proietto. Ai@unict 2023, 2023. URL https://kaggle.com/competitions/aiunict-2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944. IEEE Computer Society, 2017.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.