Image segmentation for pet identification: a semi-supervised approach

Candidate Number : ZWLD8

UCL

1 Introduction

Convolutional Neural Networks have shown remarkable results in the fields of image classification and semantic segmentation. In particular, image classification consists of establishing whether an image pertains to a certain class (e.g. whether it represents a plane, a car, or a dog), whereas semantic segmentation is a more fine-grained classification task, concerning individual pixels inside an image, all with their own label and prediction. Therefore, pixels making up, say, a human, will be labelled as such, while pixels making up the background, a car, or a different object will be labelled accordingly. Semantic segmentation has proven especially useful in some fields such as autonomous driving, where computer vision is applied in order for the car to distinguish different objects in its proximity and where they are located. In this study, we will not go as far as recognising different classes, instead, we will focus on binary classification.

One of the main issues with supervised training of NNs for semantic segmentation is the large amounts of labelled data required. Each image requires a pixel mask, denoting a label for each of the pixels present. Acquiring such datasets can be costly and time-consuming. For this reason, a lot of research has focused on finding alternative ways to obtain efficient and reliable trained networks using a smaller fraction of available data. In this paper, we delve into one of these approaches: semi-supervised learning.

Semi-supervised learning is an approach that combines a small amount of labelled data with a large amount of unlabelled data during training. The objective is to leverage the unlabelled data to improve the performance of the model by exploiting underlying patterns or structures in the data. This can be achieved using various techniques, such as:

Self-training: The model is initially trained on the labelled data and then used to predict labels for the unlabelled data. High-confidence predictions are then used to augment the training set, and the model is retrained (Papandreou et al., 2015). Consistency regularisation: This technique enforces the model's predictions to be consistent across different augmentations or perturbations of the same unlabeled data. This helps the model learn more robust and invariant features (Sajjadi et al., 2016).

Generative models such as GANs or variational autoencoders can also be used to generate additional labelled data (Odena, 2016).

Several scholarly papers have focused on these topics. For instance, Papandreou et al. (2015) proposed a weakly- and semi-supervised learning approach using the self-training technique introduced above. In another study, Tarvainen and Valpola (2017) introduced the "mean teacher" approach, which involved weight-averaged consistency targets for improved semi-supervised deep learning. We will be using the same framework, which will be explained in detail in the following section. Hung et al. (2018) explored the employment of an adversarial approach to enforce the consistency of predictions on unlabelled data.

In this study, we will be training a semi-supervised NN on the widely used Oxford IIT Pet dataset, a 37 category pet dataset with roughly 100 images for each class created by the Visual Geometry Group at Oxford. The images have large variations in scale, pose and lighting. All images have an associated ground truth annotation of the breed, head ROI, and pixel level trimap segmentation. In our case we will be removing differences between dog breeds and just focus on distinguishing between the animal and the background.

Our aim will be to implement a highly accurate image segmentation NN and analyse differences when training the network with different proportions of labelled and unlabelled data, with the hope of preserving performance while cutting down on the amount of labelled data necessary.

2 Methods

2.1 U-Net Architecture

The U-Net network architecture is a type of convolutional neural network (CNN) designed specifically for semantic segmentation tasks, first introduced by Ronneberger, Fischer, and Brox (2015).

U-Net is characterized by its symmetric shape, resembling the letter "U", which consists of an encoding (contracting) path and a decoding (expanding) path.

The encoding path consists of multiple convolutional layers followed by max-pooling layers, which downsample the input image and extract high-level features. Each convolutional layer is followed by a Rectified Linear Unit (ReLU) activation function, and the number of feature maps is doubled after each downsampling step.

The decoding path is designed to upsample the feature maps and recover the spatial information lost during the encoding process. This is achieved by using transposed convolutional layers or upsampling layers followed by convolutional layers. To preserve the spatial information from the encoding path, skip connections are introduced to concatenate the feature maps from the encoding path with the upsampled feature maps from the decoding path, allowing the network to recover the fine-grained details required for accurate segmentation (Ronneberger et al., 2015).

The final layer of the U-Net architecture is a 1x1 convolutional layer with a sigmoid activation function, which generates the pixel-wise probability map for each class.

2.2 Mean-Teacher

The mean teacher model is a semi-supervised learning technique introduced by Tarvainen and Valpola (2017). This approach is based on consistency regularisation, which aims to ensure that the model's predictions are consistent across different augmentations or perturbations of the same input data. In the mean teacher framework, there are two networks: the student network and the mean teacher network. The student network is trained using the available labelled data and minimises the consistency loss with respect to the mean teacher network. The mean teacher network is an exponential moving average of the student network's weights.

The consistency loss minimised by the student network is as follows:

$$L_{consistency} = \frac{1}{2} \cdot E_{x,y} \left[|f_{\theta_1}(x) - f_{\theta_2}(x)|^2 \right] \tag{1}$$

where $f_{\theta_1}(x)$ and $f_{\theta_2}(x)$ are the outputs of the two networks with parameters θ_1 and θ_2, respectively, on input x. The expectation is taken over the data distribution, with x and y denoting a sample input and label pair. At each time step the teacher network is updated according to a smoothing coefficient hyperparameter determining the weight of the previous network trained. We also set a wait time that determines after how many epochs the unsupervised loss is introduced in the running loss, as at first the model needs to learn from labelled data.

2.3 Data Augmentation

During the preprocessing, all images are downscaled to the size of 64x64 as they all have to be the same size to enter U-Net. We further implement a series of random data augmentations on the training data, such as Gaussian noise and random color inversions. Data augmentation is helpful in increasing the diversity and size of our training data, helping the model generalize better, preventing overfitting, and improving overall performance.

2.4 Data Loading and Training

The dataset was divided into three subsets: the training set, validation set, and testing set. The training set contained a combination of labelled and unlabelled data, with a batch size of 32. A script was developed to randomly select a specified percentage of images in the dataloader and remove their labels. This process was designed to begin after a predetermined number of epochs, allowing the model to train with more labelled data during the initial epochs before introducing unsupervised labels.

Once the predefined epoch threshold was surpassed, a certain proportion of unlabeled data was incorporated into each batch, and the model's predictions began to be used as labels for calculating the loss.

Given the imbalance between animal and background pixels in our dataset, we decide to use the Dice Loss for our model, a loss commonly used in image segmentation tasks. Introduced by [], it is defined as follows:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y} i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i} \tag{2}$$

where N is the total number of pixels in the image, y_i is the ground truth label of pixel i (either 0 or 1), and $\hat{y}_i$ is the predicted label of pixel i (also either 0 or 1).

The Dice Loss measures the overlap between the predicted segmentation mask and the ground truth segmentation mask, with a perfect overlap resulting in a value of 1 and a completely different mask resulting in a value of 0.

To train our model we used an Adam optimiser with 0.001 learning rate.

3 Experiments

For the first part of our analysis we train three different networks:

- Lower bound network: a baseline network which we train on a sample of labelled data with no additional unlabelled data (M25 L)
- Upper bound network: a network trained on all the data available including all labels (MU)
- Mean-teacher network: a semi-supervised network using the mean-teacher approach, which uses a sample of the labelled data and the remaining unlabelled data (M25)

M25L and M25 will use a proportion of labelled data equal to 25% of the total. The lower bound will serve as a baseline for evaluating our other networks. Naturally, we expect them to be better as they make use of larger amount of data and information. The upper-bound is trained with a fully supervised approach, therefore we expect it to outperform the other networks, given it has access to all of the information available. Finally, the mean-teacher network will be the object of our study, and we aim to obtain a performance that is approximately as good as the upper bound and substantially higher than the lower bound.

Our second experiment consists in expanding our range of networks by training a larger number of them with varying proportions of labelled data. We aim to discover how the performance will change in relation to the amount of labelled data we feed during training. We train an additional four networks to accomplish this:

- M10L and M05L: two baseline network trained on 10% and 5% of the labelled data respectively.
- M10 and M05: two mean teacher networks trained with 10% and 5% of the labelled data respectively, with the remaining data used for unsupervised learning.

4 Results

We now move on to displaying and evaluating our results for our experiments. In order to better evaluate the accuracy of our models, we report the IoU (Intersection over Union) score, which is a measure of the similarity between two sets of pixels: the predicted segmentation mask and the ground truth segmentation mask. It is commonly used for image segmentation tasks and widely considered to be a better metric compared to accuracy, given the class imbalance between animal and background pixels. For part 1, the three models we have trained report the following accuracies and IoU.

Model	Accuracy (%)	IoU (%)
M U	90.29	79.23
M25 L	86.93	73.01
M25	88.24	75.46

As we can see, the upper bound is the best performing model, as expected. The mean teacher model still achieves a superior performance compared to the lower bound, therefore the unsupervised component accounts for an increase of 1.31% in accuracy and 2.45% in IoU. Overall, the results are also quite convincing, as with only 1/4 of the data, we obtain scores close to the upper bound.

We now report results for the other four models trained in part 2.

Model	Accuracy (%)	IoU (%)
M10 L	85.03	70.52
M10	85.49	69.97
M05 L	83.04	67.83
M05	83.23	68.47

These results also show clear improvements at different levels of labelled data between the lower bound and mean teacher model, further proving the effectiveness of adding an unsupervised component to the training. Nevertheless, the impact of the unsupervised component seems to be negligible compared to the supervised part, between 0-0.5% for accuracy and 0.5-1% for IoU. Perhaps what is more interesting about these results, is the remarkably good performance achieved by models trained on such small amounts of labelled data. The M05 model, for example, achieves a lower IoU by 10% (-15%) around -compared to the Upper Bound, while using -95% of the data. In situations where data collection is costly and superior performance is not of the essence, this may well be a viable solution.

5 Discussion

Our mean teacher model with 25% data clearly improves over the baseline but as expected does not perform as well as the upper bound. An example of the segmentation obtained is shown in Figure 1 (see Appendix). As we can see, the prediction is quite close to the ground-truth mask, and the body of the cat is more easily identified compared to its tail and legs. Overall, the shape and location of the cat in the image is identified.

In Figure 2 (see Appendix) we show the loss change over the different training epochs. As discussed before, the loss increases when the relative weighting of the unsupervised loss is updated, explaining the jump between epochs 20-35.

We always obtain for all models trained (apart from IoU in M10) a small improvement in performance when implementing the mean teacher algorithm. Nevertheless, the main finding is probably that we can obtain satisfactory results with very small proportions of labelled training data. This means that out of all the images available, most of them actually provide a marginal increase in performance which is negligible. With more efficient pseudo-labelling methods, perhaps the performance can improve to levels closer to the upper bound even for situations with very little data.

6 Conclusion

We have successfully implemented models for image segmentation of animals from the Oxford IIT Pet dataset. In particular, we have used a mean teacher algorithm to deal with lower amounts of labelled data, an issue that often presents itself in real world scenarios. In part 1, we show how the mean teacher model improves over the baseline, but obviously cannot perform as well as the upper bound. The results obtained are nevertheless satisfactory. In part 2 we train additional models showing that even at low levels of supervised training, acceptable performance is achieved. Overall, the mean teacher model works as expected, but it is worth thinking about whether it makes sense to go through this process when the model gets only slightly better.

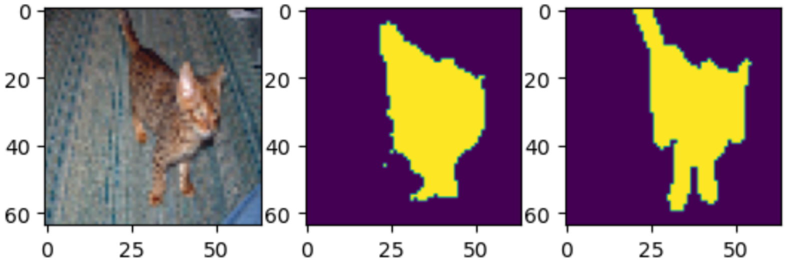
Future research could focus on finding alternative semi-supervised frameworks that can introduce sizeable improvements, or finding techniques to analyse which images are more important for training (e.g. increase accuracy the most) and prioritising their labelling. Other work related to this dataset could also be focused on multiclass recognition of the specific breeds present in the data.

References

1. Papandreou, G., Chen, L.C., Murphy, K.P. and Yuille, A.L., 2015. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In Proceedings of the IEEE international conference on computer vision (pp. 1742-1750).
2. Sajjadi, M., Javanmardi, M. and Tasdizen, T., 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. Advances in neural information processing systems, 29.
3. Odena, A., 2016. Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583.
4. Hung, W.C., Tsai, Y.H., Liou, Y.T., Lin, Y.Y. and Yang, M.H., 2018. Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv:1802.07934.)
5. Ronneberger, O., Fischer, P. and Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241). Springer International Publishing.
6. Tarvainen, A. and Valpola, H., 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems, 30.
7. Jadon, S., 2020, October. A survey of loss functions for semantic segmentation. In 2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB) (pp. 1-7). IEEE.

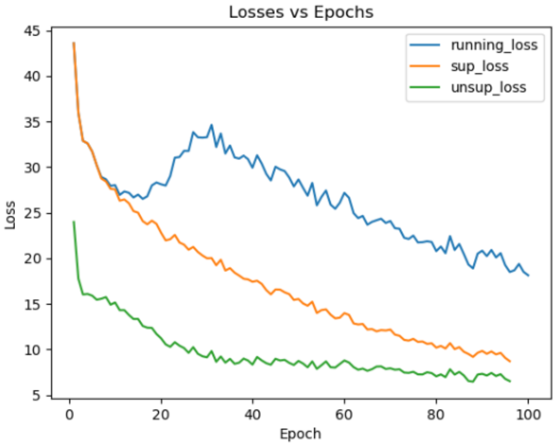
Appendix

Figure 1: Example segmentation



Left: Original image Centre: Predicted mask Right: Ground-truth mask

Figure 2: Running, supervised, unsupervised loss by epoch


Losses vs Epochs
running_loss
sup_loss
unsup_loss
Loss
Epoch