Effectiveness of Mean Teacher in Utilising Unlabeled Data

Student id:21140985

UCL

1 Introduction

This paper evaluates the performance of a semi-supervised learning framework, with a Mean Teacher algorithm, used for semantic segmentation of pet images from Oxford-IIIT Pet dataset. In particular, we compare the framework results with a baseline without unlabelled data, and we also compare it with an upper-bound with all labels being available. Furthermore, we tackle two research questions. The first question relates to the impact of different noise levels, added to the inputs, on the performance of the framework. The second question focuses on the impact of different labelled data ratios on the framework's performance. Our experiments show that the framework performs better than the baseline by achieving 78.5% test IoU accuracy which is significantly higher than the baseline which gives 48.9%. The framework with an upper-bound performs the best as it gives 91.2% accuracy. The result of our research on the noise level shows that initially increasing the noise level improves test accuracy, but after a certain level it causes the accuracy to decrease. Lastly, we observe that increasing the labelled that ratio improves model accuracy only to a certain point, suggesting that not having enough labelled data is not the only factor which limits model accuracy, but other factors also play a part.

Supervised learning performs best when we have a large amount of high quality labelled data which is often not the case, especially in some domains such as healthcare [6] One approach to tackle this problem is semi-supervised learning which alongside labelled data also utilises unlabelled data during training. Various algorithms have been developed for semi-supervised learning and the one we decided to use is called the Mean Teacher algorithm [7]. This algorithm has been shown to perform better compare to Π model [3] and temporal embedding algorithms [4], and it also performs better than pseudo label algorithm in most cases [5].

2 Methodology

2.1 Mean Teacher

Mean Teacher is a variation of consistency regularisation which applies data augmentation to semi-supervised learning by leveraging the idea that a model

should output the same class distribution for an unlabeled example even after it has been augmented [1]. The mean teacher algorithm was implemented in a similar fashion to that in the original paper [7], however the model architecture used, called DeepLabv3 [2], is different from the one used in the original paper. In the Mean Teacher algorithm, we have two identical DeepLabv3 models called student and teacher which are trained simultaneously. At each iteration, the total loss is computed by adding the classification loss for the student model and the consistency loss $J(\theta)$ eq.(1) for the teacher model. The outputs of the student model are used as targets for consistency loss. During backpropagation, only the weights θ_t of the student model are updated, unlike the Π model in which teacher model weights $\theta_t^{'}$ are also updated during backpropagation. The weights of the teacher model in Mean Teacher are updated using the exponentially moving average (EMA) of the student model weights eq.(2). This approach of updating teacher weights helps improve on temporal embedding where we do not update teacher model weights, but use EMA of student predictions to get targets for teacher model. This approach causes the learned information to be incorporated into the training process at a slow pace because the target is only updated once per epoch whereas the EMA weights in Mean Teacher are updated after every iteration [7].

$$J(\theta) = E_{x,\eta^{'},\eta}\left[||f_{teacher}(x,\theta^{'},\eta^{'}) - f_{student}(x,\theta,\eta)||^2\right] \tag{1}$$

x: input, η , $\eta^{'}$: gaussian noises

$$\theta_t^{'} = \alpha\theta_{t-1}^{'} + (1-\alpha)\theta_t \quad \alpha\text{: } EMA\ \ Weight \tag{2}$$

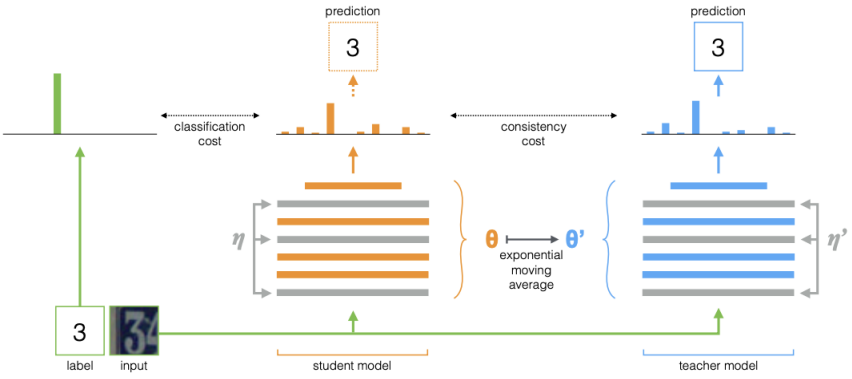

Fig. 1. The Mean Teacher Method

2.2 Noise and Labelled Data Ratio Impact

To address the impact of label noise, we train the semi-supervised framework for different levels of noise and evaluate the results. Similarly, we evaluate the results obtained from training the model for different labelled data ratios. The noise levels used are $0.001, 0.01, 0.1, 1$ and the labelled data ratios are $0.05, 0.3, 0.5, 0.7$.

3 Experiments

For all experiments, we split the data into development (70%) and test (30%) sets. The development set is further split into train (90%) and validation (10%) sets. In the training set, we consider having only (5%) labeled data and the rest unlabeled. For the baseline model, we used only the 5% labelled data from the training set and discarded the 95% unlabeled training data. For the upper-bound model, we used all the training data and provided labels for the unlabelled data. Lastly, for our semi-supervised framework, we kept all labelled and unlabelled data in our training set. To define the labeled and unlabeled samples in the training data, we define a vector of zeros **label_flag** with the same length as labeled data and a vector of ones **nolabel_flag** with the same length as unlabelled data. We concatenate the two vectors and shuffle them to get a one-hot vector **flag** which is used during training to identify label and unlabeled samples. We applied geometric transformation, including random crop, rotation, flips, to input and outputs, and also added Gaussian noise. In all the experiments, we used a pre-trained DeepLabv3 architecture with modified last classification layer.

BCE loss function was used for classification loss and MSE loss was used for consistency loss. Since initially the student model does not give good predictions, but for later epochs the model predictions improve, therefore the consistency loss is given less weight during the first few epochs (rampup phase) to allow the teacher model not to rely heavily on the student model predictions. The weight W_t defined during rampup phase is calculated in the following way:

$$W_t = e^{-5.0 * phase * phase}$$

$$phase = 1.0 - \frac{\min(epoch, rampup\_length)}{rampup\_length} \qquad rampup\_length = 5$$

Fig. 2. Consistency Loss Weight Calculation

When updating the teacher model weights, the EMA weight α_t is calculated using the formula shown below. Initially the weight is 0.9 as the student model weights change more quickly at the start, therefore the teacher model needs to rely less on student model weights. The EMA weight increases to 0.999 as training steps increase.

$$\alpha_t = \max\left(1 - \frac{1}{step_t}, \alpha_0\right) \quad step\_t : \text{iteration number}$$
$$\alpha_0 = 0.9 : \text{initial EMA weight}$$

Fig. 3. EMA Weight Calculation

4 Results

The results from our experiments are summarised in the tables below.

Model	Train IoU	Validation IoU	Test IoU
Baseline with no unlabelled data	84.8%	46.8%	48.9%
Mean Teacher with all labels available (upper-bound)	92.0%	91.4%	91.2%
Mean Teacher with labelled and unlabelled data	79.0%	78.5%	78.5%

Table 1. Performance of our three different models

From Table 1 we observe that the baseline model gives good IoU accuracy on the training set, but does poorly on the validation and test sets. This might be because we have kept only the 5% labelled training data and discarded the rest of the unlabelled data, therefore the model overfits the training data and fails to generalise to data outside its training set. The model which uses mean teacher algorithm with all labels available performs the best because Mean Teacher helps regularise the model which rather than minimizing the classification cost at the zero-dimensional data points of the input space, minimizes the cost on a manifold around each data point, thus pushing decision boundaries away from the labelled data points [7]. Lastly, the mean teacher framework with labelled and unlabelled data performs better than the baseline showing that including unlabelled data can help improve the accuracy of models where we do not have enough labelled data. The difference between the test IoU's of the upper-bound framework and the framework with labelled and unlabelled data is significant (12.7%), implying unlabelled data can help improve accuracy, but it is unlikely to have the similar effect as labelled data.

Table 2 shows the results we obtained for our study which addresses the question about the impact of label noise on the model accuracy. As it can be observed, the test IoU accuracy increases as we increase label noise from 0.001 to 0.01, but starts to decrease with further increase of label noise. This might be because initially increasing the noise helps to regularise the model so that it does not

Model	Label Noise	Train IoU	Validation IoU	Test IoU
Mean Teacher with labelled and unlabelled data	$\mu = 0.0$, $\sigma = 0.001$	80.9%	79.0%	80.1%
Mean Teacher with labelled and unlabelled data	$\mu = 0.0$, $\sigma = 0.01$	84.3%	82.2%	83.7%
Mean Teacher with labelled and unlabelled data	$\mu = 0.0$, $\sigma = 0.1$	83.7%	81.4%	82.8%
Mean Teacher with labelled and unlabelled data	$\mu = 0.0$, $\sigma = 1$	78.5%	73.1%	75.5%

Table 2. Model trained with different label noises

overfit the training data, but as we increase the noise further, we lose bias information from the training data, reducing the correlation between inputs and targets which effects the ability of the model to learn anything from the training data and therefore it underfits the data.

Model	Labelled Data Ratio	Train IoU	Validation IoU	Test IoU
Mean Teacher with labelled and unlabelled data	0.05	79.0%	78.5%	78.5%
Mean Teacher with labelled and unlabelled data	0.3	89.9%	86.8%	89.1%
Mean Teacher with labelled and unlabelled data	0.5	91.2%	88.0%	90.4%
Mean Teacher with labelled and unlabelled data	0.7	91.9%	88.4%	90.8%

Table 3. Model trained with different amount of labeled data

Table 3 shows the results of our study about the second research question. We observe that increasing labeled data ratio from 0.05 to 0.3 causes the test IoU accuracy to increase significantly which means having higher ratio of labelled data can help improve model accuracy. However, this is true to a certain point because we can see from the table that the test accuracy does not increase significantly when we change the ratio from 0.3 to 0.5 and then to 0.7.

5 Discussion and Conclusion

Although supervised learning has shown successful results in many domains, however it is important to reduce dependency on this approach and make other approaches, which require less labelled data, more efficient. This would enhance the ability to train successful models for domain-specific applications where it is expensive, difficult and time-consuming to collect a large amounts of labelled data. Our study explores the semi-supervised approach and more specifically the Mean Teacher algorithm and we found that the model accuracy does increases significantly when we utilise unlabelled data using this algorithm. However our study also shows that providing labels for unlabelled data does even better which implies that the potential of utilising unlabelled data needs to be further improved. This is unlikely to be achieved by further increasing regularisation as our research on the impact of label noise suggests that the model underfits the

data for heavy regularisation. Furthermore, our analysis of different labelled data ratios shows that increasing the labelled data ratio only improves accuracy up to a certain point and after that, it seems that the data is not an issue, but an improvement needs to be made to the Mean Teacher algorithm. For future work, a more recent algorithm called MixMatch [1] could be explored which combines consistency regularisation, entropy minimisation and traditional regularisation approaches to semi-supervised learning. The algorithm guesses low-entropy labels for data-augmented unlabeled examples and mixes labelled and unlabeled data using MixUp algorithm [8] The MixMatch algorithm is illustrated in the figure below from the original paper [1].

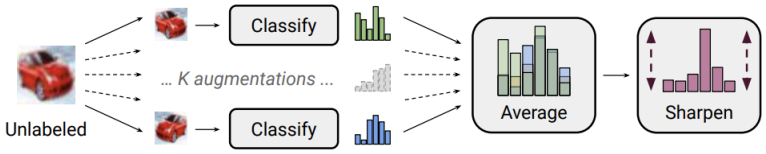

Fig. 4. Consistency Loss Weight Calculation

References

1. David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.
2. Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
3. Corinna Cortes, ND Lawarence, DD Lee, M Sugiyama, and R Garnett. Advances in neural information processing systems 28. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*, 2015.
4. Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.
5. Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.
6. Liyan Sun, Jianxiong Wu, Xinghao Ding, Yue Huang, Guisheng Wang, and Yizhou Yu. A teacher-student framework for semi-supervised medical image segmentation from mixed supervision. *arXiv preprint arXiv:2010.12219*, 2020.
7. Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.
8. Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.