International Journal of Natural and Engineering Sciences
E-ISSN: 2146-0086 19(3): 168-187, 2025
Research Article

Negotiated Representations to Prevent Overfitting in Machine Learning Applications

Nuri Korhan[1a*]

[1]*Istanbul Technical University, Maslak, Istanbul, Turkey*

*Corresponding author: korhan@itu.edu.tr

(Received: 10 December 2025 / Accepted: 19 December 2025)

a: ORCID 0000-0003-4351-2885

Abstract

Overfitting is a phenomenon that occurs when a machine learning model is trained for too long. When a model focuses too much on the exact fitness of the training samples to the provided training labels, it cannot keep track of the predictive rules that would be useful for recognizing patterns in the test data This phenomenon is commonly attributed to memorization of samples and noise. Over-parameterized networks with excessive neurons can memorize noise in small datasets rather than learning generalizable patterns. While it is true that the model encodes various peculiarities as the training process continues, it is argued that most of the overfitting occurs in the process of reconciling sharply defined membership ratios. This study presents an approach that increases the classification accuracy of machine learning models by allowing the model to negotiate output representations of the samples with previously determined class labels. By setting up a negotiation between the model along with an invitation to the machine learning community to explore the limits of the proposed paradigm. This work also aims to incentivize the machine learning community to exploit the negotiation paradigm to overcome the learning related challenges in other research fields such as continual learning. The Python code of the experimental setup is uploaded to GitHub(https://github.com/nurikorhan/Negotiated-Representations).

Keywords: Machine learning, overfitting, negotiated representations

Introduction

The general structure of the supervised deep learning [1] requires us to rely on the labels provided by humans beforehand. With these provided labels, one can create a cost function that informs the model on how far off the prediction is. By trying to decrease the cost with the help of backpropagation, the model is expected to encode the underlying input-output relationships into its weights. This approach works extraordinarily well for big data regimes. However, it becomes unreliable in low data regimes. If the model is large enough to ensure the fitness between samples and their respective labels, it tends to encode the aspects of the samples that are irrelevant to the classification process. The burden of encoding the irrelevant features makes the classifier less accurate on test samples. This phenomenon is called overfitting [2]. Researchers have developed highly sophisticated methods to prevent the data from overfitting.

Data augmentation and regularization techniques are limited in compensating for the overfitting problem [3]. Data augmentation aims to increase the number of samples by slightly vibrating the training samples in the multi-dimensional space to allow each sample to represent its corresponding neighbourhood, and most regularization methods limit the movement of the weights by adding punishment to the loss function. None explicitly addresses the problem of not having membership ratios of the samples to their classes. While it is true that a sample either belongs to a class or not, it is impossible to justifiably represent the samples by their corresponding labels if the proximity of the individual samples to all classes is not accounted for. This is also the case for humans. If an object is sufficiently distant from an observer, the observer will assign equal probabilities for each class. In other words, "If something is far enough, it may be anything." As the object gets closer, the observer gradually updates the class probabilities (say, it looks like a dog, but it could still be anything). Furthermore, if human errors are also considered, it should be kept in mind that the output representations should never be exact. In this respect, ANFIS [4] was an early attempt to break the ice on the quantification of class memberships. ANFIS was proposed for increasing the speed of learning in

backpropagation-based algorithms. However, a pre-encoded knowledge base (rule base and database) requires a deep understanding of the supervision criteria for determining the membership functions. Needless to mention the cost.

In this study, an attempt is made to deal with overfitting by setting up a negotiation between the modelit can be ensured that as the model obtains a better fitness to the labels, it is rewarded with a better position in the negotiation table. Therefore, the model comes to better fitness and does not spend much energy accounting for wrong labels, exceptions (outliers), and membership values that are justified by the quality of the observations. Also, gradually scrutinizing the categorical labels relieves us of endless hours of fitting the data into a paradigm.

The motivation for this study is rooted in the exploration of generic and specific differences in representations [5], as well as Ludwig Wittgenstein's philosophical investigations [8]. When attempting to identify an object, a child must examine the object through various dimensions, such as visual properties (e.g., shape, color, texture, size) and the object's function (e.g., taste, content, or purpose). Nonetheless, not all dimensions are consistently accessible for object recognition, and even when all the required information is available, it is not always feasible to employ every dimension for differentiating between classes. This phenomenon necessitates a careful examination of class memberships from the quality of observation standpoint.

This study aims to tackle the overfitting problem in low data regime machine learning problems by setting up a negotiation between the model's belief of training labels and the labels provided by human supervisors. The proposed method aims to balance the model's assumptions and human input, creating a more robust and accurate classifier, even when dealing with limited data

The organization of the paper is as follows: In the Materials and Methods section, the theoretical foundation of negotiated representations is presented, including a discussion of overfitting prevention strategies and the mathematical formulation of the negotiation paradigm. In the Experiments section, the network architectures and experimental setup for validation are described. In the Results and Discussion section, the performance of the proposed method is evaluated across four benchmark datasets (MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100), and the findings are analyzed. The Conclusion section summarizes the contributions and discusses future research directions.

Materials and Methods

Overfitting and Prevention Strategies

Overfitting [6] is a phenomenon that occurs when the model is trained for too long and focused too much on the exact fitness of the training samples to the provided training labels and cannot keep track of the predictive rules that would be useful on the test data. In the literature, overfitting is commonly attributed to memorization of the samples, noise, and other peculiarities of data samples by using a high number of neurons. While it is true that the model also encodes undesired aspects of the data samples as training process continues, it is argued that most of the overfitting occurs in the process of reconciling sharply defined membership ratios to specific classes.

The loss of the individual differences in hierarchical systems [7] is also of a great significance in understanding representation learning. Although neural networks can be considered hierarchical systems, individual differences may not be lost but transformed into the means of compensation for the inconvenient membership ratios. However, the main concern of the problem is related to the fidelity of the representation to the actual sample, and individual differences should be filtered out for higher representational capacity in the face of a particular objective such as classification.

The third argument that should be discussed is that the certainty of a decision depends highly on the quality of the observation. For instance, let’s assume that we try to recognize a cat or a dog by using a picture given. If the distance of the animal from the camera is long enough to cover all distinctive differences between cats and dogs, we decide that the probabilities are the same. As the camera gets closer to the animal of interest, the distinctive differences start to appear, and the probability of the sample belonging to one of the classes increases due to the increase in the observability. Trying to form an association among poorly represented memberships may cause the network to encode the exceptions and therefore inject specific and undesired noise

into the principal components of the distinction process. To prevent machine learning models from overfitting, several methods have been proposed: dropout [9], L1 [10], L2 [11], and augmentation [3]. To our knowledge, none of the proposed methods in the literature computationally scrutinizes the provided labels in supervised learning.

In this article, it is suggested that a sharp definition of the membership ratios may be the leading source of overfitting. To prevent overfitting, this work proposes enabling the model to negotiate the membership ratios of all samples to all classes by slightly adjusting the provided labels, such as changing a label from 1 to 0.98, to better represent the sample's relationship with the rest of the dataset. To test the hypothesis, several overfitting scenarios were generated to allow the model to compensate for label imprecision. The proposed training paradigm has been tested on publicly available benchmark datasets such as CIFAR-10, CIFAR-100, MNIST, and Fashion-MNIST. To generate a low data regime, a small set of training and test examples for each dataset were selected. The results on all datasets have shown that the negotiation between the model and the provided labels is a powerful method in preventing overfitting.

Negotiated Representations

The general structure of supervised learning requires us to rely on the labels provided by humans beforehand. These provided labels are then used to create a cost function that informs the model on how far off the prediction is. By trying to decrease the cost with the help of backpropagation, the model is expected to encode the underlying input-output relationships into its weights. This logic works extraordinarily well for big data regimes. However, it becomes unreliable in low data regimes as the size of the samples increases. If the model is large enough to ensure the fitness between samples and their respective labels, it tends to encode the aspects of the samples that are irrelevant to the classification process. A neural network can be represented as the mapping function in Equation 1:

$$Y: f(X, \theta, \mathrm{b})$$

Equation 1

where **X** represents the data instances in the training set, **Y** represents ground truth labels, **b** is bias, and **θ** represents the network parameters. The network parameters are updated at each epoch with backpropagation depending on the predicted labels, **Y'=f(X)** and the loss function $L: J(\boldsymbol{Y}, \boldsymbol{Y}')$ where J represents the cost function. The optimization of network parameters is shown as:

$$\theta^* = \arg min \frac{1}{L} \sum_{i=1}^{L} (y_i, y_i')$$

Equation 2

This study presents an attempt to deal with overfitting by setting up a negotiation table between the model

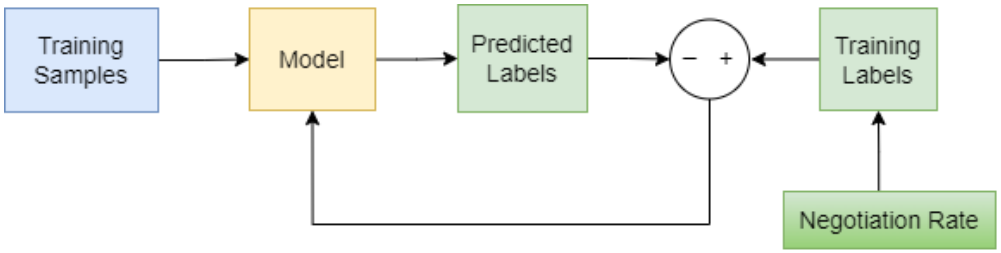


Figure 1. *The model diagram with negotiation rate.*

where negotiated labels are calculated by a weighted average of predicted labels and original labels. When the negotiation rate is included, the loss function is calculated as seen in Equation 3.

$$\boldsymbol{L} = J((1-n).\boldsymbol{Y}, n.\boldsymbol{Y}')$$

Equation 3

Thus, the optimization term shapes as seen in Equation 4.

$$\boldsymbol{\theta}^* = \arg min \frac{1}{L} \sum_{i=1}^{L} J((1-n).\boldsymbol{y}, n.y_i')$$

Equation 4

It should be kept in mind that, at the end of each negotiation phase original labels are switched with negotiated labels that are calculated in that phase. Furthermore, since the model gains more confidence as training continues, the negotiation rate is also linearly increased at the end of each epoch. This change means that the coefficient of the model's predictions will increase and the coefficient of the provided labels will decrease at the end of each negotiation phase. The linear increment in the negotiation rate limits the number of negotiations that take place throughout training. Otherwise, negotiations would arrive at a point where the model's coefficient in the weighted average (n) would be more than 1, and the previously determined labels' coefficient (1-n) would go below zero. A detailed flowchart of the model is given in Figure 2.

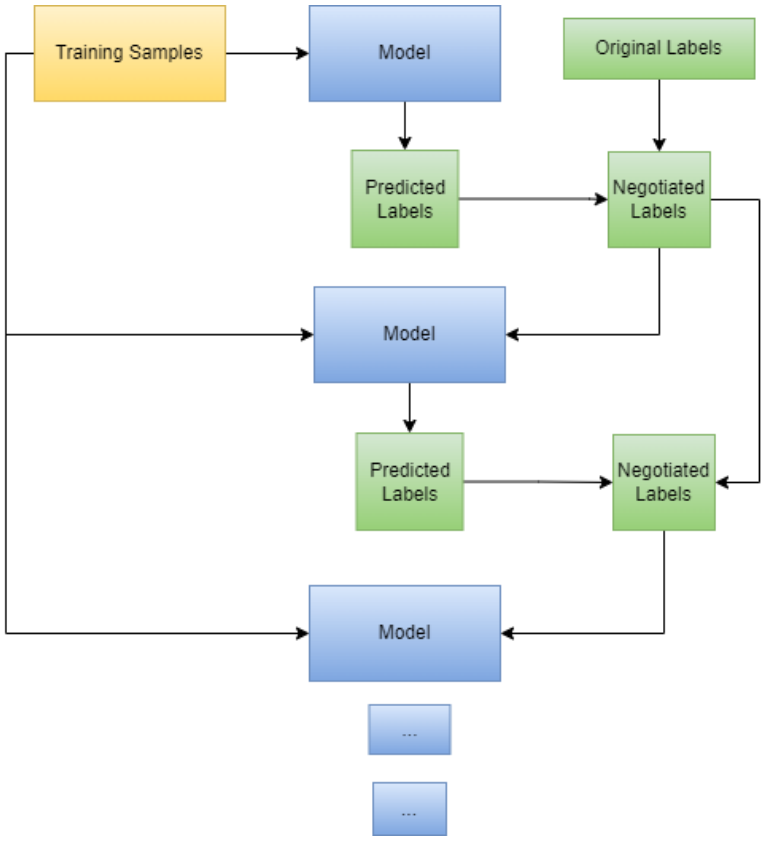


Figure 2. *Flowchart of the model.*

Experiments

To evaluate the performance of the proposed paradigm, four different models were designed. The models were provided with sufficient data to draw meaningful conclusions while limiting the amount of data used for training to induce overfitting. The experimental setups described below serve as a proof of concept and demonstrate the behaviours of the models. One downside of exceptionally high success rates in classifiers is that they can be attributed to the injection of test dataset information into the model through hyperparameter

tuning. Optimizing each part of the model for maximum test performance may also result in encoding many peculiarities of the test dataset within the model. Consequently, building upon any paradigm requires us to reverse the optimization process for a more objective evaluation of the method. For this reason, it was found it more beneficial to focus solely on demonstrating the behaviour of the model throughout the experiments. In the context of investigating overfitting induction and it

For all the convolution layers within the networks, utilization was made of the Rectified Linear Unit (ReLU) activation function, as it offers several advantages, such as reduced likelihood of the vanishing gradient problem and improved training speed. In contrast, for the final fully connected layer in each model, softmax layer was employed, as it enables the output to be constrained between the range of 0 and 1, which is particularly useful for the deployment of the negotiation paradigm in classification tasks.

Results and Discussion

This section presents evidence of the effectiveness of the proposed method for preventing overfitting in the model. First, a comprehensive analysis of the results is provided, including figures and their interpretations. Second, a comparison is made between the baseline model and the model trained with the proposed paradigm. Specifically, Table 1 and Table 2 summarize the performance metrics of the two models. Figure 4 shows that the loss reduction trend mirrors that observed in Figure 6, which confirms the performance of the proposed method. Additionally, Figure 8 and Figure 10 present the results of the model trained on the CIFAR-10 and CIFAR-100 datasets respectively. The findings are promising, even though some aspects remain unexplained. In Table 1, testing accuracy performances of the baseline model and proposed model are demonstrated.

Table 1. *Accuracy comparison of baseline model and the proposed model on test data.*

Dataset	Baseline Model's Accuracy	Proposed Model's Accuracy
MNIST	0.828	0.867
Fashion-MNIST	0.719	0.766
CIFAR-10	0.326	0.343
CIFAR-100	0.460	0.491

The comparison between the losses of the baseline model and the proposed model is demonstrated in Table 2.

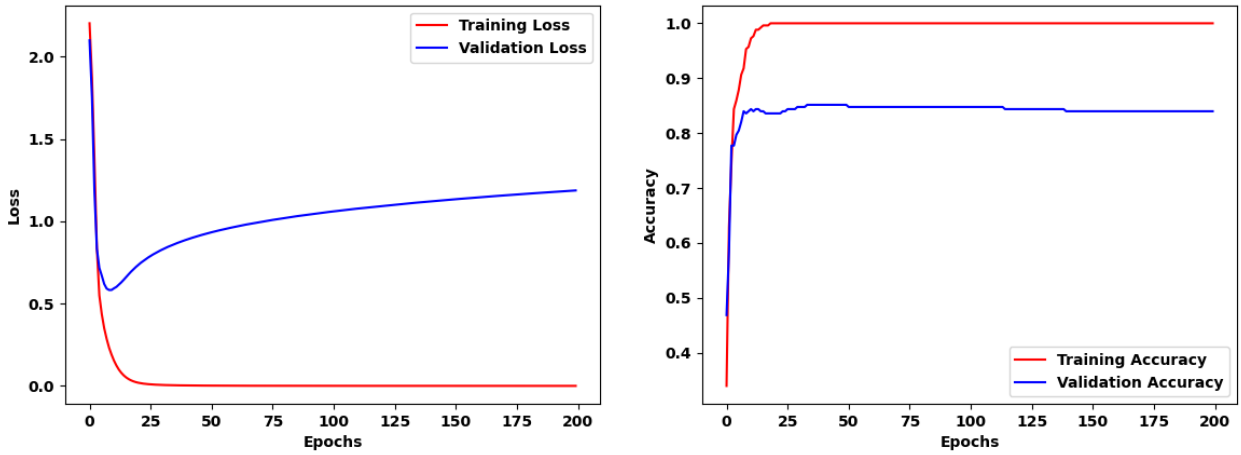
Table 2. *Loss comparison between baseline model and proposed model on test data.*

Dataset	Baseline Model's Loss	Proposed Model's Loss
MNIST	0.92	0.41
Fashion-MNIST	1.94	0.78
CIFAR-10	4.48	2.13
CIFAR-100	13.43	5.18

The proposed method outperforms the baseline model for each dataset. Plots are provided of training and validation accuracies with an increasing number of epochs in later sections to visualize overfitting and model training performances.

MNIST

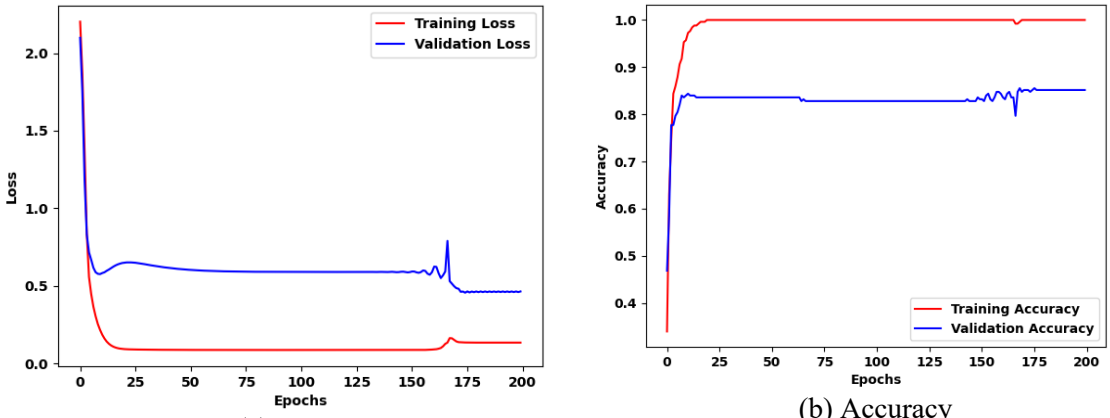
A model was constructed for MNIST dataset that consists of two convolutional layers, each having 32 and 64 filters respectively, followed by a fully connected layer containing ten neurons. The training set consisted of 256 samples, and the test set contained 256 samples. Due to the low number of samples and the simplicity of recognizing digits, it was relatively easy to generate an overfitting scenario. Moreover, there were no high-level relationships that could prevent the model from accurately classifying the samples. Additionally, since the images are single-channeled gray images, there were no color complications. The accuracy and loss values for the model are provided in Figure 3.



(a) Loss (b) Accuracy

Figure 3. Standard Network Performance on MNIST dataset.

As observed in Figure 3-a, the model starts overfitting after around ten epochs. When the proposed method was applied to the model it was observed that the overfitting was reduced and the accuracy was improved as it is seen in Figure 4.

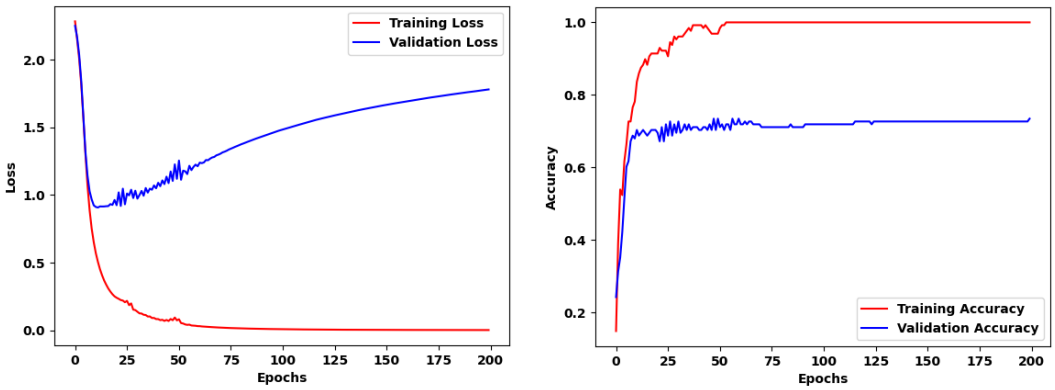


(a) Loss (b) Accuracy

Figure 4. Performance of the Network with Negotiated Representation on MNIST dataset.

Fashion-MNIST

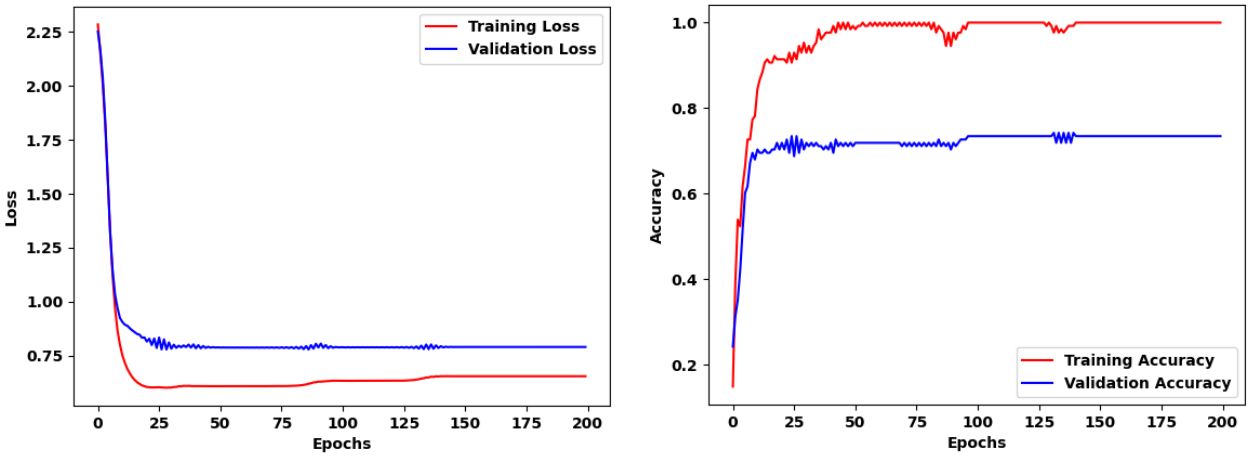
The model for Fashion-MNIST dataset has the same network parameters as were used for training MNIST dataset. The training set consisted of 128 samples, and the test set comprised 128 samples. The accuracy and loss values for the model are provided in Figure 5.



(a) Loss (b) Accuracy

Figure 5. Standard Network Performance on Fashion-MNIST dataset.

As can be noticed from Figure 5-a, the model is heavily over-fitted. The model improves after the proposed negotiation representation regarding loss and accuracy as it is seen in Figure 6.

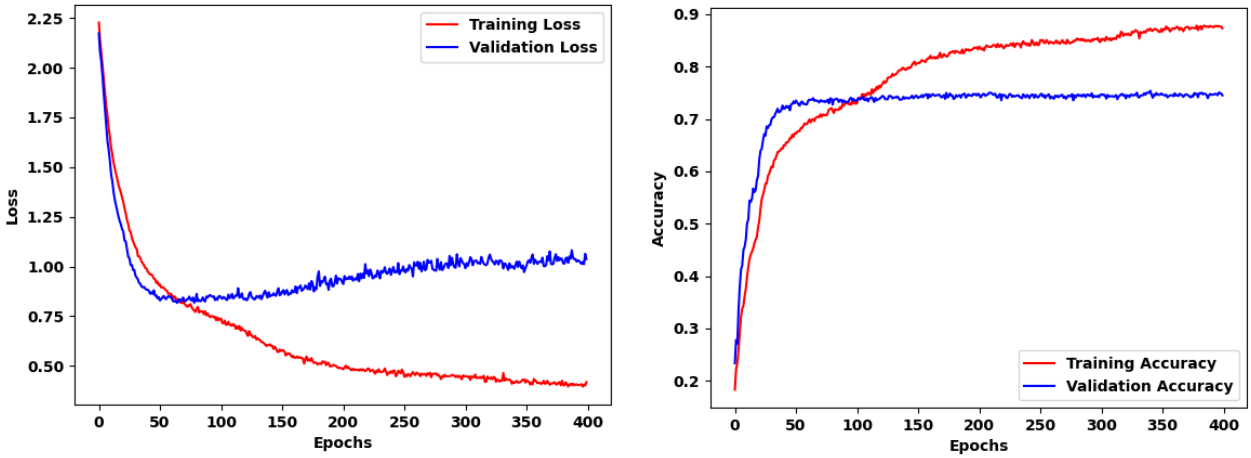


(a) Loss (b) Accuracy

Figure 6. *Performance of Network with Negotiated Representation on Fashion-MNIST dataset.*

CIFAR-10

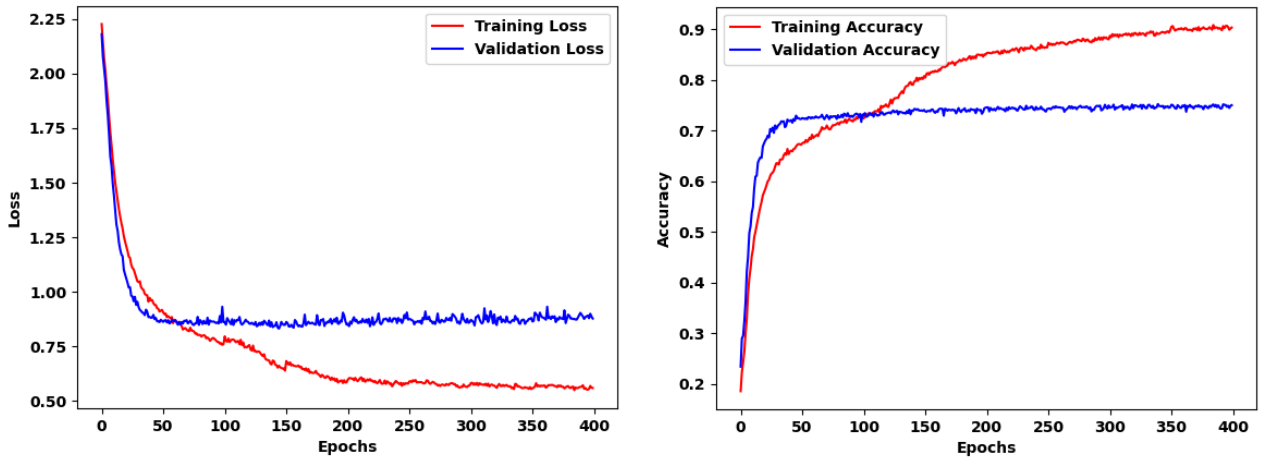
Creating an overfitting scenario for the CIFAR-10 dataset proved to be more challenging than for MNIST and Fashion-MNIST. This increased difficulty can be attributed to the higher-level relationships and color images present in the dataset, resulting in three channels of information per sample, which adds complexity to the learning task, and generating an overfitting scenario with a small network becomes more difficult. Figure 7 shows the loss and accuracy performance of regular training on CIFAR-10 dataset. The observed results clearly demonstrate that the proposed paradigm significantly reduced the validation dataset's loss. However, the increase in test accuracy was relatively minor and not indicative of the improvement in test loss. Nevertheless, any improvement is significant in the context of machine learning.



(a) Loss (b) Accuracy

Figure 7. *Standard Network Performance on CIFAR-10 dataset.*

Like the previous simulations, overfitting is observed by evaluating the loss and accuracy plots. After training with the negotiation paradigm, an improvement was obtained in the loss and accuracy of the model as demonstrated in Figure 8.

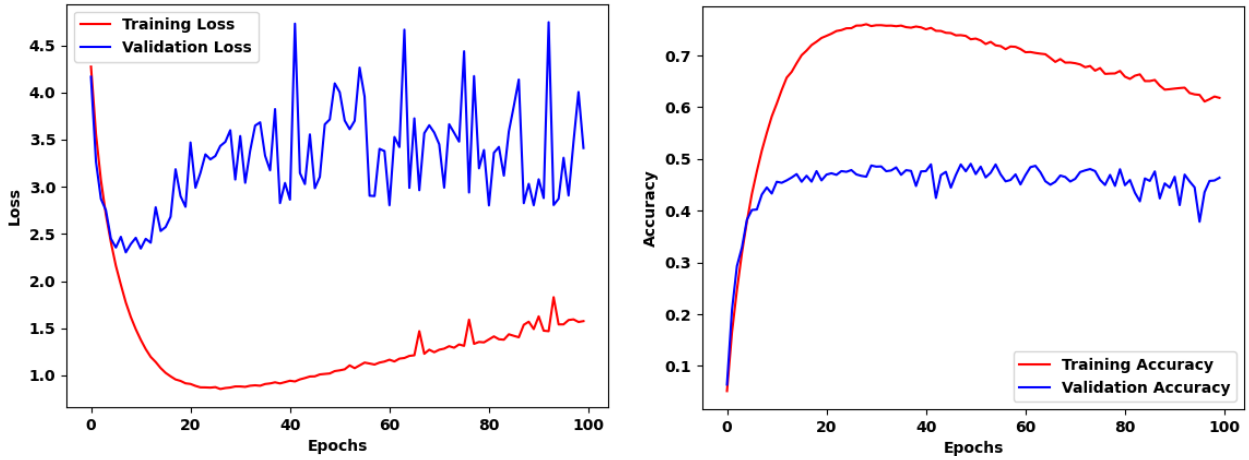


(a) Loss (b) Accuracy

Figure 8. *Performance of Network with Negotiated Representation on CIFAR-10 dataset.*

CIFAR-100

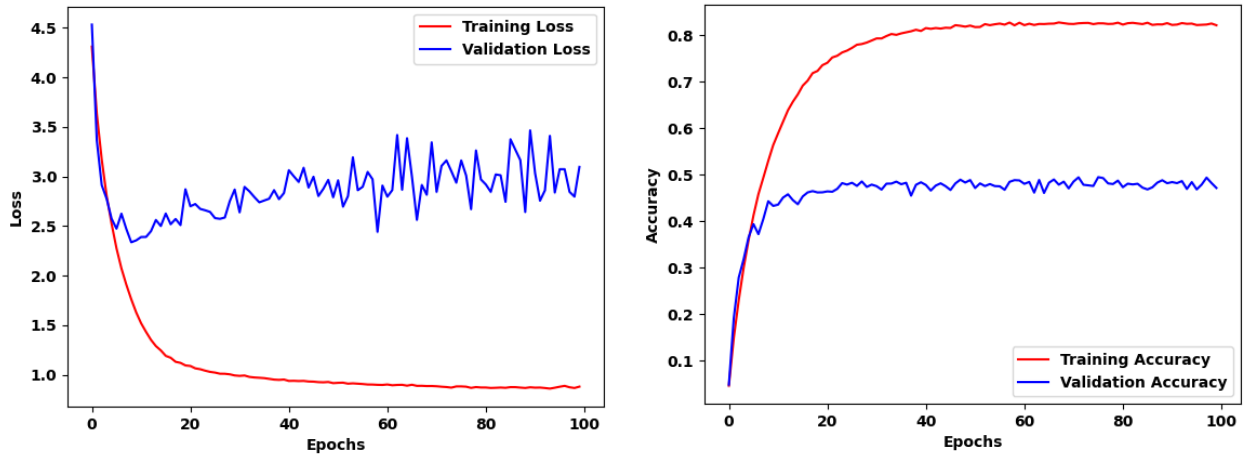
For the CIFAR-100 dataset, a more complex model was designed due to the large number of classes in the dataset. The model comprises six convolutional layers, each containing 64, 64, 128, 128, 256, and 256 filters, respectively, followed by a fully connected layer with 512 neurons and a softmax layer. The training set included 45,000 samples, while the test set consisted of 5,000 samples. A more extensive model is necessary for managing the increased number of classes, and fitting such a model requires a larger dataset. However, a larger dataset can make achieving fitness more challenging. Our choice of model and dataset size was based on these considerations, as our objective was to first generate overfitting and then mitigate it using our proposed paradigm.



(a) Loss (b) Accuracy

Figure 9. *Standard Network Performance on CIFAR-100 dataset.*

Generating an overfitting scenario for the CIFAR-100 dataset proved more difficult than for other data sets due to the large number of classes, complex relationships within the data, and the use of coloured images containing three channels of information. In the simulations for the CIFAR-100 dataset, it was found that decreasing the loss was relatively manageable while improving the accuracy was more challenging. Consequently, representational fidelity does not always guarantee high accuracy. In this scenario, overfitting is unavoidable. To mitigate overfitting, dropout and max pooling layers were incorporated. Figure 9 shows the loss and accuracy performance of the model. Figure 9 shows an obvious overfitting after a few epochs. Overfitting is reduced significantly after applying the negotiation representation as it is seen in Figure 10.



(a) Loss (b) Accuracy

Figure 10. *Performance of Network with Negotiated Representation on CIFAR-100 dataset.*

Conclusion

In this study, a novel algorithm has been presented to mitigate overfitting in classification tasks, particularly in low data regimes. The method has been applied to several data sets, including MNIST, Fashion-MNIST, CIFAR 10, and CIFAR 100, demonstrating its potential to address a broad range of low data regime challenges. The success of the method, however, is dependent on the negotiation rate, and further research is required to investigate the relationship between the dataset and the optimal negotiation rate for the best performance. The aim of this study is to draw the attention of the machine learning community towards developing novel methods for justifying assigned labels. This work proposes that the discrepancy between training and test loss could stem from the fact that the provided labels are not adequately justified by the characteristics of the observations. The justification will likely be context dependent. Considering the context of the dataset, each deviation from the most optimal representations should be injected into labels as class memberships. Doing so will enhance the model's performance and provide a more philosophically sound justification for deep learning applications. This study also suggests that the negotiated learning paradigm holds great promise for continual learning, offering a more efficient, intuitive, and sustainable approach compared to current methods in the literature [11]. By injecting the model's past experiences into future labels, one can potentially mitigate catastrophic forgetting to a new degree without compromising the plasticity of the model or relying on memory intensive replay scenarios. It can also be coupled with existing paradigms to update the state-of-the-art performances.

Acknowledgement

I would like to express my sincere gratitude to my dear friend Samet Bayram for his significant contributions and collaboration on the initial preprint of this study [14].

References

[1] LeCun Y, Bengio Y, Hinton G, et al. (2015). Deep learning. Nature. 521(7553): 436–444.

[2] Li H, Li J, Guan X, Liang B, Lai Y, Luo X. (2019). Research on overfitting of deep learning. In: Proceedings of the 15th International Conference on Computational Intelligence and Security (CIS). IEEE, pp 78–81.

[3] Balestriero, R., Bottou, L., & LeCun, Y. (2022). The effects of regularization and data augmentation are class dependent. Advances in Neural Information Processing Systems, 35, 37878-37891.

[4] Jang J-SR. (1993). ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics. 23(3): 665–685.

[5] Gilles, D., & Paul, P. (1995). Repetition and Difference. Trans. Paul Patton. New York: Columbia University Press.

[6] Dietterich TG. (1995). Overfitting and undercomputing in machine learning. ACM Computing Surveys. 27(3): 326–327.

[7] Deleuze, G., & Guattari, F. (1977). Capitalism and schizophrenia (Vol. 1). New York, NY: Viking Press.

[8] Wittgenstein, L., & Anscombe, G. E. M. (1963). Philosophical Investigations/Ludwig Wittgenstein. NY: Macmillan,[1953].–232 p.–(Philosophische Untersuchungen. English and German).

[9] Park MY, Hastie T. (2007). L1-regularization path algorithm for generalized linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 69(4): 659–677.

[10] Cortes C, Mohri M, Rostamizadeh A. (2012). L2 regularization for learning kernels. arXiv preprint arXiv:1205.2653.

[11] Salami B, Haataja K, Toivanen P. (2021). State-of-the-art techniques in artificial intelligence for continual learning: a review. In: Proceedings of the Federated Conference on Computer Science and Information Systems (FedCSIS). pp 23–32.

[12] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research. 15(1): 1929–1958.

[13] Ioffe S, Szegedy C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning (ICML). pp 448–456.

[14] Korhan, N., & Bayram, S. (2023). Negotiated Representations to Prevent Overfitting in Machine Learning Applications. arXiv preprint arXiv: 2311.11410.