Comparing Uncertainty Estimation Methods in Deep Neural Networks

by

Mehmet Arın Zeyneloğlu

Submitted to the Graduate School of Engineering and Natural Sciences
in partial fulfillment of
the requirements for the degree of
Master of Science

Sabancı University
December 2023

Acknowledgements

I would like to express my sincere gratitude to my thesis advisor, Prof. Dr. Berrin Yanıkoğlu. I am immensely grateful for her invaluable mentorship, patience, and encouragement throughout the entirety of this research endeavor. Berrin Hocam, I am both incredibly impressed and somewhat daunted by your profound intuition and deep expertise in this field. I consider myself fortunate to have absorbed even a small fraction of that knowledge.

I am immensely thankful to my family for their continuous support and encouragement. Their unwavering belief in my ability to fulfill my academic ambitions has been a constant source of inspiration, a value whose importance has only grown with time. Special thanks to my dear sister Hazal, who graciously agreed to lower the TV volume while I was writing this thesis, something she never does. I also owe a great deal of gratitude to my soon-to-be wife Begüm - whose unwavering support, and understanding have been the cornerstone of my strength throughout my master's journey - and to my friends Oğuz Eroğlu, Elif Naz Özdamar and Fırat Kızılırmak for their emotional support and encouragement during this academic pursuit.

Thank you all for being an essential part of my journey and for making this thesis possible.

Comparing Uncertainty Estimation Methods in Deep Neural Networks

Arın Zeyneloğlu

Data Science, Master's Thesis, 2023

Thesis Supervisor: Berrin YANIKOĞLU

Keywords: uncertainty estimation, cnn, evidential deep learning, monte carlo dropout, deep ensemble networks, rejection option, mislabel correction

Abstract

Convolutional Neural Networks (CNNs) is one of the mainstream paradigms in most computer vision tasks. Accurately quantifying the uncertainty in CNN's predictions is crucial as they are being used in various applications, including safety- critical domains such as medical image classification and autonomous driving. Yet, uncertainty prediction remains a challenge. Softmax probabilities are often used to model uncertainty with no solid support. Recent studies have tackled this challenge using three distinct methodologies, namely: Monte Carlo Dropout, Deep Ensembles, and Evidential Deep Learning (EDL). Although this thesis primarily focuses on EDL, the most up-to-date and computationally efficient among these approaches, each of these methods performance in uncertainty estimation along with their predictive capabilities are compared using CIFAR-10 and CelebA datasets in this work. Finally, leveraging the EDL method on the CelebA dataset, a novel approach is presented to automatically detect mislabeled samples within the dataset.

Derin Sinir Ağlarında Kullanılan Belirsizlik Ölçümleme Yöntemlerinin Karşılaştırılması

Arin Zeyneloglu

Veri Bilimi, Yüksek Lisans Tezi, 2023

Tez danışmanı: Berrin YANIKOGLU

Anahtar Kelimeler: belirslizlik ölçümleme, evrişimsel sinir ağları, monte carlo dropout, deep ensemble networks, ret seçeneği, yanlış etiketleme düzeltmesi

Özet

Evrişimsel Sinir Ağları görüntü işleme uygulamalarında en yaygın olarak kullanılan yöntemlerden biridir. Evrişimsel Sinir Ağları' nın tahminlerindeki belirsizliğin doğru bir şekilde ölçülmesi, bu yöntemin tıbbi görüntü sınıflandırması ve otonom sürüş gibi güvenlik açısından kiritik alanlar da dahil olmak üzere çeşitli uygulamalarda yaygın olarak kullanılması nedeniyle çok önemlidir. Buna rağmen, belirsizlik tahmini hala tam olarak çözülemeyen bir problem olarak kalmaya devam etmektedir. Bu yönde herhangi bir somut kanıt olmamasına rağmen softmax olasılıkları genellikle belirsizliği modellemek için kullanılmaktadır. Güncel araştırmalar belirsizlik ölçümleme problemini, Monte Carlo Dropout, Deep Ensembles ve Evidential Deep Learning (EDL) isimli üç farklı strateji kullanarak ele almıştır. Bu tez öncelikli olarak, belirtilen yaklaşımlar arasında en güncel ve hesaplama açısından en verimli olan EDL' ye odaklanmış olsa da, bu yöntemlerin her birinin belirsizlik ölçümlemedeki performansı ve tahminleme yetenekleri bu çalışmada CIFAR-10 ve CelebA veri setleri kullanılarak karşılaştırılmıştır. Son olarak, LFWA veri seti üzerinde EDL yönteminden yararlanılarak veri seti içerisinde yanlış etiketlenmiş örneklerin otomatik olarak tespit edilmesi için yeni bir yaklaşım sunulmaktadır.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Convolutional Neural Networks (CNNs) are one of the mainstream paradigms in most computer vision tasks. Accurately quantifying the uncertainty in CNN's predictions is crucial as they are being used in various applications, including safety-critical domains such as medical image classification and autonomous driving. Yet, uncertainty prediction remains a challenge. Softmax probabilities are often used to model uncertainty with no solid support.

While the importance of accurately quantifying uncertainty in CNNs for safety-critical applications has been widely acknowledged, the existing reliance on softmax probabilities has proven inadequate. Recent studies have aimed to tackle the demand for reliable uncertainty estimation by exploring Bayesian approaches and second - order probabilistic frameworks such as Evidential Deep Learning (EDL).

In a study done in 2016, it has been shown that the mean softmax probability of incorrectly classified samples was found to be greater than 0.80 on three different datasets [1]. Due to the limitations of frequentist methods to estimate uncertainty, Bayesian approaches have gained significant traction. In [2], each weight of the network is represented as a probability distribution rather than a point estimate. To approximate the predictive distribution, ensembles of different networks are used in [3] which it is computationally inefficient due to the nature of ensemble techniques. In [4], dropout is applied during testing which is computationally taxing since it requires multiple passes through the network for each data sample. As a remedy to drawbacks of mentioned techniques, Sensoy et al. [5] explicitly modeled the predic-

tions of the network as a Dirichlet distribution defined over the network outputs, by learning the parameters of the Dirichlet distribution from data. We compare these methods using public datasets that are widely used in computer vision.

Furthermore, the majority of the works in uncertainty quantification literature use small and simple benchmark datasets, such as MNIST[1] and CIFAR-10[2]. Although promising results have been obtained from these studies, the application of uncertainty quantification methods to more extensive and intricate datasets remains underexplored.

Another contribution of this thesis is to evaluate the EDL framework, the most recent and computationally efficient among the ones studied in this thesis, on two of the widely used datasets for face attribute classification, CelebA[3] and LFWA[4]. Identifying attributes from face images has been a key area of research in recent years, as it enables practical applications such as attribute based searching and video surveillance. Despite the significant improvement that has been achieved in terms of predictive performance, uncertainty estimation remains unexplored in the domain of face attribute classification. In this paper, we tackle this problem by using the Evidential Deep Learning (EDL) framework presented in [5]. The underlying model is a convolutional neural network that is trained to learn the parameters of an evidential distribution, which models a second-order probability distribution over class probabilities.

The main contributions and findings of this paper are summarized in the following:

1. We conduct comprehensive quantitative analysis on the CIFAR-10 dataset using the most commonly used approaches of uncertainty quantification, namely, Monte Carlo Dropout, Deep Ensembles and Evidential Deep Learning. We compare the ability of mentioned methods in estimating the uncertainty by:

    - Analyzing the resulting uncertainty distributions for correctly classified and misclassified samples
    - Providing rejection option based on confidence metrics obtained from each approach
    - Providing accuracy vs uncertainty curve for all uncertainty values ob-

tained in the test to assess the relationship between accuracy and confidence

2. We conduct extensive experimental analysis, both quantitatively and qualitatively, on two of the widely used datasets for face attributes classification, i.e. CelebA and LFWA

    - Applying the EDL framework to face attribute classification to estimate the uncertainty in the output predictions.
    - Illustrating the effectiveness of the proposed framework on the widely used face attributes benchmark, i.e. CelebA
    - Comparing the predictive performance of the EDL framework with the traditional softmax-based approach on 40 distinct attributes provided in CelebA
    - Demonstrating the ability of the proposed approach in estimating the uncertainty via extensive analyses and showing potential use cases
    - Utilizing the EDL framework, we introduce a novel approach to automatically detect mislabeled samples within the dataset

Chapter 2

Related Works

The problem of uncertainty quantification in deep learning have been studied for years, since "knowing what a model does not know" is a longstanding challenge, especially today with the increase in deep learning applications in our everyday lives.

In recent years, the field has seen significant advancement with approaches that seeks to approximate the posterior distribution over the parameters of the network through means of sampling. 2 commonly employed techniques in this category are Monte Carlo Dropout(MC- Dropout) [5] and Deep Ensembles [6]. Although they are popular and relatively easy to implement, these methods are computationally demanding as they require making multiple passes through the network for each data sample.

Recently, Evidential Deep Learning [7] has gained significant traction since it seeks to learn the parameters of the posterior distribution directly rather than through sampling, which reduces the computational complexity remarkably. We elaborate on the mentioned techniques in the following.

2.1 Monte Carlo Dropout

Dropout[8] is a straightforward yet effective technique used in deep learning to prevent overfitting. Essentially, dropout works by randomly deactivating a subset of neurons in the neural network based on a bernouilli random variable with a

probability p during training. This randomness ensures that the model does not become overly dependent on any specific set of neurons, thereby encouraging the network to learn more generalized features. Since it is used to combat overfitting, dropout is only used in the training process, and it is turned off during inference.

During the training process, deterministic Neural Networks learn a fixed set of weights, W, whereas Bayesian Neural Networks try to learn the posterior distribution over weights, i.e. distribution of the network parameters given input X and ground truth Y as depicted in the equation 2.1.

$$P(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}) = \frac{P(\mathbf{Y} \mid \mathbf{X}, \mathbf{W})P(\mathbf{W})}{P(\mathbf{Y} \mid \mathbf{X})} \tag{2.1}$$

Knowing this posterior distribution allows us to obtain the predictive distribution for a new, unseen test sample (denoted as **x***) by performing an integration over the posterior distribution (equation 2.2). The variance or entropy of the resulting predictive distribution can be considered as prediction uncertainty for that sample. Unfortunately, the posterior distribution is intractable to compute.

$$P(y^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int P(y^*|\mathbf{x}^*, \mathbf{W})P(\mathbf{W}|\mathbf{X}, \mathbf{Y})dw \tag{2.2}$$

Monte Carlo Dropout (MC Dropout) aims to approximate the posterior predictive distribution through sampling. The main idea behind the approach is to make multiple stochastic passes through the neural network for each sample using a different sample of weights. At its core, MC Dropout extends the dropout technique to the inference phase by utilizing the dropout mechanism to obtain multiple samples during testing. More precisely, during inference, each sample is fed into the network N times (usually taken as 100) with dropout activated so that a different output scores are produced by the network for each of these passes. This process effectively simulates sampling from a probabilistic model, creating a distribution of outputs for each input.

For classification tasks, the uncertainty can be calculated as the variance or more commonly as the entropy, which is also the preferred method in this study, across the softmax output probabilities from all forward passes.A high variance indicates

that the model is less certain about its prediction, whereas a low variance suggests higher confidence.

2.2 Deep Ensembles

Similar to MC Dropout approach described above, Deep Ensembles aims to approximate the posterior distribution by means of sampling. However, in this case, using an ensemble of independently trained networks each learning a unique sample of weights. In other words, the deep ensemble approach involves constructing an ensemble of K DNNs as M = $[M_i]_{i=1}^K$, where each DNN, i.e. M_i, is characterized by different architectural configurations.

In parallel with the MC Dropout approach, we have an ensemble of prediction distribution, $[p(y \mid x, M_i]_{i=1}^K$, which can be utilized to measure the uncertainty by computing the average entropy [9]. Intricacies regarding entropy calculation are elaborated in section 4.4.

However, it is worth mentioning the biggest drawback of the deep ensemble approach: it is very demanding in terms of both computation and memory. This challenge make the approach infeasible for real-world applications.

2.3 Evidential Deep Learning

Recently, the EDL framework [7], which is an extension of Dempster-Shafer Theory [10] and Subjective Logic (SL) [11], is proposed for use with neural networks. EDL constructs its learning objective as an evidence-gathering process by applying a Dirichlet prior defined over network outputs. It attempts to overcome the limitations of softmax-based CNNs by probabilistically estimating the predictive distribution of the network. In this setting, model outputs are interpreted as a probability distribution rather than a point estimate resulting from the traditional softmax-based approach.

For a K-class classification task with mutually exclusive classes and input x_i, the loss function is defined using the sum squares loss and the Dirichlet prior [7]:

$$\begin{aligned}\mathcal{L}_i(\theta) &= \int \|\mathbf{y}_i - \mathbf{p}_i\|_2^2 \frac{1}{\beta(\alpha_i)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d\mathbf{p}_i \\ &= \sum_{j=1}^{K} y_{ij} \left[y_{ij}^2 - 2y_i p_i + p_{ij}^2\right] \\ &= \sum_{j=1}^{K} y_{ij} \left(y_{ij}^2 - 2y_{ij}[p_{ij}] + [p_{ij}^2]\right) \end{aligned} \tag{2.3}$$

where $\mathbf{y_i}$ is the one-hot encoded target; $\mathbf{p_i}$ represents the assigned class probabilities and α_{ij} are the parameters of the evidential Dirichlet distribution.

Additionally, a Kullback-Leibler (KL) divergence term is introduced into the loss function to minimize the evidence for incorrectly classified samples, resulting in the following total loss:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \mathcal{L}_i(\theta) + \lambda_t \sum_{i=1}^{N} KL\left[D\left(p_i \mid \widetilde{\alpha}_i\right) \parallel D\left(p_i \mid \mathbf{1}\right)\right] \tag{2.4}$$

where λ_t refers to the annealing coefficient which increases the effect of the KL divergence throughout training; $D(p_i \mid \mathbf{1})$ denotes the uniform Dirichlet distribution; and $\tilde{\alpha}_i = y_i + (1 - y_i) \odot a_i$, with element-wise product, denoted by $\odot$.

Once the parameters of the Dirichlet distribution α_{ij} are obtained by applying ReLu activation function to the network outputs for input x_i, one can calculate S_i which is the Dirichlet strength, defined as $S = \sum_{k=1}^{K} \alpha_k$ and the uncertainty u, which is calculated as $u = \frac{K}{S}$ where K is the number of classes.

The modification to train a binary classifier with EDL only requires modifying the loss function, as in Eq. 2.3 and applying ReLU activation function at the output to keep the parameters of the Dirichlet distribution non-negative.

Chapter 3

Methodology

This chapter outlines the evaluation approaches used in this thesis. We first present the datasets used in our study in detail. Our strategy to analyze uncertainties is separated into three sections, namely, Comparative Analysis of Uncertainty Methods (Sec. 3.2), Qualitative Analysis of EDL Uncertainties (Sec. 3.3), and Finding Errors in Ground Truth (Sec. 3.4).

It is worth noting that, while alternative methods for uncertainty estimation have been developed and evaluated for comparison purposes, the primary emphasis of this thesis lies on EDL, which is the most recent and computationally efficient approach among those discussed. Thus, EDL is specifically employed for methods 3.3 and 3.4.

3.1 Datasets

The CIFAR-10 dataset, which is a standardized benchmark widely used in the evaluation of CNNs due to its moderate size and diversity, is utilized for comparative analysis between different uncertainty estimation methods. It consists of a total of 60k images across 10 different classes. The dataset is partitioned into training and test sets, consisting of 50k images for training and 10k images for testing purposes. Further information about the dataset is provided below:

- Number of Images: 60,000
- Image Dimensions: 32x32 pixels

- Classes: 10 (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck)
- Distribution: 6,000 images per class

In order to explore the capabilities of uncertainty estimation by the EDL framework, which is the main focus of this thesis, we conduct extensive experimental analysis on two of the widely used datasets for face attributes classification, namely, CelebA [3] and LFWA [3]. CelebA dataset [3] is the largest facial attribute dataset to date with more than 200k images, introduced by Liu et al. [3] in 2015. The dataset consists of a total of $202,599$ images with a train, validation, and test split of $162,770$ images, $19,867$ images, and $19,962$ images respectively. The dataset provides 40 binary attributes for each image. CelebA dataset is used to demonstrate the performance of the system compared to the baseline.

LFWA dataset (Labelled Faces in the Wild) [3] is a medium-sized facial attribute where the same 40 attributes as in CelebA are annotated. It consists of $13,243$ images belonging to $5,749$ subjects, whose pictures are collected from the web. The train-test splits are given, where $6,263$ images are used for training and $6,980$ images are utilized for evaluation. LFWA is used to demonstrate the use of uncertainties in catching labeling mistakes due to its relatively small size, which allows manual inspection.

3.2 Comparative Analysis of Uncertainty Quantification Methods

In this section, we explore our approach to evaluate and compare uncertainty estimation methods. The following subsections cover our strategies for analyzing the uncertainty metrics obtained from each respective method and outline our approach to their comparative analysis.

3.2.1 Softmax Output as a Confidence Score

Softmax probabilities are often used as model confidence. However, due to the exponential function employed in the softmax layer, the trained deep neural network

often produces high confidence scores even for misclassified samples, as studied extensively in [5, 12, 13]. In fact, in one study, it has been shown that the mean softmax probability of incorrectly classified samples was found to be greater than 0.80 on three different datasets [12].

We employed two approaches, similar to presented above, to showcase the drawbacks of using the softmax output as a metric of uncertainty, using the CIFAR-10 and CelebA datasets:

- Measuring the mean softmax output on wrongly classified samples in the test set
- Illustrating the distrubution of softmax probabilities on wrongly classified samples to highlight high scores for the misclassified samples

Experiments belonging to these two approaches are discussed in 4.2

3.2.2 Dropout as a Bayesian Approximation

The approach presented in [5], commonly known as **Monte Carlo Dropout**. The idea is to make multiple stochastic passes for each sample during inference time, with different sample of the weights. This approach leverages the dropout mechanism during both training and testing phases, enabling the generation of multiple stochastic predictions by activating dropout at test time.

We analyze the capabilities of this method as:

- Depicting the relationship between predictive mean resulted from N stochastic passes and entropy
- Demonstrating the test set accuracy versus reject ratio based on decreasing entropy values

The experiments associated with these methodologies will be addressed in Section 4.2.

3.2.3 Deep Ensembles

Similar to Monte Carlo Dropout approach, uncertainty estimation using Deep Ensembles approximate the predictive distribution through sampling. The difference lies in the sampling method. Unlike Monte Carlo Dropout approach which utilize dropout layers, Deep Ensembles leverage independently trained neural networks by aggregating their predictions to capture model uncertainty. We adopt the same strategy for assessing the approach's effectiveness in measuring uncertainty as used in the MC Dropout method.

3.2.4 Evidential Deep Learning

As discussed in detail in Section 2.3, EDL treats its learning objective as an evidence acquisition process by establishing a higher order distribution, i.e. evidential distribution, over the initial likelihood parameters of the network. The modification to train a classifier with EDL only requires modifying the loss function, as in equation 2.3 and applying ReLU activation function at the output to keep the parameters of the Dirichlet distribution non-negative.

For the purpose of comparing the uncertainties quantified by EDL approach, we conducted another analysis of the correlation between varying reject ratios, this time derived from uncertainties quantified by the EDL framework, and overall accuracy calculated on the test set of CIFAR-10.

We also conducted further experiments on CelebA dataset to investigate the relationship between predictive performance of the model and characteristics of uncertainty distribution estimated for different attributes belonging to the dataset. For selected few attributes with varying predictive accuracies, mean and standard deviation of uncertainty distributions are calculated on CelebA test set for positive and negative class proportions and presented in Table 4.6. The results derived from these strategies are thoroughly discussed in Section 4.2

3.3 Qualitative Analysis of EDL Uncertainties

In order to gain insights into the characteristics of samples deemed uncertain by EDL network, we conduct extensive qualitative analysis on both the training set and test set of CelebA.

First, samples in the training set are sorted by their respective uncertainty estimations. Our intuition was that samples with the highest uncertainties might possess out-of-distribution traits specific to the dataset. Results of the experiment discussed in section 4.3.3 which validated our hypothesis.

Second, samples in the test are investigated. However, in this instance, we focused on the samples associated with the smallest uncertainties. Our experiments reveal that a significant portion of these samples entail labeling errors, as expounded in Section 4.3.3. This observation lays the groundwork for our approach to identifying mislabeled samples within the training set, which is explored in the following section.

3.4 Finding Errors in Ground Truth

Many datasets have errors in the ground-truth label. Specifically, the LFWA dataset contains quite a lot of label mistakes. To see if uncertainties can be used to spot dataset label errors, we found the label mistakes in the training split of the dataset for the Male attribute. Relabeling was done only when the label error was clear and resulted in 416 new labels, out of 6263 samples.

Motivated by the observations provided in section 3.3, we plotted the ratio of ground-truth mistakes (true positives) that are caught when using different uncertainty thresholds, together with the corresponding false positive rate. The outcomes are discussed in section 4.4.1 by highlighting the ratio of ground truth mistakes caught by this approach and ratio of the false positive rate obtained by our method.

Chapter 4

Experiments

In this chapter, we elaborate on the experiments we have conducted, including architectural details, optimization techniques, training strategy, data augmentation and results. The intricacies of the training process are covered in section 4.1 while subsequent sections delve into the results.

Results are grouped into 2 main sections, Comparative Analysis of Uncertainty Methods on CIFAR-10 and Analysis of EDL Framework on CelebA & LFWA. The first section evaluates the effectiveness of different approaches in uncertainty estimation. The latter section delves into in-depth quantitative and qualitative experiments conducted on CelebA and LFWA datasets. These experiments include: (1) Accuracy comparison of the EDL against the traditional softmax based classifiers (2) Qualitative analysis to understand the characteristics of the samples that the model deems uncertain (3) Quantitative analysis to assess the weaknesses of the system in regard to class imbalance (4) Iplementing a rejection option based on uncertainty (5) Finally introducing our approach to identify ground truth errors using the EDL framework.

4.1 Experimental Setup

In this section, we provide our choice of the network architectures for each method, along with the details of our implementation. We also describe the data augmentation techniques and the training strategy we employed.

4.1.1 Network Architecture and Implementation

In all experiments, we adopted the pre-trained ResNet-50 model as the backbone feature extractor due to its relatively small size and good performance, including the state-of-art results obtained in face attribute classification problem [14]. For the experiments conducted using CIFAR-10, which consists of images of size 32x32, kernel sizes of convolutional blocks are adjusted to smaller dimensions. ResNet-50 was initially trained on ImageNet with images of size 224x224, therefore the kernel sizes and strides were designed accordingly. Main reason for our decision to reduce kernel sizes is to maintain larger feature map within the network, enhancing its descriptive capability, hence increasing its performance. It is worth noting that if the original kernel sizes are used, the size of the feature maps are reduced to 8x8 only after the 1st convolutional block. Modifications to networks convolutional parameters are summarized in Table 4.1

Table 4.1: Convolutional Settings of the original ResNet-50 and our adapted version for CIFAR-10

	Original	CIFAR-10 Adjusted
Kernel Size	7	3
Stride	2	1
Padding	3	1

As for the experiments involving CelebA dataset, original Resnet50 architecture is used without any alterations. The model is pre-trained on the ILSVRC 2012 dataset [15] with 1.2 million labeled images of $1,000$ object classes.

Pytorch is used as the deep learning development framework and each model is trained on a single Tesla V100 16GB graphics processing unit (GPU). Details regarding training strategy and implementation specifics are provided in the following for each method.

Monte Carlo Dropout.

A dropout layer with a dropout rate of 0.1 is introduced after each ReLU non-linearity presented in our adjusted version of ResNet-50. The model is trained using Adam optimizer [16], with a batch size of 128, and learning rate of $4e^{-2}$. ReduceLRonPlateau with a patience value of 10 is utilized as a scheduler and the network is trained for 50 epochs.

Deep Ensembles.

ResNet-50, ResNet-34, and ResNet-18 are utilized in deep ensembles to achieve a satisfying classification accuracy as well as powerful uncertainty representation. All models are trained using the Adam optimizer, with a batch size of 128, and learning rate of $3e^{-4}$. Once again, ReduceLRonPlateau with a patience value of 10 is used as a scheduler during training and all networks are trained for 50 epochs with an early stopping criteria.

EDL.

The system proposed in this work uses the EDL framework proposed in [7] and summarized in Section 2.3.

For EDL implementation, the softmax layer of the network is replaced with ReLU non-linearity and the loss function given in equation 2.3 is used in all experiments. For experiments involving CIFAR-10 we use 10 output nodes for each class, whereas for tests conducted using CelebA dataset, we use two output nodes per binary attribute (K = 2) and train 40 models for the 40 binary attributes separately.

As for the model optimization, we trained each model using Adam [16] optimizer with batch size of 128, learning rate of 3e4 and default momentum coefficients of (0.9, 0.999).

4.1.2 Data Augmentation

Deep Neural Networks are often characterized by a vast quantity of free parameters, amounting to hundreds of billions if we consider the recent LLMs, rendering them prone to overfitting. A common strategy to mitigate this issue is through the application of data augmentation techniques. Recently, a variety of sophisticated approaches for image augmentation have been developed, among which the procedure called RandAugment detailed in [17] was selected as our preferred augmentation policy for CelebA dataset following comparative evaluations with other alternatives. In addition to RandAugment, we employed an augmentation technique named Random Erasing, in which the pixel values of a random rectangle region in the input image are replaced with random values as elaborated in [18].

A less aggressive data augmentation is employed for experiments performed on the CIFAR-10 dataset, which can be considered as a small dataset, especially in today's standards. We observed that a more conservative augmentation techniques yield more stable uncertainty values for small datasets. Consequently, we employ the following straightforward yet efficient data augmentation methods for models trained on CIFAR-10: (1) Random Horizontal Flip: images are flipped horizontally with a probability of 0.3. (2) Random Affine Transformation: an affine transformation with degrees ranging from -3 to +3, horizontal and vertical shift in the range [-3.2, +3.2], and finally with a scaling factor ranging from 0.8 to 1.2 is applied to input images randomly with a probability of 0.4. (3) Color Jitter: Hue property of input images are jittered using a hue factor chosen uniformly from [-0.5, +0.5], while saturation is jittered with a factor chosen uniformly from [0.5, 1.5].

4.2 Comparative Analysis of Uncertainty Methods on CIFAR-10

In this section, we first provide an intuition behind fundamental problems associated with treating softmax outputs as a confidence metric. This exploration includes not just a theoretical discussion, but is also reinforced with empirical evidence obtained on the CIFAR-10 dataset. Then, the uncertainty estimation methods employed in this study are analyzed individually, culminating in a comparison that highlights the relationship between uncertainty vs accuracy.

It is important to highlight that our objective is not to achieve the state-of-the-art predictive performance in these problems, but rather to assess the effectiveness of the approaches in quantifying uncertainty. However, choosing ResNet-50 as the backbone, resizing input images to 224x224 and following a training regime that is used for the experiments in CelebA yielded 96.23% accuracy for softmax based approach, and 96.16% accuracy for EDL on the test set of CIFAR-10.

Softmax Output as a Confidence Metric. Disadvantage of considering softmax scores as confidence metric is heavily studied in deep learning literature as discussed in section 3.2.1 and is detailed in the following.

For a multi-class classification problems, the objective function is to optimize the cross entropy loss between the predicted distribution and ground truth distribution, as formulated in equation 4.2.

$$-\log \sigma(\mathbf{f}(x, \theta)_y \tag{4.1}$$

where x is the input, y is the corresponding ground truth label, and f_y is the output corresponding to the neural network with parameters θ. The outputs of the network are interpreted as the posterior probabilities of each class:

$$p(y \mid x, \theta) = \sigma(\mathbf{f}(x, \theta)_y) - \log \sigma(\mathbf{f}(x, \theta)_y \tag{4.2}$$

It is important to recognize that the probabilistic interpretation of cross-entropy loss essentially equates to Maximum Likelihood Estimation (MLE). As rooted in

frequentist approach, MLE lacks the ability to determine the variance in predictive distributions [7].

Furthermore, the softmax function is commonly recognized for its tendency to amplify the probability of the predicted class. The culprit behind this phenomenon is the exponential applied on the logits, represented as z in equation 4.3 of the neural network which is shown in:

$$\sigma(\mathbf{z})_i = \frac{e^{\mathbf{z}_i}}{\sum_{j=1}^{K} e^{\mathbf{z}_j}} \tag{4.3}$$

Moving on to the experiments, similar to a test conducted in [12], the average softmax probability of misclassified samples are measured as **0.8011**, which highlights the limitations of relying on softmax outputs as a metric of uncertainty. Furthermore, distribution of the softmax score of the winning class is plotted in Figure 4.1 for incorrectly classified samples on the test set of CIFAR-10. This distribution showcases the networks' overconfidence on incorrectly classified samples.

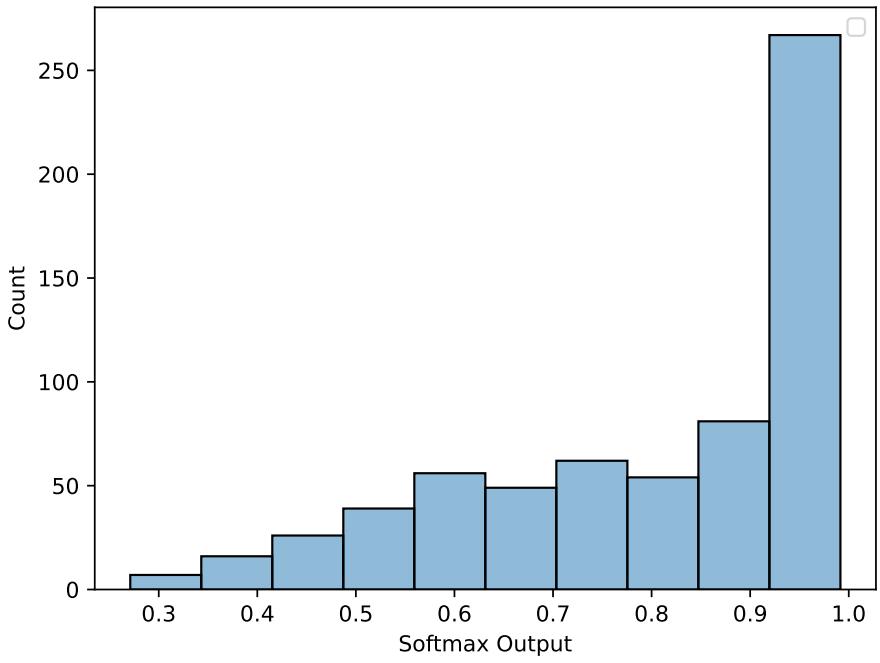


Figure 4.1: Histogram of softmax probability for misclassified samples, belonging to winner class, obtained on the test set of CIFAR-10. Distribution peaks at the tail end of the softmax probability, highlighting its limitations as an indicator of confidence.

EDL.

Extensive qualitative and quantitative analysis conducted on the CelebA dataset regarding uncertainty estimation capabilities of EDL framework are presented in the following sections. Here, we specifically provide our observations on CIFAR-10 for comparison with the other approaches.

A common way to assess the capability of uncertainty estimation is to analyze its distribution. For samples that are correctly predicted, the uncertainty distribution is expected to show a peaking trend, indicating a higher confidence level in these predictions. On the other hand, for misclassified samples, the distribution is expected to become more flattened. This behavior is illustrated in Figure 4.2 by plotting the histogram of uncertainties measured using the test set of CIFAR-10.

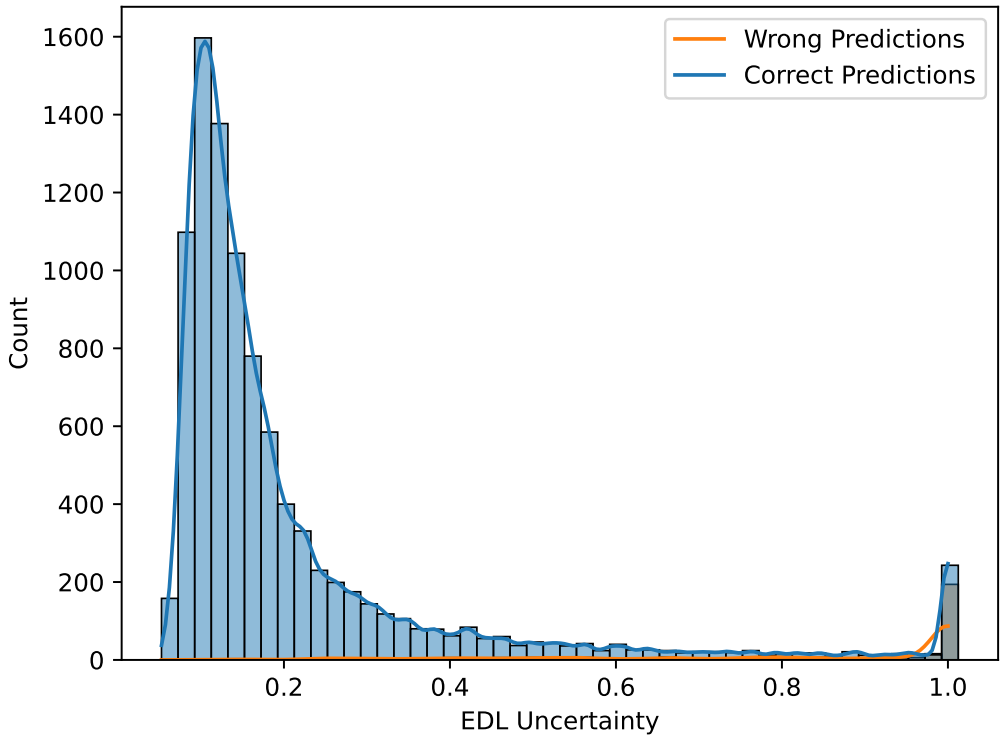


Figure 4.2: Histogram of EDL uncertainties measured on the test set of CIFAR-10, grouped by correct(blue) and wrong(orange) predictions. Distribution of correct predictions peaks at the lower uncertainty values, while it peaks at the tail end for the misclassified samples.

Additionally, the relationship between predictive accuracy and uncertainty measurement is evaluated. To this end, we calculate accuracy across varying uncertainty thresholds, ranging from 0 to 1, by considering only those test samples associated with uncertainty values exceeding the threshold. For instance, examining only the samples with uncertainty > 0.9, we expect lower accuracy compared to samples associated with uncertainty > 0.1. Consequently, this behavior should result in a decreasing curve, which is depicted in 4.3.

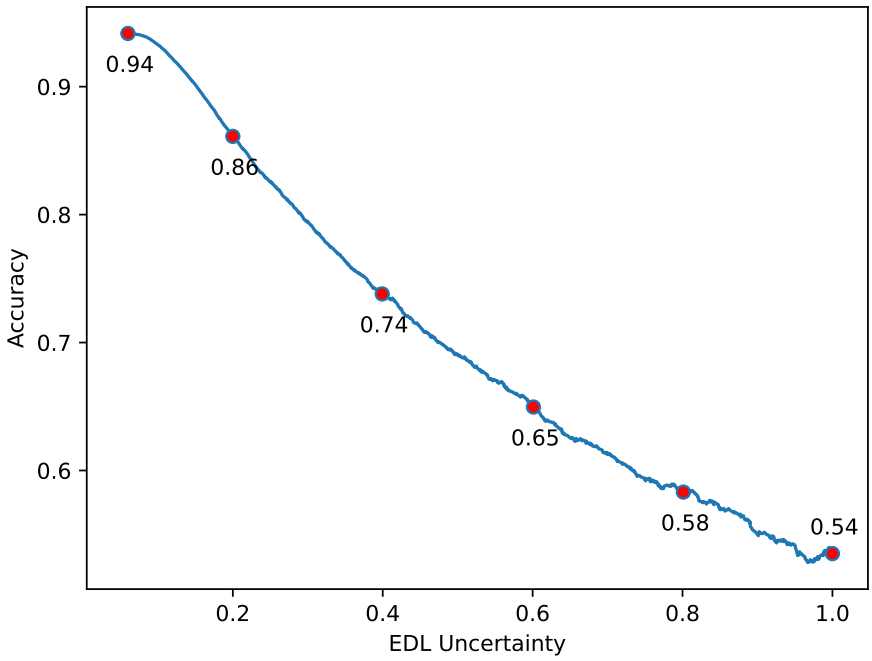


Figure 4.3: EDL uncertainty vs accuracy curve for varying uncertainty values. For each uncertainty threshold t on the x-axis, accuracy is calculated for samples with uncertainty $>$ t. Accuracy reaches its minimum value for the maximum uncertainty threshold.

Finally, we conclude this part by implementing an EDL uncertainty based reject option. We explored a range of reject ratios based on uncertainty values, spanning from 1% to 20% of the most uncertain portion of the test set, and calculated the accuracy for each of these varying reject ratios. The line plot resulted from this approach is demonstrated in 4.4. Filtering out the top 1% uncertain portion of the dataset results in **1.9%** increase in accuracy.

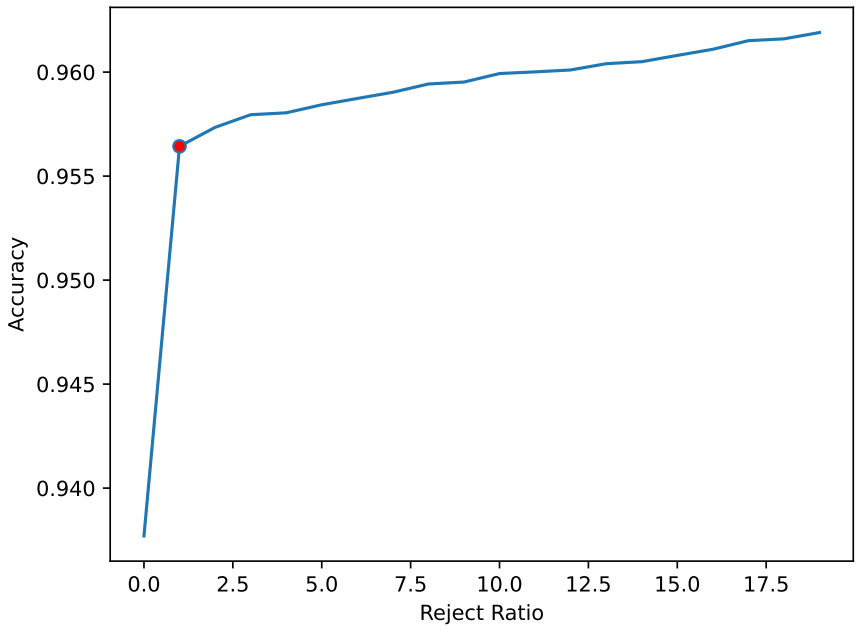


Figure 4.4: EDL uncertainty based rejection option (from 1% to 20%) acquired on test set of CIFAR-10. for instance, for reject rate of 1%, accuracy is measured by filtering out the top 1% uncertain samples, increasing the accuracy of the remaining samples by 1.9% compared to baseline(i.e. from 93.7% to 95.6% as marked with red point in the plot).

Monte Carlo Dropout.

As outlined in section 3.2.2, we implemented a dropout layer with a rate of 0.1 after each ReLU activation in the ResNet-50 architecture and trained the network for 50 epochs. During inference, each sample is passed through the network for 100 times to generate a predictive distribution. For each 100 forward pass, output probabilities for each class are stored. The average probability across all these passes represents the final prediction of the model for that sample.

Entropy is utilized as a metric for uncertainty in model's predictions, a method commonly used in the literature of uncertainty quantification [5, 6, 7, 9]. Defined by the formula in equation 4.4, entropy is a fundamental concept in information theory that quantifies the amount of unpredictability in the outcomes of a random variable. In the context of machine learning, it is used to assess the level of uncertainty in a model's predictions. In other words, high entropy for a sample indicates higher

uncertainty in the prediction, while lower entropy suggests that the model is confident in its predictions. Entropy reaches its maximum value if the model outputs probability of 0.1 for all 10 classes.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{4.4}$$

We first start our experiments by evaluating the uncertainty distribution for correctly predicted and mispredicted samples from the test set. In Figure 4.5, a scatter plot is provided to illustrate the relationship between predictive mean and entropy for two classes of CIFAR-10, namely, automobile and truck. Predictive mean is calculated by averaging over the 100 softmax score for the corresponding class, while entropy is measured according to equation 4.4. False predictions are associated with high uncertainty, while their predictive mean is relative low as expected. Among all the misclassified samples in the automobile and truck classes, the model associates only 3 and 2 samples, lower than the average entropy calculated across the entire test set, as depicted in Figure 4.5.

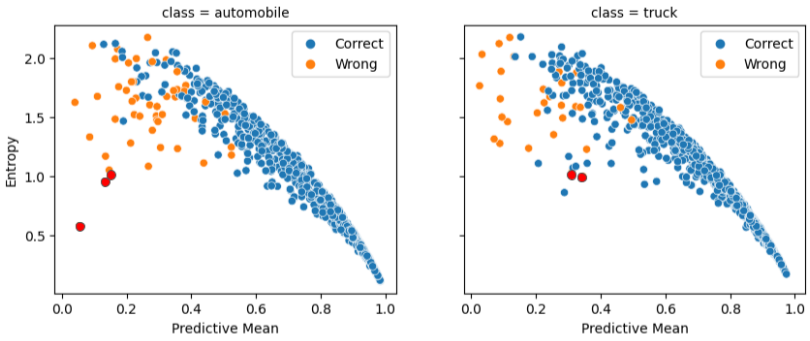


Figure 4.5: Predictive mean of the corresponding class and entropy obtained over the test set of CIFAR-10, for automobile and truck classes. Higher entropy indicates higher uncertainty. Among all the misclassified samples in the automobile and truck classes, the model associates only 3 and 2 samples(marked in red) with entropy lower than the average entropy calculated across the entire test set.

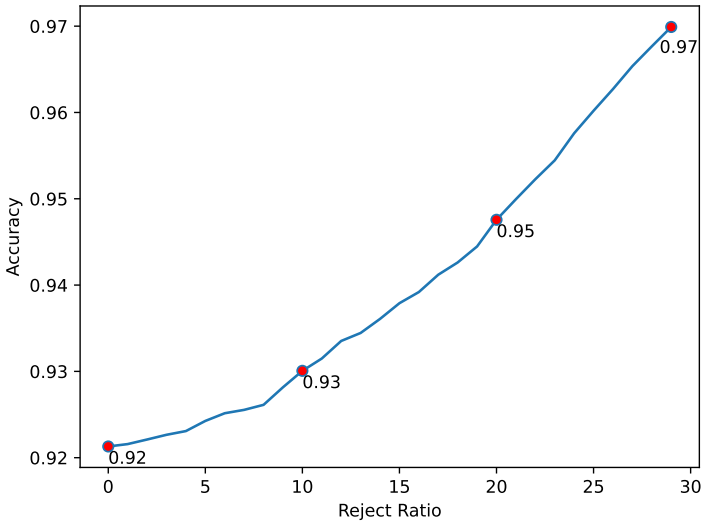


Figure 4.6: MC-DO entropy based rejection option (from 1% to 30%) acquired on test set of CIFAR-10. i.e. for reject rate of 20%, samples with entropy values greater than 80% of the highest entropy calculated in the test is filtered out, and accuracy is calculated considering only those samples.

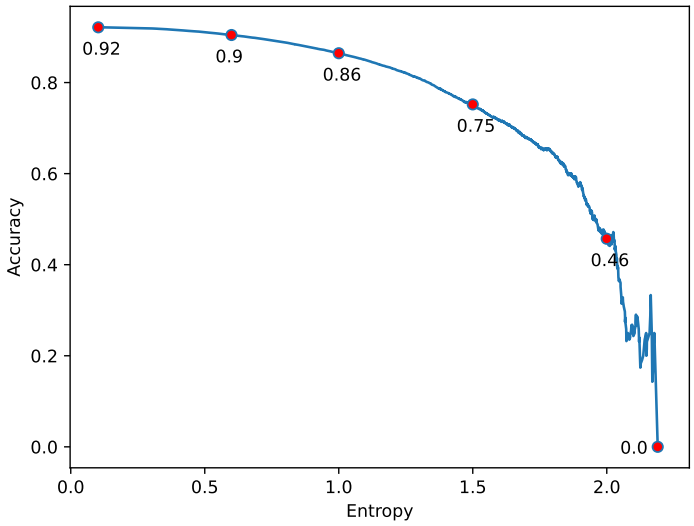


Figure 4.7: MC-DO entropy vs accuracy curve for varying entropy values. For each entropy threshold t on the x axis, accuracy is calculated considering only the samples with entropy $>$ t. Accuracy reaches its minimum value (0) for the maximum entropy measured on the test set.

Next, a rejection option mechanism using the entropy values is implemented akin to the method used in EDL approach as depicted in 4.6. This curve again exhibits an increasing trend, demonstrating that the measured entropy values distinguish between correct and incorrect samples.

To conclude, we provide our results for the correlation between accuracy and measured entropy values. Unlike the uncertainty metric used in the EDL framework, entropy values are not bounded between $[0, 1]$. Confidence values obtained by the MC-DO approach is more spread out from the mean as it exhibits more variation compared to EDL method. This is summarized in Table 4.2, by providing descriptive statistics regarding the confidence metric for all methods. Hence, rather than using a rejection option based on the proportion of confidence values, we offer a rejection criterion that utilizes the statistical characteristics of the confidence metric across different uncertainty quantification methods for comparison purposes. Details of this approach are discussed in the subsection **Observations** and the result is illustrated in Figure 4.12.

Table 4.2: Summary statistics of the confidence metrics obtained on the test set of CIFAR-10, for EDL, MC-DO and Deep Ensemble approach

	EDL	MC-DO	Deep Ensembles
mean	0.23	1.10	0.20
variance	0.05	0.22	0.04
std	0.23	0.47	0.21

Deep Ensembles.

As elaborated in section 4.1.1, 3 ResNet variations are used for deep ensembles, namely, ResNet-18, ResNet-34 and ResNet-50. Accuracy obtained over the test set for each model are measured as 92.6%, 93%, 93.5% respectively. It is important to note that the use of multiple CNNs in this approach demands significantly more memory and computational resources than any other method analyzed in this study, making it impractical for real-world applications. For the remaining of this section, we adopt the same experimental structure that aligns with the methodologies used in other approaches

First, in order to evaluate approach's capability to measure uncertainty, we explore whether the model can distinguish between correctly and incorrectly classified samples based on the uncertainty metric, i.e. entropy. Following this investigation, we find that the average entropy for the correctly classified samples is **0.17**, in contrast to a significantly higher average entropy of **0.72** for misclassified samples.

Second, we implement a rejection option based on entropy values in the ensemble method, paralleling the approach used in other methods. The resulting curve, illustrated in Figure 4.8, shows the impact of excluding the most uncertain portion of the test set on overall accuracy. When we reject the top 1% of the test set to which the model assigns the highest uncertainty, there is a 1.8% increase in accuracy.

Finally, the relationship between predictive accuracy and confidence is analyzed. Similar to other approaches, we compute accuracy at various entropy thresholds, expecting a correlation between higher entropy and lower accuracy. When we filter out the samples with the highest entropy values (for instance, the top 10% corresponding to entropy of 1.56), we observe that the accuracy for samples in this portion reaches zero. However, it's important to note that this outcome primarily stems from the fact that only 4 samples are present in this segment. This is also the reason of the noisy segments locating at the tail of the curve.

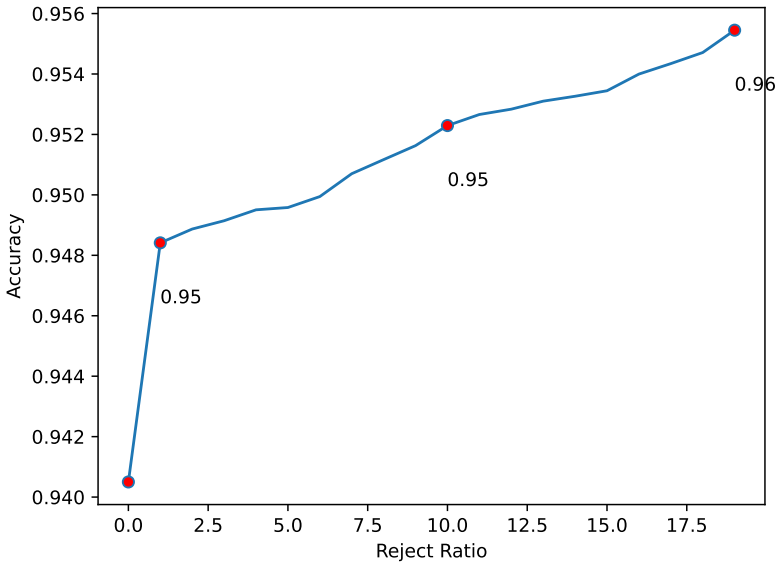


Figure 4.8: Deep Ensembles entropy based rejection option (from 1% to 20%) acquired on test set of CIFAR-10. i.e. for reject rate of 20%, samples with entropy values greater than 80% of the highest entropy calculated in the test is filtered out, and accuracy is calculated considering only those samples.

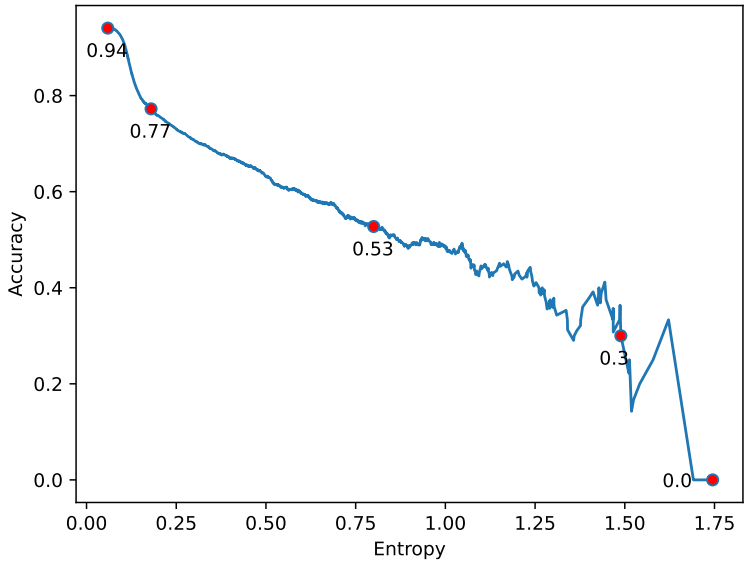
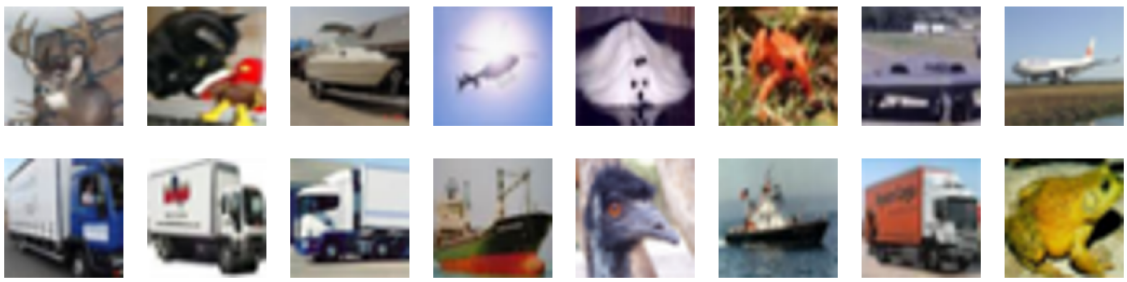
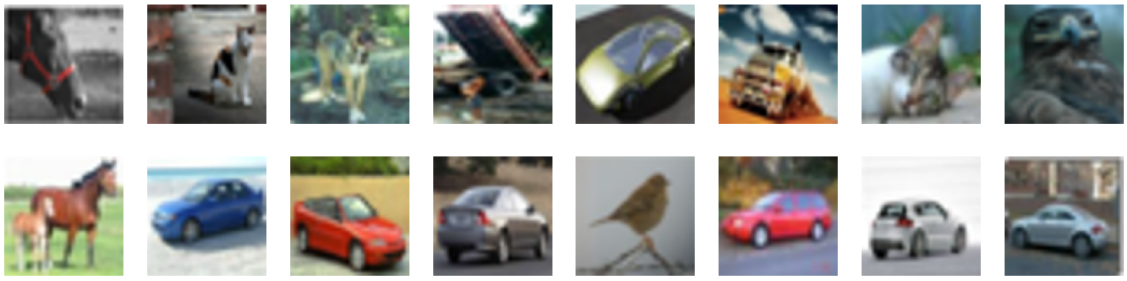


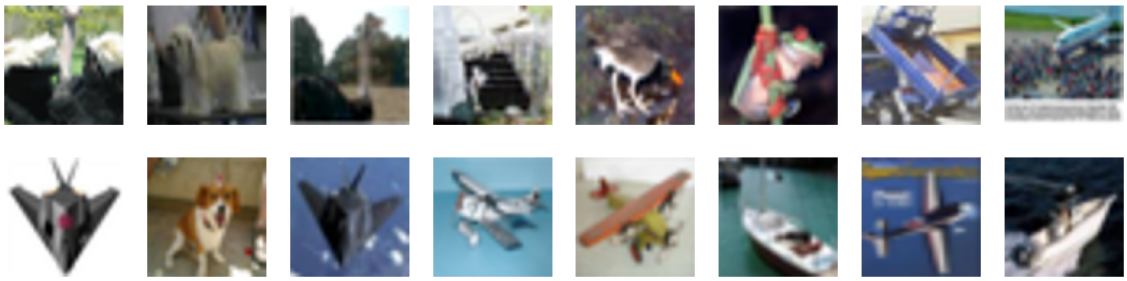
Figure 4.9: Deep Ensembles entropy vs accuracy curve for varying entropy values. For each entropy threshold t on the x axis, accuracy is calculated considering only the samples with entropy $>$ t. Accuracy reaches its minimum value (0) for the maximum entropy measured on the test set.



EDL



MC-Dropout



Deep Ensembles

Figure 4.10: Samples with the highest uncertainties(top row) and the lowest uncertainties (bottom row) measured on the test set of CIFAR-10 for each approach.

Observations.

Here, we extend the individual analysis conducted for each approach in the previous sections to overall comparison and offer our insights. To begin with, we present a qualitative analysis performed on the CIFAR-10 test set for each method. Next, we particularly focus our attention into discrepancy regarding the value range of the uncertainty metrics, which hinders a direct comparison between methods in terms of reject ratios. A short summary of the mentioned problem is provided before presenting our approach to comparison in the subsequent sections.

We explore the test set of CIFAR-10 in Figure 4.10, according to the uncertainty values assigned to the samples by each method. Samples with the highest uncertainties are displayed in the top row, while the bottom row shows samples with the lowest, which is an intuitive result, verifying our implementation qualitatively.

As discussed regarding the experiments for MC-Dropout approach, the uncertainty metric of the EDL framework can take values in the range of [0,1], independent of the number of classes involved in the problem. In contrast, the maximum value of entropy is influenced by the number of classes; with 10 classes in this instance, the upper limit for entropy reaches 3.32. In addition, no test sample reached this upper limit in our experiments, as sample with the maximum entropy has a value of 2.20 and 1.74 for MC-Dropout and Deep Ensembles approaches respectively. Consequently, for instance, if we consider the top top 1% uncertain samples in the test set of CIFAR-10, EDL approach have 440 samples whereas MC-Dropout and Deep Ensembles approaches contain only three and two samples respectively in this portion. This is summarized in Table 4.3 for different methods. This issue is illustrated further in Figure 4.11 by dividing the uncertainty range of each method to 100 evenly spaced intervals, and reporting the corresponding number of samples within each interval. This issue is illustrated further in Figure 4.11 by dividing the uncertainty range of each method to 100 evenly spaced intervals, and reporting the corresponding number of samples within each interval.

Table 4.3: Number of samples that belongs to top K% uncertain portion of the test set, for K $\in \{1, 2, 5, 10\}$ for EDL, MC-Dropout and Deep Ensembles method

	Count		
	EDL	MC-Dropout	Deep Ensembles
top 1%	442	3	2
top 2%	457	12	4
top 5%	493	47	10
top 10%	538	199	24
top 20%	641	914	67

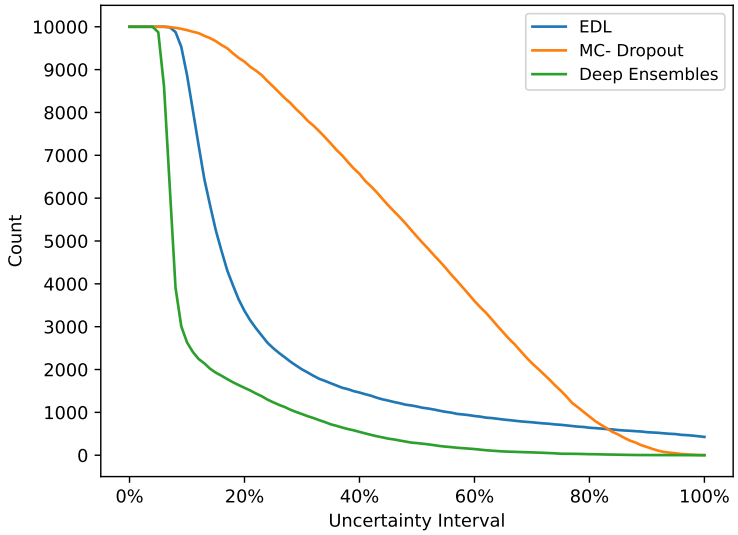


Figure 4.11: Count of samples for varying uncertainty values with evenly spaced intervals , 0% corresponding to minimum uncertainty and 100% indicating the maximum uncertainty value observed on the test set of CIFAR-10 for each approach.

As a result of this observation, we suggest a rejection option based on the descriptive statistics of the uncertainty metrics, measured on the test set. We computed the mean and standard deviation (std) of the uncertainty values in the test set and employed the mean, along with values one and two standard deviations above the mean as threshold values. Table 4.4 shows the number of samples left and their corresponding accuracy when samples with uncertainty greater than the specified thresholds are removed from the dataset. EDL emerges as the most suitable method for this approach, given its higher accuracy across these thresholds.

A similar plot is created by considering all samples that exceed an uncertainty threshold and then applying a rejection criterion, as shown in Figure 4.12.

Table 4.4: Accuracy and the count of samples that remain after excluding samples with uncertainty exceeding the specified thresholds.

Threshold Values	EDL		MC-DO		Deep Ensembles	
	Acc	Count	Acc	Count	Acc	Count
mean	99.6%	7451	99.4%	4994	99.2%	7803
mean + 1 std	99.1%	8808	96.8%	8070	97.9%	8764
mean + 2 std	98.5%	9240	92.5%	9925	96.48%	9369

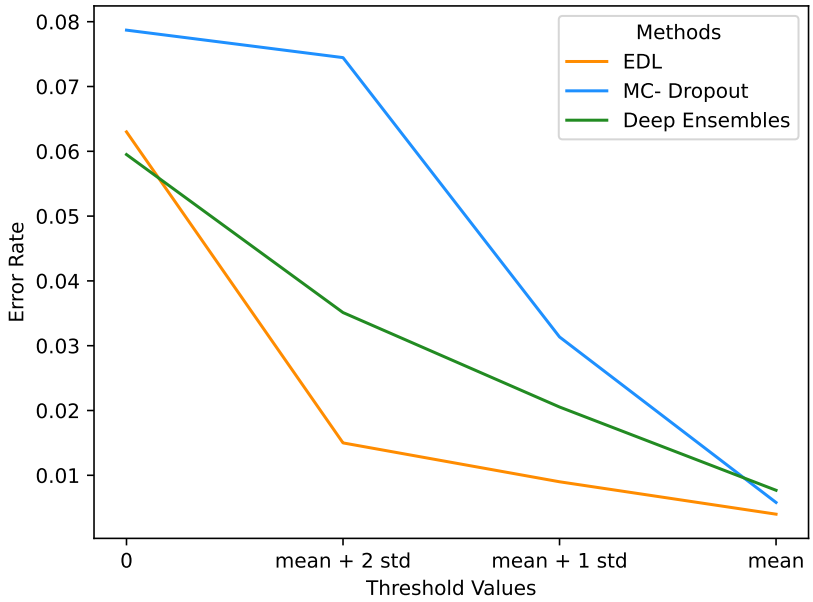


Figure 4.12: Error rate for for the samples that remain if the samples associated with uncertainty exceeding the specified threshold is rejected.

4.3 Analysis of EDL Framework on CelebA & LFWA

In this section, we first provide a literature review for face attribute classification then evaluate the EDL framework in a series of experiments, including both quantitative and qualitative analysis. A significant portion of this section has been previously published in [19].

4.3.1 Face Attribute Classification

Face attributes classification is the task of describing visual characteristics of facial images, including gender and facial expressions, which are therefore beneficial for identifying individuals. It has gained significant attention in wide range of applications, such as face recognition [20, 21, 22], image search and retrieval [23, 24, 25], face verification [26, 27], etc.

Until recent years, facial attributes classification has been addressed by extracting handcrafted features at predefined locations/ landmarks as in [28, 29, 30, 31]. Kumer et al. [28] trained binary classifiers for face attributes based on low-level features extracted from different regions of the face. Li et al. [31] employed multi-scale Gabor features [32] for facial attributes encoding, which are then converted by a learned hashing process for attributes prediction. Even though these kinds of approaches lead to reasonable results in various applications, these handcrafted features are not tuned for the target task and may fail with unconstrained backgrounds and complex facial variations.

Due to the rapid development and the ability of deep learning to learn discriminative features, CNN has shown great success in face attributes classification [14, 33, 34, 35, 36]. In [33], Hand and Chellappa propose a multi-task deep CNN sharing the lowest layers amongst all attributes. In [14], Atito and Yanikoglu take advantage of attributes relationship by training attributes in groups based on their localization in an end-to-end framework and incorporating an ensemble learning technique within the network itself to reduce the training time. In [36], Chen et al. propose an attribute grouping strategy to divide the attributes into task groups based on their

correlation. A recent survey on the topic can be found in [37].

While the above methods learn effective classifiers, the trained models are in general not effective in estimating the confidence/uncertainty of the output. In this thesis, we explore using the EDL framework introduced by Sensoy et al. [7] to represent the prediction uncertainty in face attribute classification.

4.3.2 Accuracy Evaluation

The application of the EDL framework is demonstrated on the face attribute classification problem, using the widely used CelebA dataset. Specifically, we trained 40 binary classifiers for each attribute independently using EDL loss function, employing ResNet-50 as the backbone network. As a baseline, we train the same backbone network using sigmoid activation function, one for each attribute. The training of the binary classifiers are done independently, using the binary cross entropy loss. As shown in Table 1, the models trained with EDL approach outperform the independently trained models with traditional sigmoid layer (denoted as Baseline by a margin of 0.63% points (91.34 vs 90.70%) and obtain better results among on 24 out of the 40 attributes. Considering that the EDL model improved the performance over the simple baseline, we conclude that there is no *disadvantage* of using the EDL framework.

The proposed approach also obtained comparable results to the state-of-the-art methods, which use advanced approaches to improve performance. For instance, both systems [14, 33] use the multi-task learning approach (in addition to other novelties), which has been found to improve accuracy compared to the independent training of 40 binary classifiers due to the regularization brought by the more general learning task. Considering the results in Table 4.5, we see that EDL results are within 2% points of the state-of-art [14], even though the aim was not to beat the state-of-art, but to show that models that are trained with EDL objective are better suited to estimate the prediction uncertainty as shown in the extensive qualitative analysis given below.

We have also implemented a baseline with the multi-task learning approach where the target is the 40-dimensional label corresponding to the 40 labeled attributes that

Table 4.5: State-of-the-art accuracies on CELEBA dataset under two settings; multi-task learning and independent classifiers. Bold figures indicate the best results among the proposed (EDL) and the baseline with independent binary classifiers.

Attribute	Multi-task Learning			Independent Classifiers	
	Baseline-MTL	[33]	[14]	Baseline	EDL (ours)
5_o_Clock_Shadow	94.82%	94.51%	97.18%	94.42%	**94.61%**
Arched_Eyebrows	84.14%	83.42%	85.79%	**83.89%**	83.36%
Attractive	83.04%	83.06%	85.68%	79.29%	**82.76%**
Bags_Under_Eyes	84.95%	84.92%	86.33%	84.46%	**85.11%**
Bald	99.02%	98.90%	99.57%	**98.98%**	98.78%
Bangs	96.16%	96.05%	96.32%	94.42%	**96.26%**
Big_Lips	71.51%	71.47%	92.70%	69.94%	**70.83%**
Big_Nose	84.29%	84.53%	83.36%	**84.43%**	84.18%
Black_Hair	90.42%	89.78%	94.00%	89.12%	**89.32%**
Blond_Hair	96.16%	96.01%	97.89%	94.44%	**95.88%**
Blurry	96.23%	96.17%	96.84%	**96.28%**	96.15%
Brown_Hair	89.51%	89.15%	89.61%	**93.81%**	89.43%
Bushy_Eyebrows	92.86%	92.84%	94.41%	**93.01%**	92.80%
Chubby	95.75%	95.67%	97.54%	95.11%	**95.81%**
Double_Chin	96.50%	96.32%	97.56%	96.37 %	**96.40%**
Eyeglasses	99.66%	99.63%	99.13%	98.91%	**99.65%**
Goatee	97.60%	97.24%	98.41%	95.54%	**97.42%**
Gray_Hair	98.27%	98.20%	98.96%	**98.39%**	98.11%
Heavy_Makeup	91.83%	91.55%	94.19%	89.94%	**90.91%**
High_Cheekbones	87.94%	87.58%	88.69%	**88.01%**	87.45%
Male	98.75%	98.17%	99.13%	**98.71%**	98.37%
Mouth_Slightly_Open	94.07%	93.74%	96.27%	85.67%	**93.60%**
Mustache	96.91%	96.88%	98.75%	95.82%	**96.97%**
Narrow_Eyes	87.51%	87.23%	89.21%	**87.86%**	87.64%
No_Beard	96.36%	96.05%	98.36%	**96.41%**	95.43%
Oval_Face	75.11%	75.84%	77.07%	**75.38%**	75.25%
Pale_Skin	97.05%	97.05%	99.30%	96.47%	**96.92%**
Pointy_Nose	77.83%	77.47%	78.54%	76.48%	**77.36%**
Receding_Hairline	93.95%	93.81%	94.90%	**93.88%**	93.42%
Rosy_Cheeks	95.26%	95.16%	95.66%	**96.13%**	95.17%
Sideburns	97.96%	97.85%	98.05%	96.02%	**97.96%**
Smiling	93.15%	92.73%	95.15%	92.14%	**93.23%**
Straight_Hair	84.21%	83.58%	85.21%	83.58%	**84.56%**
Wavy_Hair	85.46%	83.91%	85.53%	84.79%	**85.61%**
Wearing_Earrings	90.68%	90.43%	91.34%	88.84%	**90.75%**
Wearing_Hat	99.09%	99.05%	99.13%	98.57%	**98.92%**
Wearing_Lipstick	94.28%	94.11%	97.11%	91.04%	**94.34%**
Wearing_Necklace	87.20%	86.63%	88.32%	**88.61%**	87.14%
Wearing_Necktie	97.00%	96.51%	97.58%	94.39%	**96.78%**
Young	88.98%	88.83%	89.84%	88.66%	**88.96%**
Mean	91.56%	91.29%	93.20%	90.70%	**91.34%**

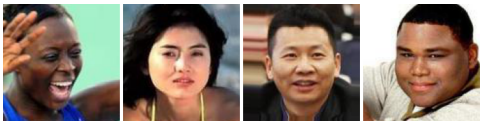
are trained at once, with the same backbone network but no other improvement. The results of this model is given as Baseline-MTL in Table 4.5. Despite the regularization advantage of the MTL training, the EDL approach obtained very similar results in comparison (91.34% vs 91.56%).

4.3.3 Qualitative Analysis of EDL Uncertainties

When we analyzed the samples with the highest uncertainties on the **training set** (uncertainty of 0.9 or larger), we observed that they contain both out-of-distribution and challenging characteristics. These samples are roughly divided into 4 groups: (1) Samples in which faces are occluded by items such as eyeglasses and hats, (2) Samples with under-represented races in the dataset, (3) Samples with non-stereotypical gender traits, and (4) Samples with incorrect ground-truth labels. Examples of samples that are associated with an uncertainty of 0.9 or larger from the training portion of the CelebA dataset are shown in Fig. 4.13, to illustrate these issues. Note that while the first three (a-c) issues are challenges in the problem or the class distributions, the last row (d) indicates labeling problems in the dataset.

On the other hand, when we analyzed the uncertainties associated with the prediction errors on the **test set**, we saw that 72 samples are associated with an uncertainty of 0.1 or lower; in other words, the system is quite certain about its predictions. Upon closer inspection, we realized that more than half of these (39 samples, 54.1%) are actually mislabeled in the ground-truth (indicated by a red cross mark), as shown in Fig. 4.14.

(a) Occluded Faces



(b) Under-represented Races

(c) Non-stereotypical Gender Traits

(d) Errors in the Ground-truth

Figure 4.13: Sample images with high uncertainty for the Male attribute from the training set of CelebA.

Figure 4.14: Sample images corresponding to prediction errors with *lowest* uncertainty for the Male attribute (top-left: most certain; bottom-right:most uncertain). Most of these errors turned out to be ground-truth mistakes (indicated by a red cross mark), while others are genuine mistakes (indicated by green tick).

4.3.4 Uncertainty Distributions

When we analyze the distribution of the uncertainties over the whole test set (rather than just the mistakes), we see that the model is certain in its predictions for most of the samples. Furthermore, the distribution is peaked for the attributes for which the network is more accurate. This is illustrated in Fig. 4.15 for the test set of CelebA dataset over two facial attributes which are easier to visually inspect (i.e. Male, Blond Hair).

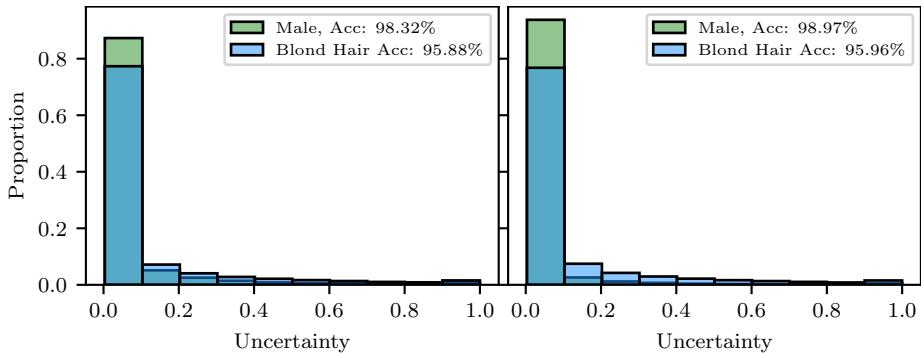


Figure 4.15: Histogram of uncertainties, obtained over the test set(left) and training set(right) of CelebA for the Male and Blond Hair attributes.

In addition, looking at the distribution for positive and negative samples in each attribute, as shown in Table 4.6 for four attributes in CelebA dataset, we see that the system displays more uncertainty towards the class with fewer samples. For instance, the Male attribute is very accurate (98.32%) and well-balanced (62/38) and the resulting uncertainties are very low for either class.On the other hand,

uncertainties are higher for attributes learned with low accuracy (Young) or for under-represented classes in Eyeglasses (positive) and Bald (positive).

In summary, the uncertainties assigned by the EDL system are lower for attributes that are more accurately learned and for classes that are dominant in each attribute. These results show the promise of this framework to assess confidences in a classification task.

Table 4.6: Mean and standard deviation of uncertainties calculated on CelebA test set for the selected attributes, shown with positive and negative class proportions in parentheses. Accuracy for the attributes are 98.32%, 99.65%, 98.78%, and 88.96% from left to right, respectively.

	Male		Eyeglasses		Bald		Young	
	Positive (%62)	Negative (%38)	Positive (%7)	Negative (%93)	Positive (%2)	Negative (%98)	Positive (%75)	Negative (%25)
Mean	0.045	0.032	0.981	0.010	0.895	0.025	0.217	0.432
Std	0.091	0.101	0.132	0.056	0.251	0.112	0.232	0.270

4.4 Using Uncertainties for Reject Option

Motivated by the findings above, we wanted to analyze whether uncertainty scores would be useful as confidence measures that can be used in determining inputs for which the system is unsure and reject to make a decision, during *test* time. First, we considered how many images are rejected and the corresponding accuracies for different rejection thresholds, as shown in Table 4.7. As can be seen there, if we reject all test samples with an uncertainty 0.9 or above, 126 samples are rejected out of 19,962, corresponding to a reject rate of 0.6%, while accuracy increases 0.25% points (98.37 to 98.62%). Hence, using uncertainties that are learned using the EDL framework seems effective for implementing a reject decision, without rejecting too many.

Secondly, we plotted the proportion of wrong predictions among the top-k most uncertain images, in Figure 4.16. Considering the Male attribute, we see that the error rate among the top-100 most uncertain images is 43%, while the error rate for

Table 4.7: Reject rate and resulting accuracy increase (in percentage points) for different uncertainty thresholds, measured over the test set of CelebA for the Male attribute.

Uncertainty Threshold	# of Samples Rejected	Reject Rate	Accuracy Increase
0.5	302	1.51%	0.53
0.6	235	1.17%	0.44
0.7	194	0.97%	0.39
0.8	151	0.75%	0.31
0.9	126	0.63%	0.25

the attribute is only 1.6% overall. When we consider prediction probabilities of the baseline model that does not use EDL, we see that the error rate among the top-100 most uncertain images is smaller (38.5% compared to 43%). In other words, when considering the same number of images to reject, EDL uncertainties better align with prediction mistakes.

Hence, it seems that we can indeed use the EDL uncertainties for denoting a reject region. We can also obtain a similar plot by considering all images above an uncertainty threshold and choose a reject. This is further demonstrated in Fig. 4.17 by plotting the receiver operating characteristic (ROC) curve for increasing uncertainty thresholds.

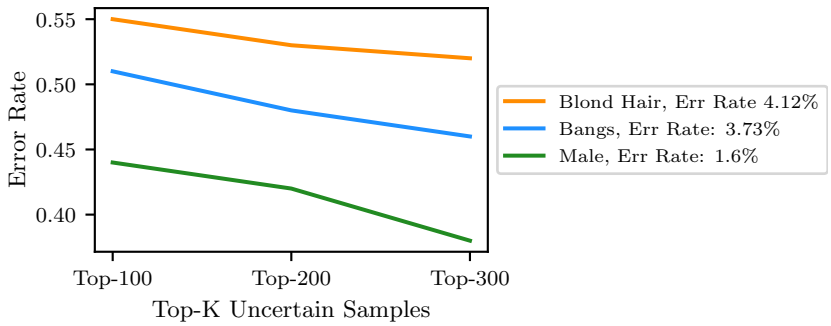


Figure 4.16: Error rate for top-K uncertain samples in CelebA test set for Blond Hair, Bangs, Male attributes. K $\in \{100, 200, 300\}$

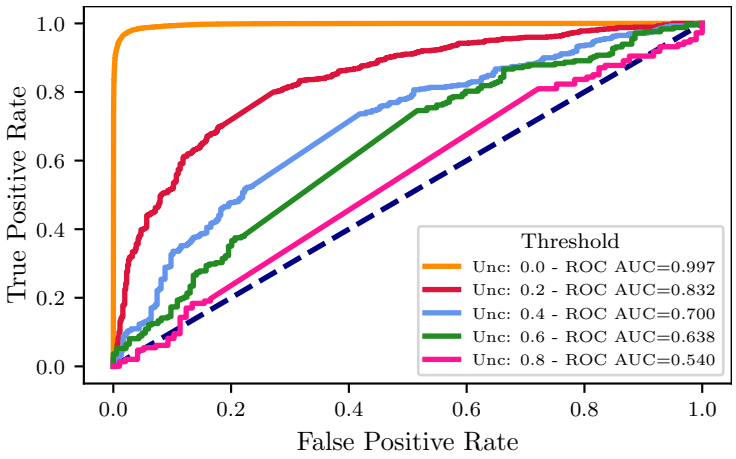


Figure 4.17: ROC curve for the Male Attribute with different threshold values based on uncertainty, from the test set of CelebA

4.4.1 Finding Errors in Ground Truth

Many datasets contain mistakes in the ground-truth label that makes it difficult to benchmark algorithms. This is especially true for the LFWA dataset.

To see if uncertainties estimated by EDL can be used to spot dataset label errors, we found the label mistakes in the train portion of the dataset for the Male attribute. Relabeling was done only when the label error was clear and resulted in 416 new labels, out of 6,263 samples (6.6%).

We then plotted the ratio of ground-truth mistakes (true positives) that are caught when using different uncertainty thresholds, together with the corresponding false positive rate, as shown in Fig. 4.18.

For instance, if we consider the samples that are associated with an uncertainty of 0.5 or above (605 samples), 170 of them are mislabeled in the ground-truth (28%) and 435 are correctly labeled. Thus, **40.86%** of the ground-truth mistakes are caught at the expense of **7.43%** false positive rate.

We conclude that browsing the highly uncertain training samples after the training of the EDL model can be effective, to see if there are any mislabeled samples in the dataset.

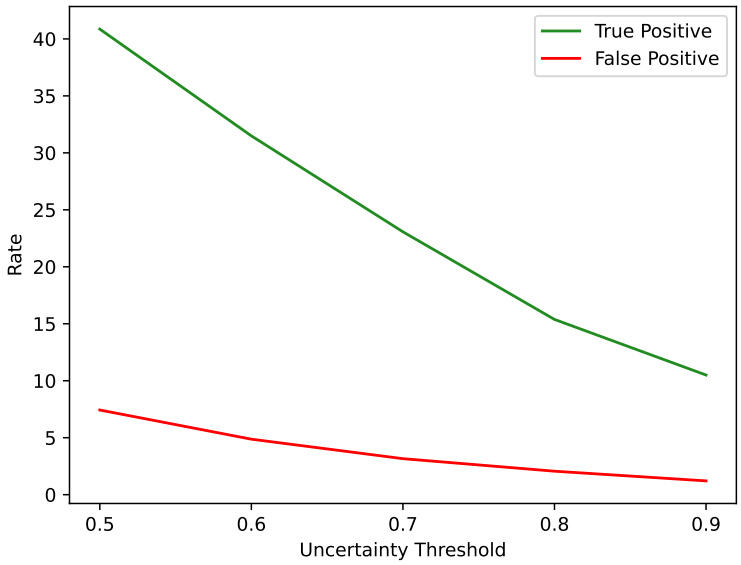


Figure 4.18: True positive (samples with wrong ground-truth caught according to uncertainties) and False positive (samples with correct ground truth labels) rates obtained with varying uncertainty thresholds.

Chapter 5

Conclusion and Future Work

In this thesis, we have explored the problem of uncertainty quantification in deep learning, which gained importance as the deep learning models are being utilized in more applications that significantly influence our daily lives. In particular, our study focused on widely-used sampling based approaches, such as Monte Carlo Dropout and Deep Ensembles, alongside the more recent Evidential Deep Learning method. The latter approach aims to directly learn the parameters of the predictive distribution instead of approximating it through sampling, hence rendering it more efficient. Initially, we conducted comprehensive analysis of each approach on the CIFAR-10 dataset in order to evaluate their effectiveness in measuring uncertainty. Then, we focused our attention into the EDL framework, which showed promising results in the initial experiments. We carried out extensive experimental analysis using significantly larger datasets for face attribute classification, specifically CelebA and LFWA. Our observations regarding these experiments are summarized in the remaining.

To begin our initial exploration work on CIFAR-10, we evaluated the resulting uncertainty distributions on the test set, demonstrating that that models can differentiate between correctly and incorrectly classified samples by the uncertainty values that it assigns. Furthermore, the relationship between accuracy and the measured uncertainty is explored in order to implement a reject region. For this, we offered 2 options: (1) based on the numerical values that uncertainty metrics can take in each method (2) using the summary statistics measured on the test set. In the light of

these experiments, for a system that precision is the priority, we suggest that Deep Ensembles approach could be more appropriate since samples that it associates with the highest uncertainty, though very few in number compared to other approaches, contains a greater proportion of incorrectly classified samples. EDL performed better for the second option as it resulted in higher accuracy for each specified threshold values. It is important to acknowledge that our test setup may not fully capture the efficacy of MC-Dropout. Although we followed the suggestion of the original author and introduced dropout before every weight layer, this approach, in conjunction with the configuration of ResNet, results in an overly strong regularization effect for CIFAR-10. Therefore, we conclude that while MC-Dropout is fairly straightforward to implement, it demands specific architectural considerations based on the nature of the problem.

As for the experiments conducted on CelebA using EDL method, we first evaluated the predictive performance of traditional softmax-based approach against the EDL-based approach on the 40 unique attributes provided in the dataset. For this, we trained 40 separate classifiers for both methods, each utilizing the same backbone. Additionally, we implemented a baseline using the multi-task learning approach for comparison. We conclude that the EDL framework does not compromise predictive performance in binary classification tasks due to

- Models trained with EDL approach outperform the independently trained models with traditional sigmoid layer by a margin of 0.63% points (91.34% compared to 90.70%) and achieve better results in 24 of the 40 attributes.
- Despite the added benefit of regularization in MTL training, the results from the EDL approach were very similar (91.34% compared to 91.56%).

It is worth highlighting the measures we have taken to maintain fairness in a comparison involving 80 distinct models. To this end, we set the initial learning rate of each model according to individual learning rate range tests, trained all models for 100 epochs and applied early stopping with a patience value of 10, and used the same optimizer, scheduler and batch size as stated in Section 4.1. We intend to extend this comparison to multi-class classification problems in future to reach a more general conclusion.

The comprehensive qualitative study we conducted on the CelebA dataset revealed that the training set samples associated with the highest uncertainty values often display out-of-distribution and challenging features such as occluded faces, under-represented races as well as samples with incorrect ground truth labels. Furthermore, we demonstrated that prediction uncertainties learned by the system can be used for indicating potential ground-truth mistakes in the dataset and weaknesses of the system with regard to class imbalance and challenges present in the data. We find it important to emphasize the usefulness of uncertainty estimation methods as a tool to identify the potential weaknesses in the dataset. This proves particularly invaluable when curating real-world datasets, as it allows for necessary adjustments to ensure that the distribution of the training data aligns with that of the broader real-world dataset. Building on these insights, we proposed a method to detect labeling errors using our relabeled version of the LFWA dataset. Our approach managed to catch 40.86% of the ground-truth mistakes at the expense of eliminating 7.43% of correctly labeled samples.

We suggest that the EDL framework can be used in many classification problems as it requires only a small change in the general network architecture and loss term, and the assessed uncertainties are good indicators of prediction confidences. In future work, we will extend the EDL formulation to multi-label problems and compare it with the other uncertainty qualification methods on larger and more complex datasets.

Bibliography

[1] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[2] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[3] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[4] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[5] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[6] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in neural information processing systems*, vol. 30, 2017.

[7] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *arXiv preprint arXiv:1806.01768*, 2018.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[9] W. He and Z. Jiang, "A survey on uncertainty quantification methods for deep neural networks: An uncertainty source perspective," *arXiv preprint arXiv:2302.13425*, 2023.

[10] K. Sentz, S. Ferson *et al.*, *Combination of evidence in Dempster-Shafer theory*. Sandia National Laboratories Albuquerque, 2002, vol. 4015.

[11] A. Jøsang, *Subjective Logic*. Springer, 2016.

[12] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[13] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," *arXiv preprint arXiv:1905.11001*, 2019.

[14] S. A. A. Ahmed and B. Yanikoglu, "Within-network ensemble for face attributes classification," in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 466–476.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[16] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.

[18] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[19] A. Zeyneloglu, S. A. A. Ahmed, and B. Yanikoglu, "Face attribute classifica-

tion with evidential deep learning," in *Fifteenth International Conference on Machine Vision (ICMV 2022)*, vol. 12701. SPIE, 2023, pp. 488–495.

[20] D. Liu, N. Wang, C. Peng, J. Li, and X. Gao, "Deep attribute guided representation for heterogeneous face recognition." in *IJCAI*, 2018, pp. 835–841.

[21] C. Huang, Y. Li, C. C. Loy, and X. Tang, "Deep Imbalanced Learning for Face Recognition and Attribute Prediction," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 11, pp. 2781–2794, 2019.

[22] M. A. Diniz and W. R. Schwartz, "Face attributes as cues for deep face recognition understanding," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 307–313.

[23] X. Zhao, H. Li, X. Shen, X. Liang, and Y. Wu, "A modulation module for multi-task learning with applications in image retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 401–416.

[24] S. Suchitra and R. Poovaraghan, "Dynamic multi-attribute priority based face attribute detection for robust face image retrieval system," *Multimedia Tools and Applications*, vol. 79, no. 33, pp. 24 825–24 849, 2020.

[25] A. Zaeemzadeh, S. Ghadar, B. Faieta, Z. Lin, N. Rahnavard, M. Shah, and R. Kalarot, "Face image retrieval with attribute manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 116–12 125.

[26] Y. Lu, Y.-W. Tai, and C.-K. Tang, "Attribute-guided face generation using conditional cyclegan," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 282–297.

[27] X. Di, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Multi-scale thermal to visible face verification via attribute guided synthesis," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 266–280, 2021.

[28] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 365–372.

[29] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1543–1550.

[30] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose Aligned Networks for Deep Attribute Modeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1637–1644.

[31] Y. Li, R. Wang, H. Liu, H. Jiang, S. Shan, and X. Chen, "Two Birds, One Stone: Jointly Learning Binary Code for Large-scale Face Image Retrieval and Attributes prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3819–3827.

[32] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Transactions on Image processing*, vol. 11, no. 4, pp. 467–476, 2002.

[33] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[34] S. A. Aly and B. Yanikoglu, "Multi-label networks for face attributes classification," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2018, pp. 1–6.

[35] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2597–2609, 2017.

[36] Z. Chen, F. Liu, and Z. Zhao, "Let them choose what they want: A multi-task cnn architecture leveraging mid-level deep representations for face attribute classification," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 879–883.

[37] X. Zheng, Y. Guo, H. Huang, Y. Li, and R. He, "A Survey of Deep Facial

Attribute Analysis," *International Journal of Computer Vision*, vol. 128, no. 8, pp. 2002–2034, 2020.