How to Stop Off-the-Shelf Deep Neural Networks from Overthinking

Yigitcan Kaya, Tudor Dumitras
University of Maryland, College Park

Abstract

While deep neural networks (DNNs) can perform complex classification tasks, most of their natural inputs do not necessitate the depth of the modern architectures. This leads to wasted computation, as the network overthinks on the simpler inputs. The overthinking problem could be prevented if standard DNNs could produce early predictions. However, prior work suggests that this is challenging in existing architectures, such as ResNet, as their internal layers are not trained for classification and optimizing them for accurate predictions hurts the end performance. In this paper, we explore the overthinking problem, and, as a remedy, we propose a generic modification to off-the-shelf DNNs—the Shallow-Deep Network (SDN). With this modification, a DNN can efficiently produce predictions from either shallow or deep layers, as appropriate for the given input. We employ feature reduction and a layer-wise objective function to train these progressively deeper internal classifiers while preserving the end-performance. We can apply the SDN modification either by training from scratch or by tuning a pre-trained model. Experiments on four architectures (VGG, ResNet, WideResNet, and MobileNet) and three image classifications tasks suggest that, for an average input, an SDN can produce a correct prediction before its middle layer. By avoiding unnecessary computation, the SDN can reduce the required number of operations for an input by 41% over the original network. Finally, we observe that disagreements among the early classifiers reliably indicate inputs where the network is likely to make a mistake. Building on this observation we propose an internal confusion metric and a method to diagnose misclassifications by visualizing these disagreements.

1 Introduction

Deep neural networks (DNNs) have enabled recent breakthroughs on many tasks, such as image classification (Krizhevsky, Sutskever, and Hinton 2012) and speech recognition (Hinton et al. 2012). In such tasks, a DNN’s sequence of layers resembles human perception in the way it combines simple (shallower) representations, such as edges, into more complex (deeper) ones, such as faces (Zeiler and Fergus 2014). A fundamental difference is that people can learn simpler heuristics that allow them to perform even complex tasks, such as driving or playing the piano, with little mental effort (Kahneman 2011). When such heuristics are adequate, human overthinking is considered as harmful, because it leads to slow decision making, when people think too much or for too long. Overthinking can also cause confusion and mistakes when irrelevant details divert the attention. However, in a standard DNN, the decision-making process—the forward pass—always remains the same, whether the input is simple or difficult to classify.

In this paper, we ask the question: Do deep neural networks overthink on some of their inputs? We define neural network overthinking1 as a DNN model’s potential to make a correct classification in its internal layers but delaying the decision because of the lack of an early prediction mechanism. While it is known that a network’s depth does not necessarily correlate with its performance (He et al. 2016), this does not necessarily imply that the network overthinks, as accurate early predictions may not be feasible. A generic mechanism to enable such predictions and prevent overthinking would conserve computation at test time: practical applications include adapting DNNs to devices with reduced computational resources and improving their inference time.

The prior work does not paint an optimistic picture regarding early predictions. Researchers have proposed the use of internal classifiers against the vanishing gradients problem (Lee et al. 2015; Szegedy et al. 2015). These approaches discarded the internal classifiers, as the main objective was to improve the end performance. Moreover, researchers suggested that, for off-the-shelf architectures such as ResNet, having accurate internal classifiers is inherently at odds with preserving the original performance (Huang et al. 2017). To overcome this problem, researchers proposed runtime network pruning (Lin et al. 2017), brand new architectures (Huang et al. 2017), or cascading simple models with more complex ones based on the input difficulty (Wang et al. 2017). However, enabling accurate early predictions in off-the-shelf DNNs remains an open question.

We present results suggesting that, for some inputs, off-the-shelf convolutional neural networks (CNNs) are in fact capable of making a correct classification early in the forward pass; in other words, these CNNs sometimes overthink.

1Prior work has utilized this term to describe a property of the whole network, regardless of its internal state (Wang et al. 2017). In contrast, we consider that a network overthinks only when it performs more computation than it needed to, as this is analogous to human overthinking.
We investigate a universal remedy for overthinking by introducing internal classifiers, which can be trained to perform well, in generic DNN architectures. Additionally, we show that accurate internal classifiers can provide insights into how a network reaches to its end prediction.

Our first contribution is the Shallow-Deep Network (SDN), a modification to existing DNNs for introducing internal classifiers. As illustrated in Figure 1, an SDN combines shallow and deep internal classifiers in a single network and makes predictions employing various feature complexity levels. We overcome the challenges reported in the prior work by (i) performing feature reduction before each internal classifier (see Figure 1), to regularize these classifiers and to ensure that our method can scale up to large DNNs; and (ii) utilizing a layer-wise objective function, to train these internal classifiers. These techniques make the SDN modification practical for a range of neural network models, including VGG (Simonyan and Zisserman 2014), which is commonly used as a feature extractor, ResNets (He et al. 2016), which are the building blocks for more advanced architectures, and MobileNets (Howard et al. 2017), which are designed for constrained mobile devices. We can apply the modification either by training the network from scratch or by tuning a pre-trained model. Our experiments suggest that our modification results in accurate internal classifiers while preserving—and sometimes improving—the original network’s performance.

Our second contribution is to utilize SDNs as a vehicle for exploring overthinking in off-the-shelf CNNs. First, we show that most natural inputs do not require the CNN’s full depth. For these simple inputs, overthinking leads to slow inferences and wasted computation. Second, we show that some complex inputs are recognized correctly only by the

SDNs’ first internal classifier, which focuses on simple features (e.g. the general structure of a beaker) while the subsequent layers discover finer details that are irrelevant (e.g. text on the beaker). For these inputs, overthinking actually leads to errors. We also show that SDNs can mitigate the first pitfall of overthinking by preventing unnecessary computation. Using an entropy-based metric to quantify prediction confidence, we can reliably determine when the network should stop thinking and make an early prediction at an internal classifier. This allows SDNs to reduce computation by 25% without loss of accuracy.

Our third contribution is to demonstrate that SDNs can help the diagnosis of the errors made by neural networks. The second pitfall of overthinking yields the intuition that the internal classifiers can reveal how confused a network is when making a prediction. In a confused decision, the internal predictions are inconsistent, as they disagree with each other. We propose a new confusion metric, which quantifies the disagreement among the internal predictions of an SDN. We observe experimentally that this metric is an indicator of the cases where the network is likely to misclassify the input. Furthermore, by visualizing the internal disagreements we can investigate the input elements that cause the confusion. In addition to their practical application regarding diagnosing DNN errors, these visualizations provide a new perspective for reasoning about model interpretability.

We evaluate our techniques on three common benchmarks: CIFAR-10, CIFAR-100 (Krizhevsky and Hinton 2009), and Tiny ImageNet by applying the SDN modification to four off-the-shelf CNN architectures: VGG (Simonyan and Zisserman 2014), ResNet (He et al. 2016), Wide ResNet (Zagoruyko and Komodakis 2016) and MobileNets (Howard et al. 2017). On an average CIFAR-10 input, an SDN makes an early prediction before its middle layer. This reduces the required test time computation by 52%, 44%, 37% and 67% for VGG-16, ResNet-56, Wide-ResNet-32-4 and MobileNet, respectively. Furthermore, SDN tuning reduces the error rate by up to 9% over the original network, by acting as a regularizer. Finally, our normalized confusion scores suggest that a network is significantly less confused in a correct prediction (−0.0002 on average) than in a wrong prediction (1.84 on average).

2 Related work

Internal Classifiers. Deeply Supervised Networks (Lee et al. 2015) and Inception architecture (Szegedy et al. 2015) have proposed using internal classifiers for improving the end performance. These studies discard the internal classifiers, as they are not designed, or trained, for accurate predictions. Recent work describes a new MSDNet architecture for any time predictions, in the case of insufficient computing resources for a full forward pass (Huang et al. 2017). MSDNets is designed in response to the finding that claims attaching internal classifiers to existing architectures hurts the end performance. We propose Shallow-Deep Networks as a modification, instead of a new architecture, for intro-

Figure 1: Neural-network overthinking on an image recognition task. The dotted components show the SDN modification to the original convolutional neural network. The SDN makes two internal predictions and an end prediction. Even though the end prediction is correct, the network actually overthinks, as it could make a correct prediction earlier.

http://tiny-imagenet.herokuapp.com/
Reducing internal classifiers to pre-trained off-the-shelf networks. An SDN aims to achieve two goals: previous studies have suggested that it is contradictory: having accurate internal classifiers while preserving the original performance. Furthermore, previous work has proposed cascading different models for reducing the inference costs. In IDK Cascades (Wang et al. 2017), authors cascade simple networks with more complex ones based on the prediction confidence. In these cascades, the simple model makes an early prediction, if it has enough confidence in its prediction. However, each model in the cascade relies on similar representations to make a prediction (Yosinski et al. 2014); thus, essentially, they redo the previous computation when invoked. An SDN makes in-network early predictions by chaining its internal prediction, thus, reducing this redundancy as each internal layer reuses the preceding computation.

Neural Network Certainty Measures. Widespread usage of DNNs makes indications of prediction correctness crucial for handling the potential errors. Prior work has shown that predicted probability distribution—softmax scores—makes an unreliable metric, as it is over-confidence towards a single class (Szegedy et al. 2013). To overcome the problem, prior work has described a calibration scheme (Guo et al. 2017) and Bayesian model uncertainty estimation (Gal and Ghahramani 2016). The credibility metric (Papernot and McDaniel 2018) uses the representational distances to neighboring training points. Our confusion metric is inherently to SDNs and orthogonal to these approaches as it essentially captures the how consistently a network reaches to an end prediction. Confusion is a non-expensive metric, as it does not require any calibration, reconfiguration or computing distances.

Visualizing Neural Networks. Because of their complex structures, DNNs are hard to interpret by humans (Olah et al. 2018). Existing visualization techniques aim to gain intuition about a network for more interpretable models. Deconvolutional networks are designed (Zeiler and Fergus 2014) to provide projections of features a DNN learns. Similarly, Grad-CAM (Selvaraju et al. 2017) uses the gradients to highlight the important input regions for any prediction. Instead of explaining a prediction, we first detect the confusion in the network; then use an existing technique to visualize and investigate the input elements that cause this confusion.

3 The Proposed Method

We consider the supervised learning setting and the standard DNN structure: a sequence of internal layers ending with an end classifier. Let \( S = (x, y) \in (X, Y) \) be our input-output pairs, where sample \( x \) denotes the input data, an instance, and \( y \in \{1, \ldots, K\} \) is its correct ground-truth label. A DNN with \( M \) internal layers, given an instance \( x \), performs the following classification to predict its probability of belonging to each class: \( F_{(M+1)}(F_M(\ldots F_1(x))) \). Here \( F_m \) denotes the learnable function layer \( m \) applies and \( F_{(M+1)} \) denotes the end classifier. To simplify the notation, we write the output of the layer \( m \) as \( F_m(x) \). We denote end prediction on the instance \( x \) as \( \hat{y}_{(M+1)} = \arg\max_{y} F_{(M+1)}(x) \), i.e. the class with the highest probability.

3.1 Shallow-Deep Networks

Figure 1 describes a motivating scenario for Shallow-Deep Networks, and how we modify an off-the-shelf convolutional neural network (CNN) as an SDN. Modern CNNs achieve groundbreaking results especially on perception tasks, but they also come with a significant computational burden. Our experiments show that standard CNNs overthink on simpler instances, however, mitigating overthinking requires a classification mechanism before \( F_{M+1} \). We specifically focus on modifying off-the-shelf CNNs as Shallow-Deep Networks for establishing such mechanism; and leave the extension to other architectures for future work.

Formally, we define overthinking as follows: for \( 1 \leq m \leq M \), consider a learnable function \( F'_m \), an internal classifier, that makes a prediction using \( F_m(x) \). We denote the internal prediction at the layer \( m \) as \( F'_m(F_m(x)) \), or simply as \( F'_m(x) \). We say that a network overthinks on input \((x, y)\), if \( y = y_m = \arg\max_{y} F'_m(x) \), namely network can make a correct prediction at the internal layer \( m \). Overthinking is distinct from the overfitting problem, as the former leads to wasted computation, but, the latter hurts the classification performance. Furthermore, whereas regularization, such as Dropout (Srivastava et al. 2014), can mitigate overfitting; there is no general solution for overthinking.

Researchers suggest the difficulty of having viable internal classifiers in off-the-shelf CNNs (Huang et al. 2017), in the next sections, we describe our design for overcoming this difficulty. We first describe modifying an off-the-shelf CNN by attaching classifiers to its internal layers. Next, we specify the feature reduction of the internal layer outputs for keeping internal classifiers practical. Finally, we formulate our layer-wise objective function to train the internal classifiers by tuning a pre-trained network.

Modifying an Off-The-Shelf CNN. Figure 1 shows the SDN modification and the internal classifiers it introduces to a CNN. An internal classifier consists of two parts: a feature reduction module and a fully connected \( F'_m \) layer. A feature reduction module takes \( F_m(x) \) and returns a smaller \( \tilde{F}_m(x) \). Following the reduction, \( F'_m \) takes \( \tilde{F}_m(x) \) and makes an internal prediction.

We highlight the overhead of modifying off-the-shelf CNNs in Table 1. Considering the inferences at the end classifier, the overhead on the required computation—FLOPs per instance—is minor; as the fully connected layers are computationally efficient. Similarly, the training time of an SDN is comparable to the original network, with around 10% overhead. However, the modification causes an increase in the network size, proportional to the number of classes, because of \( F_m \) layers. To reduce this overhead, we incorporate feature reduction modules before each \( F'_m \).

Furthermore, we only attach internal classifiers to a subset of internal layers, i.e. \( \mathcal{A} \subset \{1 \ldots M\} \), \(|\mathcal{A}| = N\). We attach a \( F'_m \) once every few layers with the exception of the initial layers; focusing on the layers where feature map size changes, i.e. after pooling or strided convolution. We make this se-
In this section, we describe the ways we utilize the accurate internal classifiers we introduce to the off-the-shelf CNNs.

### Obtaining Early Predictions

An SDN, unless stopped early, makes sequential predictions at each internal classifier, and a last one at the end. We observe that for the majority of correct end predictions, an SDN can actually make a correct internal prediction as well. However, determining if a prediction is correct requires an unrealistic ground truth. As a practical solution, we propose using the certainty of a prediction as an indicator of its correctness. If an internal prediction is certain enough, then we stop the forward pass and make an early prediction. Conversely, if a prediction is uncertain, then we forward the instance to the following layer for another prediction at the next classifier. To quantify the prediction certainty, we measure its entropy of the predicted probability distribution.

Given an instance \( x \), the uncertainty of the \( i \)th prediction is \( H_i[F_i^g(x)] = -\sum_{j=1}^K F_i^g(x)^j \log F_i^g(x)^j \); where \( F_i^g(x)^j \) denotes the predicted probability of instance \( x \) belonging to the class \( j \). To find the threshold between certainty and uncertainty, first, we search for a \( 0 < q < 1 \): the probability that an instance reaching to an internal classifier obtains an early prediction. Based on \( q \), using the training set, we calculate the entropy threshold \( \theta \) for the \( i \)th internal classifier as \( p(H_i[F_i^g(x)]) < \theta \). In our experiments, we do a search to find ideal \( q \) in three settings that prioritize either the prediction performance or the required computation.

### Internal Confusion Metric

An SDN’s internal predictions throughout the forward pass reveal how consistent a network reaches to its end prediction. A disagreement among them hints that a network is making an inconsistent, a confused, prediction; whereas an agreement indicates consistency. To quantify this inconsistency, we propose the confusion metric by correlating the disagreement among the internal predictions and the end prediction. First, given the instance \( x \), we measure the disagreement between the prediction pair \( i \) and \( j \) as: \( c_{(i,j)}(x) = ||F_i^g(x) - F_j^g(x)||_1 \). After experimenting with different distance metrics, such as \( L_2 \) or \( D_{KL} \), we see that \( L_1 \) produces the most ideal scores. The summation of all pairwise disagreement scores, i.e. \( C(x) = \sum_{i,j} c_{(i,j)}(x) \), gives us the unbounded confusion score, on the instance \( x \). To normalize these scores, we compute the \( \bar{C} \) and \( \sigma_C \): the mean and standard deviation of the confusion scores on the training set instances.

### Visualizing Internal Confusion

We observe that in a significant portion of misclassifications an SDN makes, there is a disagreement among the internal predictions. We quantify these disagreements in our confusion metric and hypothesize that confusion can cause wrong predictions. Towards error mitigation, our confusion metric can indicate whether the network is likely to make a mistake. As a further step, we propose visualizing the confusion for investigating its source in the input. Specifically, given an input \((x, y)\), we visualize the disagreement in the prediction pair \( i \) and \( j \), i.e. \( y = \hat{y}_i \neq \hat{y}_j \). We use Grad-CAM (Selvaraju et al. 2017) for visualizing the input elements that are influential in a prediction. With this visualization, we aim to interpret network confusion and the mistakes it triggers.
4 Experiments

4.1 Experiment Settings

Datasets. In our experiments, we use three datasets for benchmarking: CIFAR-10, CIFAR-100 (Krizhevsky and Hinton 2009) and Tiny ImageNet. The two CIFAR datasets consist of 32x32 pixels, colored natural images. CIFAR-10 and CIFAR-100 images are drawn from 10 and 100 classes, respectively; containing 50,000 training and 10,000 validation images. We use a standard data augmentation scheme: padding, random cropping, and random horizontal mirroring. The Tiny ImageNet dataset consists of a subset of ImageNet images (Deng et al. 2009), resized at 64x64 pixels. There are 200 classes, each of which has 500 training and 50 validation images. We augment the data with random crops, horizontal mirroring, and RGB intensity scaling.

Architectures. We experiment with four off-the-shelf CNNs: VGG (Simonyan and Zisserman 2014), ResNet (He et al. 2016), Wide-ResNet (WRN) (Zagoruyko and Komodakis 2016) and MobileNet (Howard et al. 2017). Specifically, we use VGG-16, ResNet-56 and WRN-32-4 configurations of these architectures. We denote an SDN variant of a network by appending SDN to its name; e.g. VGG-16 becomes VGG-16-SDN. We train these CNNs for 100 training epochs, using the hyper-parameters in original studies. We treat the resulting CNNs as our pre-trained, off-the-shelf networks; and after the modification, we tune their corresponding SDNs for 25 epochs. For a fair comparison with respect to the total training time, instead of the CNNs after epoch 100, we apply the SDN modification to the CNNs after epoch 75. For comparing the test time computational requirements of CNNs and SDNs, we report avg. FLOPs: the average floating point operations a network performs per input, in billions. As the classification performance, we simply report the Top-1 accuracy on the validation data, as a percentage.

4.2 The Overthinking Problem

In this section, we illustrate the overthinking problem of an off-the-shelf CNN architecture, VGG-16, on Tiny ImageNet task. To this end, we first obtain the corresponding SDN, VGG-16-SDN. The SDN contains 6 internal classifiers and an end classifier; thus, producing a total of 7 predictions per instance. Figure 2 presents the validation accuracy the SDN achieves in each classifier compared to the original CNN. The first internal classifier, even though it requires only 37% of the FLOPs the VGG-16 performs, achieves 74% of the original accuracy and produces correct predictions for 62% of the instances the CNN correctly classifies. The third internal classifier achieves comparable accuracy to the CNN, while requiring 67% of the original FLOPs. These results suggest that, after a certain depth, adding more layers does not improve the accuracy. Furthermore, for the majority of instances, we can reliably produce correct predictions at the internal layers, with significantly less computation compared to the original CNN. In Section 4.4, we show that these internal classifiers facilitate early predictions on instances where the CNN would overthink.

![Figure 2: Accuracy of VGG-16-SDN on the Tiny ImageNet task. Each blue marker denotes the accuracy of a classifier in the SDN, the last one being the end classifier; we plot the accuracy (left axis) against average FLOPs. Dotted lines denote the accuracy and average FLOPs of the baseline VGG-16. Bars represent the fraction of the baseline’s correct predictions that are classified correctly by our SDN (right axis).](image2)

![Figure 3: Sample images from the Tiny ImageNet classes Golden Retriever and Persian Cat. On the left, we present the simple images that the first internal classifier of the VGG-16-SDN classifies correctly. On the right, we present the difficult images that the first internal classifier misclassifies, but the end classifier correctly classifies.](image3)

While a CNN can overthink on the simpler instances, the internal classifiers can underthink on the difficult ones. To illustrate this phenomenon, in Figure 3, we show randomly sampled validation images from two classes. The left side shows the simple images the first internal classifier of the VGG-16-SDN classifies correctly. On the right side, we present the difficult images the first internal classifier misclassifies. These images suggest that the first internal classifier learns simple representations that allow it to recognize typical instances from a class, but fail on atypical instances, e.g. where the subject is occluded or zoomed out. Overthinking leads to wasted computation as the network applies unnecessarily complex representations. On the other hand, not having enough representational complexity causes underthinking, thus, wrong internal predictions. Interestingly, simple representations appear to work best on some of the difficult images. We identified 102 instances in the validation data that can only be correctly classified by the first internal classifier. All other classifiers in the SDN and the original CNN misclassify these instances. Moreover, this behavior is consistent among the other networks we experimented with. Figure 4 shows a random selection of these
images along with their ground truth labels and misclassified labels by the end classifier. We hypothesize that these images consist of confusing elements, sometimes belonging more than one class; while the first internal classifier recognizes objects displayed prominently (e.g. the uniform and the snorkel), subsequent layers discover finer details that are irrelevant (e.g. the binoculars and the bikini). As it does in humans, overthinking leads to confusion in these cases and potentially causes the network to misclassify. These results suggest that correct classifications require the appropriate level of representational complexity for each instance, which further illustrates the utility of the SDN’s internal classifiers. Building on this idea, in Section 4.6 we visually investigate the sources of the confusion and diagnose these mistakes.

4.3 The Accuracy of Internal Classifiers
Table 2 presents the accuracy of the internal classifiers compared to the original CNNs. Our findings here support our hypothesis that overthinking is prevalent across off-the-shelf models. We see that internal classifiers exceed 80% of the original accuracy but need less than half of the computation. Furthermore, SDN tuning, by imposing additional constraints, actually improves the performance consistently for all off-the-shelf models, throughout all tasks. The improvement is even more emphasized in the more difficult tasks; reducing the error rate by up to 9.1%. SDN tuning, similar to [Lee et al. 2015], forces highly discriminative features in internal layers; therefore, a classifier trained on such features perform better than a classifier trained on less discriminative features. Moreover, in almost all cases, the accuracy actually peaks before the end, at an internal classifier. This enables pruning of the layers preceding the peak, without hurting the original performance. Pruning these layers can reduce the size of an SDN by up to 50%; combined with the early predictions, this can further reduce the computation needed.

4.4 Making Early Predictions
Training accurate internal classifiers is not enough for making early predictions; we must also decide when to stop thinking. We consider three settings for determining early prediction thresholds, \( \theta \): efficient, balanced and accurate. We set the thresholds to achieve at least 95%, 99% and 100% of the original CNN’s accuracy in the efficient, balanced and accurate settings, respectively, while reducing the required operations the most. We set these thresholds by searching for the appropriate \( q \) values, i.e probability that an instance reaching to an internal classifier obtains an early prediction, that achieve the desired accuracy in each setting. These thresholds facilitate on-the-fly adjusting the network based on the resource availability and the performance requirements. We present these results in Table 3 comparing the accuracy and average FLOPs in each setting to the off-the-shelf CNNs. We observe that SDNs reduce the required computation by an average of 41%, 34% in 25% in the efficient, balanced and accurate settings, respectively. Even in our most difficult benchmark, Tiny ImageNet, SDNs can reliably make early predictions, thus, reducing the required computation by 26% in the balanced setting. Furthermore, we also see that network complexity leads to diminishing returns. A 13% increase in average FLOPs pushes the thresholds from the efficient to the balanced setting, resulting in a 5% accuracy increase. We must increase the average FLOPs for another 13% to reach the accurate setting, which only improves the accuracy by an additional 1%.

4.5 Internal Confusion as an Error Indicator
Here, we present the confusion scores the VGG-16-SDN produces on the Tiny ImageNet validation data. As a baseline, we use the prediction probabilities, \( \max_y F_{(M+1)}(x) \), the original VGG-16 produces on the validation set. This probability is commonly referred as the confidence score of a DNN, and it constitutes an alternative to our confusion scores for determining the likelihood of prediction error.

Table 2: Accuracy comparison between CNNs and corresponding SDNs. In the column Orig, we present the accuracy of the original CNN. In the columns +80% and +99%, we present the accuracy and the avg. FLOPs, relative to the original CNN (as a percentage), of the internal classifier that the first exceeds 80% and 99% of the original accuracy, respectively. Finally, in the column Max, we present the internal classifier with the maximum accuracy and its relative avg. FLOPs. We highlight the cases where an internal classifier actually outperforms the original CNN.
Table 3: Comparing the CNNs and corresponding SDNs with early predictions, based on the average gigaFLOPs required per instance. Column Orig presents the average GFLOPs for the original CNN. The next three columns present the average GFLOPs for SDNs with efficient, balanced or accurate early prediction thresholds, along with their corresponding accuracies.

| Ntw. | Orig | Efficient | Balanced | Accurate |
|------|------|-----------|----------|----------|
|      |      |           |          |          |
| CIFAR-10 |      |           |          |          |
| VGG  | 0.63 | 0.29 | 92.7  | 0.30 | 93.2  | 0.36 | 93.5  |
| Res. | 0.25 | 0.12 | 87.6  | 0.14 | 90.9  | 0.16 | 91.6  |
| WRN  | 2.37 | 1.22 | 90.9  | 1.49 | 93.8  | 1.93 | 94.5  |
| Mob. | 0.75 | 0.18 | 85.8  | 0.25 | 89.8  | 0.31 | 90.6  |
| CIFAR-100 |      |           |          |          |
| VGG  | 0.63 | 0.28 | 68.4  | 0.31 | 69.9  | 0.33 | 70.6  |
| Res. | 0.25 | 0.16 | 64.9  | 0.19 | 68.1  | 0.20 | 68.9  |
| WRN  | 2.37 | 1.52 | 72.4  | 1.62 | 73.9  | 1.73 | 75.0  |
| Mob. | 0.75 | 0.25 | 61.4  | 0.34 | 64.2  | 0.39 | 64.8  |
| Tiny ImageNet |      |           |          |          |
| VGG  | 2.52 | 1.58 | 57.2  | 1.68 | 58.4  | 1.78 | 59.4  |
| Res. | 0.25 | 0.19 | 51.4  | 0.20 | 53.4  | 0.23 | 54.1  |
| WRN  | 2.37 | 1.72 | 57.4  | 1.89 | 59.7  | 2.01 | 60.1  |
| Mob. | 3.03 | 1.96 | 57.1  | 2.27 | 58.6  | 2.60 | 58.7  |

In our experiment, we aim to investigate if confusion scores is a reliable indicator for the cases when the network misclassifies an input. Such indications have practical significance for handling errors; for example, they can alert users about cases where the network is unable to make a good prediction. Figure 5 compares the distributions of confusion scores for correct and wrong predictions. While correct predictions tend to have low confusion scores (∼0.14 on average), the misclassifications are concentrated among instances with high confusion (with an average score of 1.02). In contrast, the difference in confidence scores is less pronounced: 0.93 vs. 0.68 on the correct and wrong predictions, respectively (distribution omitted owing to space limits). When used as an indicator for likely misclassifications, confusion produces fewer false negatives than confidence. Compared to an average correct prediction, 6% of the misclassified instances (220 out of 3991) cause less confusion for the SDN, whereas 23% (941 out of 4104) misclassified instances obtain more confidence in the original CNN.

4.6 Visualizing Confusion for Diagnosing Errors

Figure 4 shows some images that confuse our VGG-16-SDN. On these images, the first internal classifier, which predicts the correct labels, and the end classifier, which misclassifies, disagree. We hypothesize that the existence of confusing elements in the input triggers disagreements and potential misclassifications. In Figure 5, we aim to identify these elements, by utilizing the Grad-CAM visualization (Selvaraju et al. 2017) of the disagreeing first and end predictions. As an example, consider the beaker vs. pill bottle image. Here, we see that the first internal classifier focuses on the simple structure of the beaker, which leads to a correct prediction. The finer details, such as text on the beaker and its top opening, cause the end classifier to confuse the beaker with a pill bottle. Visualizing the confusion helps us to better understand how the network makes incorrect decisions, providing a new perspective towards interpretable deep learning.

5 Conclusions

We propose the Shallow-Deep Network (SDN), a modification to off-the-shelf deep neural networks for obtaining predictions from a network’s internal layers. Our modification consists of attaching classifiers to the internal layers after reducing their output size and using a layer-wise objective function to train these classifiers. In a range of popular convolutional neural network architectures, SDNs enable early predictions for simpler inputs without retraining; thus, significantly reducing the average inference cost while preserving the original performance. Furthermore, we utilize SDNs for defining and exploring a notion of overthinking in deep neural networks. We also define a confusion metric based on the level of disagreement among internal classifiers, and we show experimentally that high confusion is associated with misclassifications. This allows us to interpret these errors by diagnosing the input elements that trigger them. In the future, we plan to explore means to avoid confusion, e.g. by...
making predictions from the simplest possible features in the case of confusing input. We also plan to apply SDN to other neural network architectures, such as generative models.

6 Acknowledgements

We would like to thank Dr. Tom Goldstein, Dr. Nicolas Papernot, Virinchi Srinivas and Sanghyun Hong for their valuable feedback. This research was partially supported by the Department of Defense.

References

[Deng et al. 2009] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, 248-255. Ieee.

[Gal and Ghahramani 2016] Gal, Y., and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In International Conference on Machine Learning, 1050-1059.

[Guo et al. 2017] Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. arXiv preprint arXiv:1706.04599.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.

[Hinton et al. 2012] Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine 29(6):82–97.

[Howard et al. 2017] Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861.

[Huang et al. 2017] Huang, G.; Chen, D.; Li, T.; Wu, F.; van der Maaten, L.; and Weinberger, K. Q. 2017. Multi-scale dense convolutional networks for efficient prediction. CoRR, abs/1703.09844 2.

[Kahneman 2011] Kahneman, D. 2011. Thinking, fast and slow. Farrar, Straus and Giroux / New York.

[Krizhevsky and Hinton 2009] Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical report, Citeseer.

[Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 1097-1105.

[Lee et al. 2015] Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; and Tu, Z. 2015. Deeply-supervised nets. In Artificial Intelligence and Statistics, 562-570.

[Lee, Gallagher, and Tu 2016] Lee, C.-Y.; Gallagher, P. W.; and Tu, Z. 2016. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In Artificial Intelligence and Statistics, 464–472.

[Lin et al. 2017] Lin, J.; Rao, Y.; Lu, J.; and Zhou, J. 2017. Runtime neural pruning. In Advances in Neural Information Processing Systems, 2181–2191.

[Olah et al. 2018] Olah, C.; Satyanarayan, A.; Johnson, I.; Carter, S.; Schubert, L.; Ye, K.; and Mordvintsev, A. 2018. The building blocks of interpretability. Distill. https://distill.pub/2018/building-blocks.

[Papernot and McDaniel 2018] Papernot, N., and McDaniel, P. 2018. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. arXiv preprint arXiv:1803.04765.

[Selvaraju et al. 2017] Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In ICCV, 618–626.

[Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

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

[Szegedy et al. 2013] Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[Szegedy et al. 2015] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1–9.

[Wang et al. 2017] Wang, X.; Luo, Y.; Crankshaw, D.; Tumanov, A.; Yu, F.; and Gonzalez, J. E. 2017. Idk cascades: Fast deep learning by learning not to overthink. arXiv preprint arXiv:1706.00885.

[Yosinski et al. 2014] Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? In Advances in neural information processing systems, 3320–3328.

[Zagoruyko and Komodakis 2016] Zagoruyko, S., and Komodakis, N. 2016. Wide residual networks. arXiv preprint arXiv:1605.07146.

[Zeiler and Fergus 2014] Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In European conference on computer vision, 818–833. Springer.