Shallow-Deep Networks: Understanding and Mitigating Network Overthinking

Yigitcan Kaya, Sanghyun Hong, Tudor Dumitras
University of Maryland, College Park

THE OVERTHINKING PROBLEM

Humans are susceptible to overthinking.

- Thinking more than needed to solve a problem
- Wastes our energy and leads to slow decision-making.
- Causes confusion and potential mistakes

We ask *are deep neural networks also susceptible to overthinking too?*

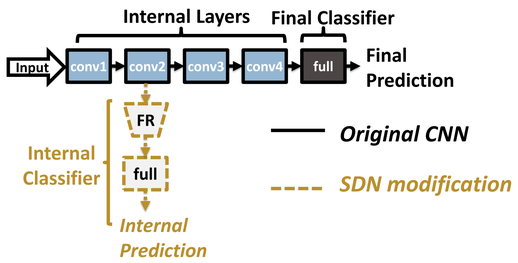
- Experiments on four recent convolutional neural networks and three common image classification tasks
- The answer is **YES.** Without requiring their full depth, DNNs can correctly classify the majority of samples.
- Leads to wasted computation for up to **95%** of the samples. (the *wasteful* effect of overthinking)
- Occurs in **~50%** of all misclassfications. (the *destructive* effect of overthinking)

SHALLOW-DEEP NETWORKS

Detecting overthinking requires producing *internal predictions* from the earlier layers of a DNN.

We propose Shallow-Deep Networks (SDNs)

- A *generic* modification that introduces internal classifiers to *off-the-shelf* DNNs
- *Internal classifiers* throughout the DNN's forward-pass
- Applied to both pre-trained and untrained DNNs
- *Feature reduction* (FR) for scalability



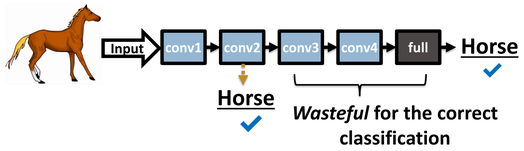
Training the internal classifiers

If the *off-the-shelf* DNN is pre-trained, we only train the weights in the internal classifiers.

If the *off-the-shelf* DNN is untrained, we jointly train the original weights and the internal classifiers

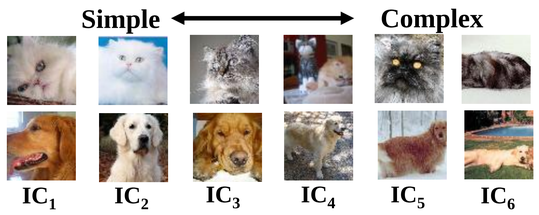
- Prior work[1] claims this is challenging without sacrificing accuracy.
- Our *weighted objective function* even *improves* the original accuracy by up to **10%**.

THE WASTEFUL EFFECT OF OVERTHINKING



The majority of the samples **do not require** the full depth of the DNN.

- For these samples a full forward-pass is *wasteful*.
- Only **5%** of the samples for CIFAR-10, **19%** for CIFAR-100 and **31%** for Tiny ImageNet **do require** the full depth.



VGG16-SDN-TinyImageNet. The test samples that are first correctly classified by each internal classifier (IC). Notice how progressively complex the samples get over ICs. Using the full forward-pass for simple samples would be wasteful.

Confidence-based early exits mitigate the wasteful effect and reduce the average inference cost by up to 50%

It is not possible to know whether an internal prediction is correct.

- How can we tell whether we should stop the forward-pass and make an *early exit*?
- How can we realistically prevent the computational waste?

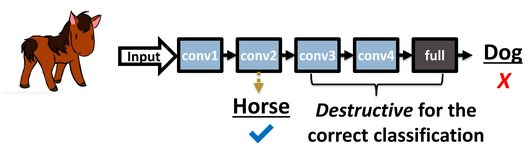
We use *internal prediction confidence* for making early exits.

- Tunable *confidence threshold* to determine whether to stop or continue.

Network (Tiny ImageNet)	Orig. Acc.	<25% Cost	<50% Cost	<75% Cost	Max. Acc.
VGG16	59%	37%	56%	**63%**	**63%**
ResNet	54%	39%	39%	53%	**55%**
WRN	60%	37%	55%	**63%**	**63%**
MobileNet	59%	45%	57%	**62%**	**62%**

Accuracies of the SDNs with early exits. We set the confidence threshold to adjust average inference cost.

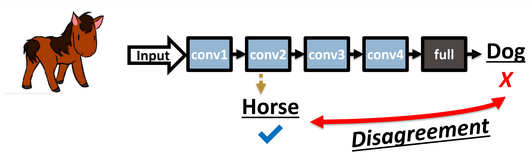
THE DESTRUCTIVE EFFECT OF OVERTHINKING



The *cumulative accuracy* of an SDN also includes the correctly classified samples at the internal classifiers.

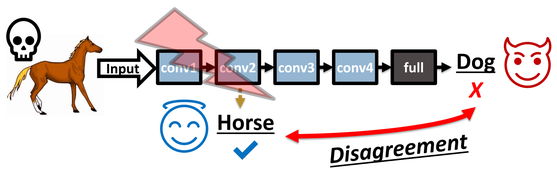
- The difference between cumulative and regular accuracy quantifies the destructive effect. (**4%** on CIFAR-10, **13%** on CIFAR and **14%** on Tiny ImageNet)
- The destructive effect occurs in **~50%** of all misclassifications.

The destructive effect causes internal disagreement



Notice how the destructive effect manifests itself as an *internal disagreement* between an internal classifier and the final classifier.

Backdoor attacks maliciously induce disagreement



A recent backdoor attack [2] also induces disagreements on the victim DNN for the malicious samples

- The victim's accuracy on the malicious test samples is only **12%**, however, with the confidence-based early exits the accuracy increases to **84%**.

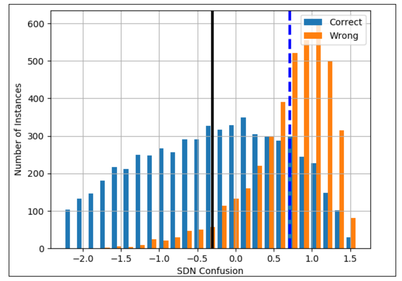
THE CONFUSION METRIC

The destructive effect induces disagreements and it is prevalent in misclassifications.

The *confusion* metric quantifies *how much the final prediction diverges from the internal predictions.*

Confusion is a reliable error indicator

VGG16-SDN-TinyImageNet. Network shows significantly less confusion when it correctly classifies sample (**-0.29** for average correct vs **0.71** for average wrong).



Visualizing the final classifier and the disagreeing internal classifier reveals confusing input features.

CONCLUSION & IMPLICATIONS

Deep neural networks overthink. This causes computational waste and avoidable mistakes.

- We mitigate overthinking to some extent.
- This improves the inference cost and accuracy

Eliminating overthinking would lead to *a significant boost* in accuracy and inference cost.

- We need DNNs that can *adjust their complexity* based on the *required feature complexity*.

CONTACT & CODE & EXTRAS

http://shallowdeep.network

REFERENCES

[1] Huang, Gao, et al. "Multi-Scale Dense Convolutional Networks for Efficient Prediction." *ICLR 2018*

[2] Gu, Tianyu, et al. "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks." *IEEE Access* 7 (2019): 47230-47244.