EECS 182 Deep Neural Networks
Fall 2025 Anant Sahai and Gireeja Ranade

Discussion 12

1. Entropy, Cross-Entropy, Kullback - Leibler (KL)-divergence

Entropy is a measure of expected surprise. For a given discrete Random variable Y, we know that from Information Theory that a measure the surprise of observing that Y takes the value k by computing:

$$\log \frac{1}{p(Y=k)} = -\log[p(Y=k)]$$

As given:

- if $p(Y=k) \to 0$, the surprise of observing k approaches ∞
- if $p(Y=k) \to 1$, the surprise of observing k approaches 0

The Entropy of the distribution of Y is then the expected surprise given by:

$$H(Y) = E_Y\Big[-\log\big(p(Y=k)\big)\Big] = -\Sigma_k\Big[p(Y=k)\log[p(Y=k)]\Big]$$

On the other hand, Cross-entropy is a measure building upon entropy, generally calculating the difference between two probability distributions p and q. it is given by:

$$\begin{aligned} H(p,q) &= E_{p(x)}\Big[\frac{1}{\log\big(q(x)\big)}\Big] \\ &= \Sigma_x\Big[p(x)\log[\frac{1}{q(x)}]\Big] \end{aligned}$$

Relative Entropy also known as KL Divergence measures how much one distribution diverges from another. For two discrete probability distributions, p and q, it is defined as:

$$D_{KL}(p||q) = \Sigma_x\Big[p(x)\log[\frac{p(x)}{q(x)}]\Big]$$

(a) Let's define the following probability distributions given by:

$$p(x) = \begin{cases} 0.5 & \text{when x = 1} \\ 0.5 & \text{when x = -1} \end{cases}, \text{ and } q(x) = \begin{cases} 0.1 & \text{when x = 1} \\ 0.9 & \text{when x = -1} \end{cases}.$$

Note that $H(p) = -\log 0.5 = 1$ bit and $H(q) = -0.1\log 0.1 - 0.9\log 0.9 \approx 0.47$ bits. Show that KL-divergence is not symmetric and hence does not satisfy some intuitive attributes of distances.

(b) Re-write $D_{KL}(p||q)$ in term of the Entropy $H(p)$ and the cross entropy $H(p,q)$.

2. Catastrophic Forgetting

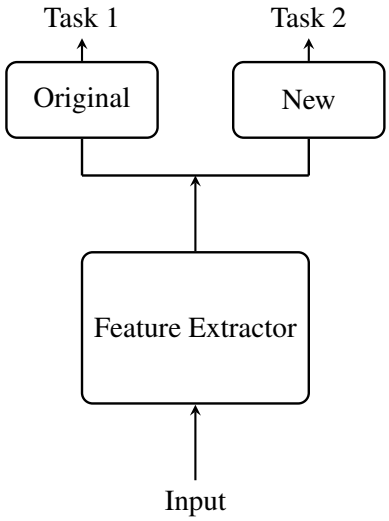
The neural networks are vulnerable to the distributional shift. Many questions in AI/ML are related to the distributional shift: out-of-distribution (OoD), domain adaptation/generalization, meta-learning, and so on. In this discussion, we study one of those problems, catastrophic forgetting, alternatively called catastrophic interference. The catastrophic forgetting is the tendency of neural networks to lose information about previously learned tasks when learning the new one. This is also referred to as stability-plasticity dilemma[1]. One potential drawback of a model that is too stable is that it will not be able to consume new information from the future training data. Conversely, a model with too much plasticity may suffer from large weight changes and forgetting of previously learned representations.

(a) What happens in parameter space

Figure 1a is the cartoon parameter space of the given model and tasks. The dark orange oval is the simplified low-loss region of task 0, and the light blue oval is that of task 1. The model is trained on the task 0, and θ_0^* is the trained parameter of the model, which minimizes the loss for that task. We now train the model with task 1 data.

i. Mark the θ_1^* on the Figure 1a, which is the trained parameter for task 1 if the model is too plastic.
ii. Mark the $\tilde{\theta_1}$ on the Figure 1a, which is the trained parameter for task 1 if the model is too stable.
iii. What are the problems if the model is too plastic or stable?
iv. Where do we want to go?

(a) The cartoon parameter space for feature extractor.



(b) Feature extractor architecture with task-specific heads.

Figure 1: Catastrophic forgetting: (a) Parameter space visualization showing loss landscapes for different tasks, and (b) Feature extractor architecture with task-specific heads.

[1] https://www.sciencedirect.com/science/article/pii/S1364661399012942

(b) Dealing with catastrophic forgetting

We consider three traditional approaches for learning the new tasks: feature extraction, fine-tuning, and joint training.

i. frozen feature extraction - The model trained with the previous tasks is frozen. The output of this model with the new task input is used for to train the new classifier for the new task.

ii. full fine-tune on new task - The model trained with the previous tasks is trained with the new task. To prevent the large shift, the learning rate is typically low.

iii. joint training full fine-tune - The model is trained with both previous task data and new task data

Let's study pros and cons of those methods. Fill in the table below.

Category	frozen feature extraction	full fine-tune on new task	joint training
New task performance			Good
Old task performance	Good		Good
Storing old task data	No	No	

3. How to read research papers

One critical skill for improving yourself in deep learning is to quickly read papers. Being able to read papers efficiently and to identify the key contributions regardless of what the authors claimed are some critical skills as you work in a field where some fundamental theory is still an open problem. In this question, you will be reading Learning without Forgetting (Li et al.), a paper that proposes a simple but effective strategy to alleviate the problem of Catastrophic Forgetting.

But first, let's talk about how to read a deep learning paper. Every researcher has a different strategy, but one recurring pattern is to *read with multiple passes*.

- First pass: Read the **title**, **abstract** and **figures**
- Second pass: Read the **introduction**, **conclusion** and **figures** again
- Third pass: Read the **method** (skip or skim the math) and skim over the **results** (skip ablation study)
- Fourth pass: Read everything else but skip parts that do not make sense (unless you are doing research in that particular field)

Now spend some time to reading the Learning without Forgetting paper with this multiple-pass strategy, and discuss with your classmates for the following questions.

(a) What was the challenge authors are trying to overcome?

(b) How is Learning without Forgetting (LwF) different from standard fine-tuning and joint training (multitask learning)?

(c) Is LwF better than feature extraction in both new and old tasks? How does LwF compare with joint training in terms of accuracy?

(d) Does the new task dataset size affect LwF's effectiveness? What about the old task dataset size?

Contributors:

- Jerome Quenum.
- Anant Sahai.
- Suhong Moon.
- Kumar Krishna Agrawal.
- Sultan Daniels.
- Kevin Li.

