CS 7150: Deep Learning — Summer-Full 2020 — Paul Hand

Week 7 — Preparation Questions For Class
Due: Monday June 22, 2020 at 12:00 PM Eastern time via Gradescope

Name: Saurabh Vaidya
Collaborators: Sumeet Gajjar

You may consult any and all resources in answering the questions. Your goal is to have answers that are ready to be shared with the class (or on a hypothetical job interview) as written. **Make sure to tag each question when you submit to Gradescope.**

Directions: Read the article 'Overcoming catastrophic forgetting in neural networks'.

Question 1. *Provide a summary of the contributions of this paper.*

Response:

Answer 1. *This paper provides an algorithm that aims to make continual learning possible without having to learn the joint distributions of two different task and rather allows for sequential learning of two tasks. This paper presented an algorithm EWC, which determines weights crucial for a task, retains them for the next task such that our models perform better on both tasks. They ran this algorithm on supervised and reinforcement learning setting.*

This paper also sheds light on how our artificial networks can take inspiration or in some cases a direct implementation of biological neurological processes.

Question 2. *Explain Figure 1. Be sure to mention what the ambient space represents, why the shapes are depicted as ellipses, and why these ellipses intersect.*

Response:

Answer 2. *The space represents parameter space θ for a task such that the error due to parameters in that region is low.*

The shapes are ellipses because the authors approximate the posterior distribution of θ with respect to Data D as a multi variate gaussian distribution. Since the precision of the gaussian is defined by a matrix which is Positive Semi Definite(PSD), the projections of contour of a multivariate gaussian are elliptical.

*PSD matrix allows for eigen decomposition and positive eigen values. Level set $\mathcal{X}$ is a set of vectors x where every vector $x \in \mathcal{X}$ has the same value for its density function $c = p(x)$. Every contour of a density function is the shape of **this** level set. Level set for gaussian distribution is given by*

$$L(x) = x\Sigma^{-1}x^T \tag{1}$$

where $x = x + \mu$ and μ is the mean of our gaussian

Since fisher matrix is a covariance of score of our mdoel and PSD, an eigen decomposition of a fisher matrix allows the Level Set of our density function to be

$$L(x) = x\Lambda x^T \tag{2}$$

where Λ is diagonal matrix of positive eigen values.

Because Λ is a diagonal matrix we can write

$$L(x) = \sum_{d=1}^{D} \lambda_d . x_d^2 \tag{3}$$

where D is dimensions of our vector x.

Now finding the level set in 2D,

$$L(x) = \lambda_1 x_1^2 + \lambda_2 x_2^2 \tag{4}$$

which transforms into equation of ellipse if we denote L(x) with a constant C.

The equation becomes

$$1 = \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} \tag{5}$$

where $a^2 = \frac{C}{\lambda_1} b^2 = \frac{C}{\lambda_2}$

See Reference

As the neural networks are over-parameterized, there exist multiple configurations of θ_B that have same performance on task B. This means that the parameters spaces for our tasks A and B would be quite big thereby increasing chances of them intersecting. Thus EWC can find the θ_B in parameter space of A by constraining the θ_A parameters to change for task B while being in the space of θ_A.

Question 3. *In this paper, the authors modify the loss function for task B based on the parameters of a network trained for task A. How does this formulation identify which weights of the neural network are most important for task A? Make sure to comment on whether high Fisher information or low Fisher information indicates importance for task A.*

Response:

Answer 3. *The loss of the overall problem of learning two tasks will be minimized when both parts of the equation are minimized i.e. $\mathcal{L}(\theta_B)$ and $\sum_i \frac{\lambda}{2} . \mathcal{F}_i (\theta_i - \theta^*_{A,i})^2$*

The second term of the equation will be minimized when θ_i is close to $\theta_{A,i}$ since the difference between them is squared. However weighing these squared distances is done where each weight determines the importance of weight for task A.

A high fisher information value denotes that this particular value of parameter tell us a lot about the observed data. In other words higher fisher information is obtained when our observed data is actually modelled by our parameter θ_i i.e true $\theta^* = \theta_i$. Thus we want to retain those $\theta_{A,i}$'s that gave more fischer information

Question 4. *How is the algorithm in this paper biologically inspired?*

Response:

Answer 4. *Biologically when an animal learns a new skill, the excitatory synaptic connections are strengthened by increasing more volume of dendritic spines of a neuron. It has been revealed that these connections are retained on learning new tasks and erasing these connections also leads to forgetting of*

previously learned skill. Thus continual learning in mamals relies on process specific synaptic connections. The knowledge gained in previous task is encoded in proportion of synapses.

The EWC algorithm inspires from the same ideology where synaptic connections correspond to weight parameters in artificial neural networks. Thus retention(strengthening) of weights important to previously learned tasks allows for continual learning on other tasks.

Question 5. *Why are randomly permuted MNIST classification problems claimed to be equally hard? Wouldn't a human find it much more difficult to classify images with randomly permuted pixels?*

Response:

Answer 5. *When we shuffle all pixels, but not change the value of each pixel, neural network weights associated with those pixels will learn the same values are before. Thus even though the weight vector would look different it will still be from the same solution space. Thus for the network both tasks are equally difficult because in both cases the network has to learn new sets of weights for each pixel 'position'. Moreover because the network was MLP which does not learn high level features such as spatial relationships unlike CNNs, the permutated problem is equally hard for it since it has focus on each pixel level learning.*

For humans the pixel values at a particular position are not discriminative rather the spatial relationships between pixels is what helps same as CNNs. Thus to humans randomly permutated pixel values of digits are hard to classify.

Question 6. *Explain Figure 2. Make sure to include the context, a statement of what literally is plotted, what is to be observed, and what is concluded. Separately explain panels A, B, C.*

Response:

Answer 6. *The authors want to confirm whether EWC allows deep neural networks to learn different tasks sequentially without catastrophic forgetting and if so how it achieves that.*

Figure 2A, plots the performance of SGD, L2 regularization and EWC training to see the extent of catastrophic forgetting in all three algorithms. The fig plots training vs performance on test set on 3 tasks A,B,C. We have to observe how the performance of each algorithms fares on first task as we sequentially keep learning different tasks. The conclusion is SGD performs poorly on task A after it starts training for task B and worsens as task C resumes. It also performs poorly on task B as it trains for task C. L2 regularization seems to maintain its performance level within considerable limits for old tasks but performs poorly on current tasks. EWC maintains steady performance on old tasks and also performs better on current tasks.

Fig 2B compares how standard regularization methods with SGD perform in comparison to EWC when the number of sequential tasks increase. We observe how the accuracy of the model changes as the number of continual tasks increase. Its concluded that EWC has relatively stable performance with only a slight dip towards the end whereas SGD+dropout method's performance starts to drop significantly as number of tasks increases. EWC stays relatively same which is similar to how a model would have performed on a single task.

Fig 2C, the authors want to find whether EWC allocates different set of weights for different tasks or the weights are shared for those tasks. The figure shows the overlap of fisher information between same weights for two tasks in 2 different settings where the permutation of pixels is lot in MNIST versus less

permutation. We want to observe how much is the overlap(i.e fraction of weights that are shared) as the depth of the layer increases for tasks with higher and lesser pixel permutations. Its observed that the overlap between deeper layers is more than shallower layers for tasks that have higher permutation and is almost steady for tasks with lesser permutation. It's concluded that for lesser permutation tasks, weights are shared among two tasks. In higher permutation scenario, earlier layers have distinct sets of weight for two different tasks but as the network gets deeper, the layers close to the output have more overlap and just more weight sharing.

Question 7. *The paper says that the EWC algorithm "can be grounded in Bayesian approaches to learning." Explain in your own words what this comment means.*

Response:

Answer 7. *Bayesian approaches to learning involve learning the posterior distribution of our target given our input ($p(\theta|X)$. In doing so, we use generative processes where we determine the likelihood of our input given a particular target ($p(X|\theta)$) and multiply that with a prior probability of knowing the target ($p(\theta)$).*

EWC roots from the same philosophy where the posterior distribution of weight parameters from previous tasks ($p(\theta_{n_0}|x_0)$) are used as prior probabilities of the same weights(target) for next task i.e $p(\theta_{n_1}) = p(\theta_{n_0}|x_0)$ in order to learn the posterior distribution of weights for next task i.e $p(\theta_{n_1}|x_1)$.