**Red- should be deleted.**

**Yellow – take a look, should be change?**

**Red- I wrote in your section**

**Dry part:**

1

A close-up of a text

Description automatically generated

(a)

The Universal Approximation Theorem (UAT) states that given enough parameters, a multi-layer perceptron (MLP) with one hidden layer and a non-linear activation function can approximate any continuous function on a compact subset of to arbitrary precision.

Specifically, for (a continuous function

What is the following ? a proof of a finite sum of sigmoidal functions can densely approximate any continuous function ? (take a look on this : <https://medium.com/@ML-STATS/understanding-the-universal-approximation-theorem-8bd55c619e30> )

there exists a function of the form that can approximate to any given precision. Here, are parameters that adjust to the required level of precision.

This theorem supports the claim that you can achieve optimal error over any dataset and loss criterion. If there exists an optimal function that minimizes the loss on a given dataset, he UAT guarantees that there exists a single hidden layer MLP that can approximate to any desired precision. This means that such an MLP can approximate

closely enough to behave in the same way on the dataset. As a result, using an MLP with one hidden layer provides a potential pathway to achieving the minimal loss.

\*If the loss function does not have a finite minimum, we can refer to as the function that achieves the desired level of loss.

(b)

The conclusion that one should "never use more than one hidden layer" is incorrect because the UAT is simply an existence theorem. While it guarantees that an MLP with one hidden layer can approximate any continuous function, it does not provide a practical way to find this approximation. There are several reasons why using only one hidden layer might not be the best choice in practice:

* Unknown Number of Neurons: The UAT does not specify the number of neurons required in the hidden layer to achieve the desired precision. In practice, k could be extremely large, making the network impractically wide and computationally expensive.
* Optimization Challenges: The way we train neural networks (e.g., with backpropagation and gradient descent) does not guarantee finding the optimal solution. The optimization process can get stuck in local minima, saddle points, or regions with poor gradients, preventing the network from achieving the desired approximation.(you said that )
* Empirical Success of Deep Networks: While the UAT applies to single hidden layer networks, deeper networks (those with more than one hidden layer) have been found empirically to perform better. They can learn hierarchical features, leading to better generalization, more efficient learning, and often better performance in real-world tasks.

Rephrase your answer in (b) : (on (a) I should work afterwards) :

The conclusion that one should "never use more than one hidden layer" is incorrect because the Universal Approximation Theorem (UAT) is primarily an **existence theorem**. While it guarantees that a single-hidden-layer neural network can approximate any continuous function on a compact domain, it does not offer practical guidance on how to construct or train such a network for real-world applications. Here are several reasons why using more than one hidden layer is often necessary:

* **Existence vs. Practical Implementation**: The UAT is an **existence theorem**. It guarantees that a single-hidden-layer network **can** approximate any continuous function given the right parameters and enough neurons. However, it doesn't provide a practical method for constructing such a network or specify how to find the optimal weights and biases needed for the approximation.
* **Unknown and Potentially Large Number of Neurons**: The theorem does not indicate how many neurons are required to achieve the desired level of precision. In practice, approximating complex functions might require an extremely large number of neurons in the hidden layer. This makes the network **impractically wide**, leading to high computational costs, increased training time, and greater resource consumption.
* **Optimization Challenges**: Training a very wide, shallow network introduces significant optimization difficulties. Gradient-based learning algorithms like backpropagation may struggle with such networks due to issues like vanishing or exploding gradients. The optimization process can get stuck in local minima, saddle points, or flat regions in the error landscape, preventing the network from finding the optimal solution and achieving the desired approximation.
* **Hierarchical Feature Learning**: Deep neural networks with multiple hidden layers are capable of learning **hierarchical representations** of data. Early layers capture low-level features, while subsequent layers build upon them to recognize more complex patterns. This layered learning is crucial for tasks involving high-dimensional data with intricate structures, such as image and speech recognition. Single-hidden-layer networks lack the depth to model these hierarchical and compositional patterns effectively.
* **Empirical Success of Deep Networks**: Empirically, deep networks have been found to outperform shallow networks across various domains. They tend to **generalize better** to unseen data and achieve higher performance metrics. This success is attributed to their ability to model complex, non-linear relationships more effectively than shallow networks, even those with a large number of neurons in a single hidden layer.

A white paper with black text

Description automatically generated

(a)

Bob can convince Alice by explaining that the main advantage of using a Convolutional Neural Network (CNN) over a Multilayer Perceptron (MLP) for image classification is that CNNs are specifically designed to exploit the spatial structure of images. CNNs utilize local connectivity and weight sharing to effectively capture spatial hierarchies and local patterns such as edges, textures, and shapes, which are crucial for understanding and classifying images.

Main Difficulties Alice Might Encounter If She Uses an MLP:

* **Loss of Spatial Relationships**: MLPs require images to be flattened into one-dimensional vectors, which destroys the two-dimensional spatial arrangement of pixels. This loss of spatial information makes it challenging for the network to learn and recognize patterns that depend on the relative positions of pixels, leading to poorer performance in image classification tasks.
* **High Computational Complexity and Number of Parameters**: flattening high-resolution images results in extremely large input vectors. An MLP would need to connect each input pixel to neurons in the next layer, leading to a massive number of weights and biases. This not only increases computational requirements and memory usage but also makes the network more difficult to train effectively.
* **Risk of Overfitting**: with such a large number of parameters, the MLP is more prone to overfitting the training data. It may learn to memorize the training images rather than generalizing to new, unseen images, resulting in poor performance on validation or test datasets.

(b)

I disagree with Alice's conclusion. While it's true that the convolution operation is linear, the crucial difference lies in **how these linear operations are structured and utilized within the network architectures**, which significantly impacts their capabilities. As mentioned earlier, CNNs are designed to **preserve the spatial structure of images** through local connectivity and weight sharing. This architectural design enables CNNs to efficiently learn spatial hierarchies and patterns essential for image classification, such as edges and textures, regardless of their position in the image.

In contrast, MLPs require images to be flattened into one-dimensional vectors, which destroys the spatial relationships between pixels. This loss of spatial information prevents MLPs from effectively learning patterns that depend on pixel proximity. Additionally, MLPs do not employ weight sharing, resulting in a much larger number of parameters compared to CNNs, which can lead to increased computational complexity and a higher risk of overfitting.

Therefore, even though both CNNs and MLPs involve linear operations between activations, the **architectural differences**—specifically, how these linear operations are applied—make CNNs far more effective for image classification tasks. CNNs can capture and utilize spatial features in images that MLPs cannot, due to their ability to maintain spatial hierarchies and apply learned features across different parts of the image. This fundamental difference means that CNNs and MLPs are not equivalent in practice for image classification, despite the linearity of their operations.

A black text on a white background

Description automatically generated

3.

We would expect momentum to help in the optimization process, even for convex optimization problems. The momentum method utilizes an exponential moving average of gradients from previous steps to update the parameters in the current step. This means the current parameter is adjusted based on both the current gradient and past gradients. In convex optimization, the negative gradient consistently points toward the same global minimum. As a result, the momentum method accelerates the updates in the direction of the minimum, leading to faster convergence.  
In addition, the momentum method can smooth out updates in cases of oscillations—such as when the learning rate is too large, and the gradients change direction in each step. This smoothing effect leads to a more stable optimization process.

For example, as we learned in the course on numerical algorithms, when applying the steepest descent method to least squares problems (which are convex), using a normalized learning rate in each step can cause the gradient at each step to become perpendicular to the previous gradient. This phenomenon slows down convergence because the optimizer tends to zigzag across the loss landscape rather than moving directly toward the minimum. Introducing momentum helps combat this issue by accumulating the direction of previous gradients, allowing the optimizer to maintain a more consistent and directed path toward the global minimum.

However, it’s important to note that momentum is not always superior to standard gradient descent. If the learning rate or momentum factor is not well-tuned, it can lead to overshooting or slower convergence.

3.

Yes, using momentum can help the optimization process when dealing with a convex loss function that has a single global minimum. momentum accelerates convergence by incorporating an exponential moving average of past gradients into the current update, adjusting parameters based on both current and previous gradients.

in convex optimization, where the negative gradient consistently points toward the global minimum, this accumulation of gradients leads to faster and more direct movement toward the minimum.

momentum also smooths out oscillations that may occur when gradients change direction between steps, especially if the learning rate is large, resulting in a more stable optimization process. For example, in steepest descent methods applied to least squares problems, gradients can become nearly perpendicular between iterations, causing the optimizer to zigzag and slow down.

Introducing momentum helps combat this issue by accumulating the direction of previous gradients, allowing the optimizer to maintain a more consistent and directed path toward the global minimum.

However, it’s important to note that momentum is not always superior to standard gradient descent. If the learning rate or momentum factor is not well-tuned, it can lead to overshooting or slower convergence.

A black text on a white background

Description automatically generated

The **loss function** is a **scalar** value that summarizes the performance of the model. When training a model, we usually want to minimize this scalar loss, which measures how well the model is doing (e.g., mean squared error, cross-entropy loss). The goal of calling .backward() is to compute the **gradient of the loss** with respect to the model's parameters.

The gradient of a scalar with respect to a vector of parameters is well-defined: it's a vector (or matrix for multiple layers).

However, if the output is **not a scalar** (i.e., it's a vector or matrix), then technically, the gradient would be a **tensor of higher dimensions** (i.e., you'd get a Jacobian matrix rather than just a gradient vector). This would involve calculating and storing the gradient for every output with respect to every parameter, which would be much more complex and expensive.

so, when the output is scalar, you calculate a gradient (a vector), but if the output were a vector, you'd have to calculate a much more complex and expensive structure (i.e. Jacobian matrix), PyTorch simplifies this by requiring a scalar output when calling .backward(),

A text on a white background

Description automatically generated

A white background with text and numbers

Description automatically generated

Given single channel image with pixels denote as x, a model consists from 2D convolutional layer with input and output layer equal to 1 and kernel .

For each input pixel at position in the image, the convolution layer applies a set of learnable filters to compute an output

Since the **kernel size** is , the filter covers the **entire input image** in a single step, therefore the output is a scalar (size of ).

Let:

- The kernel be:

- The input be:

- The bias is a scalar .

The model output is given by:

.

Since the loss computes the difference between an y, and is a scalar while y is a 32x32 matrix, broadcasting is applied. Broadcasting repeats the scalar ​ over all cells of the matrix y during the subtraction.

The loss term is therefore:

Using the chain rule :

Given that the kernel size has dimensions 32x32 (the same as the dimensions of the input feature map), this convolutional layer is equivalent to a **linear layer with one neuron.** The output of the model is a scalar.

Let:

- The kernel be:

- The input be:

- The bias is a scalar .

The model output is given by:

.

Since the loss computes the difference between an y, and is a scalar while y is a 32x32 matrix, broadcasting is applied. Broadcasting repeats the scalar ​ over all cells of the matrix y during the subtraction.

The loss term is therefore:

Using the chain rule :

A close-up of a text

Description automatically generated

(a + b)

Yes, it is possible to add positional embeddings when using a Recurrent Neural Network (RNN) model.

The idea behind positional encoding is to add a vector to each token’s embedding that represent it’s position in the sequence.

In the context of RNNs, you can augment the input embeddings with positional embeddings before feeding them into the RNN layers, this means that for each time step t instead of feeding only the input ​, you feed ​, where is the positional embedding corresponding to position t.

However, while it is possible to add positional embeddings to an RNN, it is generally not necessary because RNNs naturally incorporate positional information due to their sequential processing of tokens.

RNNs read the input sequence one token at a time, and the hidden state at each time step inherently depends on all the previous tokens and their positions in the sequence. This sequential processing ensures that the model "knows" the position of each token based on when it is processed, making explicit positional embeddings redundant. However, if desired, positional embeddings could still be added to the input before feeding it into the RNN, though this would likely offer limited benefit.

A black and white image of a square with white and gray squares

Description automatically generatedNote: The Transformer processes input sequences in parallel, meaning that it has no inherent notion of the order of tokens. To give the model a sense of order, positional encoding is added to the token embeddings.

A black text on a white background

Description automatically generated

7

(a)

Each pixel in the map represents the attention weight between a word in the source sequence (English) and a word in the target sequence (French). The attention weight shows how much influence the English token has when creating the context-based representation for the French token by applying the attention weights. The attention weight is based on the similarity between the query (target token) and the keys (source tokens). A higher weight means the English token played a more significant role in forming the representation of the French token. The color of the pixel represents the strength of attention: lighter colors indicate stronger relationships (higher attention weights), while darker colors indicate weaker relationships.

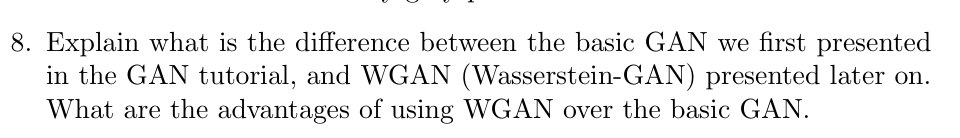
(b)   
Rows with only one non-zero pixel mean that the attention for that particular French word is focused entirely on a single English word. This typically occurs when the model identifies a direct, straightforward translation between the French and English tokens. In this case, the model does not need to distribute attention across multiple words and has clearly identified that the English token is the most relevant for translating the French token. As shown in the image, these single-focus attention rows often correspond to one-to-one word translations between the two languages.

(c)

Rows with several non-zero pixels mean that the attention for that particular French word is distributed across multiple English tokens. This indicates that the translation of the word is not a straightforward one-to-one relationship, or that more context is needed to understand the French token. For example, in the case of "L'", the attention is distributed between the matching word in English ("the") and the objects it refers to, such as "agreement" and "signed". This shows that the model is using information from multiple English words to form the correct French translation.

(d)  
 The attention weights for each French token must sum to 1 (the weights are normalized during calculation using SoftMax). When a row has only one non-zero pixel, that single token receives all the attention (with a weight of 1), which is why the pixel appears white, indicating maximum attention. In contrast, when the attention is distributed across multiple tokens in the source sequence, each token is assigned a fraction of the total attention, with individual weights less than 1. Since the attention is divided, these pixels will have lower values and therefore appear as darker pixels (gray) rather than white.

(I add this “ as darker pixels”)



A white paper with black text

Description automatically generated

9.

We saw that , meaning the log-likelihood of is constructed from a KL-divergence term and an ELBO term. The KL divergence is always non-negative. Therefore, we can conclude that . This means the ELBO serves as a lower bound for .

Since our goal is to maximize (in line with the maximum likelihood estimation approach), if we maximize the ELBO term, we also maximize what we want to optimize: the likelihood.

1. To compute the KL-divergence between and , we can either estimate it empirically (by sampling from the distributions) or use an analytical solution if the distributions are explicitly defined. In this case, we don’t have an exact form for these distributions.

While we could attempt to approximateusing Markov rules, this would require estimating intermediate values , which is generally avoided during training. Instead, we typically compute directly , as it's more practical.

On the other hand, is estimated through a neural network. While we could use this neural network to estimate the full distribution ,it becomes challenging because we would need additional unknown values, such as the joint distribution or marginal probabilities like

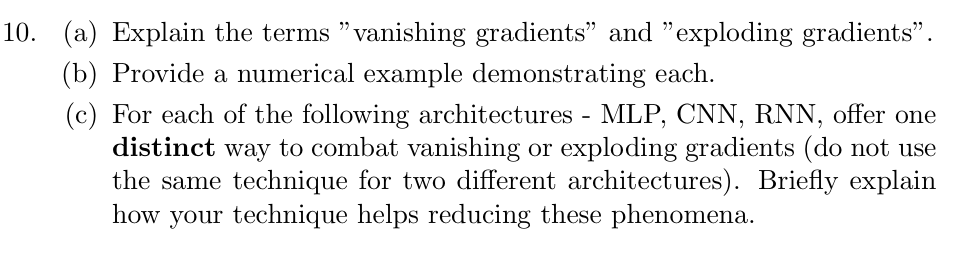
Moreover, even if we had these distributions**, computing the expectation required for the KL divergence itself can be difficult.** The expectation involves integrating over all possible sequences, which can be computationally expensive or intractable, particularly when working with high-dimensional distributions

**Another answer:**  models the probability of a sequence given the final image This is hard to compute since models the reverse process .

Even if we use formulas like we will still stuck with problematic terms we don’t have like (which is essentially what we try to compute).

1. We ignore this term because it is constant with respect to the learned parameters

Since is not learned and is a known prior, this KL-divergence does not depend on . Therefore, its gradient is zero, meaning it does not affect the optimization process, allowing us to safely ignore it during training.

****