1 Modern achitecture

1. A CNN takes a 128×128 RGB image (3 channels) as input. It is passed through the following sequence of layers:

   (a) Conv layer with 16 filters, each of size 5×5, stride=1, padding='same'

   (b) Max Pooling layer with 2×2 window, stride=2

   (c) Conv layer with 32 filters, each of size 3×3, stride=1, padding='valid'

   (d) Max Pooling layer with 2×2 window, stride=2

   (e) A Flatten layer

   (f) A final Dense (Fully Connected) layer with 100 units

   - Calculate the dimensions (height, width, number of channels) of the output after each of the first four layers. Show your work.
   - What is the total number of elements after the Flatten layer?
   - Calculate the number of trainable parameters in the first convolutional layer. (Remember: parameters include weights and biases).

2. Build your own residual net and train on the CIFAR10 dataset, try your best (within your computing capability) to get the test accuracy to 90 percent. Submit your code and training log.

3. Consider a simple RNN cell defined by the following equations:

$$\begin{aligned} \text{Hidden State Update:} \quad & h_t = \tanh(W_{xh} x_t + W_{hh} h_{t-1} + b_h) \\ \text{Output:} \quad & y_t = W_{hy} h_t + b_y \end{aligned}$$

   Where:

   - x_t is the input at time step t
   - h_t is the hidden state at time step t (h_0 is initialized to zeros)
   - y_t is the output at time step t

   Given an input sequence $\mathbf{x} = [x_1, x_2, x_3]$:

   (a) Draw the computational graph by **unrolling the RNN** for all three time steps. Clearly show the inputs x_t, hidden states h_t, outputs y_t, and the sharing of weights W_{xh}, W_{hh}, and W_{hy} across time.

   (b) Using your graph, explain how the hidden state h_3 depends on the entire input sequence $[x_1, x_2, x_3]$. Why is the hidden state considered the "memory" of the network?

4. The RNN network was described in last problem.

   (a) During Backpropagation Through Time (BPTT), the gradient of the loss L with respect to an early hidden state h_1 involves a long chain of derivatives. Write the expression for $\frac{\partial L}{\partial h_1}$ in terms of $\frac{\partial L}{\partial h_3}$, showing the chain rule through h_2.

   (b) The tanh derivative is $\frac{d}{dz}\tanh(z) = 1 - \tanh^2(z)$, which is always ≤ 1. Explain how repeated multiplication of these derivatives (and the weight matrix W_{hh}) during BPTT can cause the **vanishing gradient problem**.

   (c) What is the negative consequence of vanishing gradients for an RNN's ability to learn long-range dependencies in a sequence?

5. Using the attached dataset (input.txt), and using the RNN with the GRU cell to train a language model. Submit your model and the sampling results.

6. Given the fundamental attention equation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Consider the following input matrices:

$$Q = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}, \quad K = \begin{bmatrix} 2 & 1 \\ 4 & 3 \end{bmatrix}, \quad V = \begin{bmatrix} 0.5 & 1.0 \\ 1.5 & 2.0 \end{bmatrix}$$

where $Q \in \mathbb{R}^{3\times2}$, $K \in \mathbb{R}^{2\times2}$, $V \in \mathbb{R}^{2\times2}$, and $d_k = 2$.

(a) Compute the raw attention scores QK^T and show all steps.

(b) Apply the scaling factor $\frac{1}{\sqrt{d_k}}$ to the raw scores.

(c) Calculate the attention weights by applying the softmax function to each row of the scaled scores. Show your work for at least one row.

(d) Compute the final output by multiplying the attention weights with V.

(e) What is the shape of the final output matrix? Explain why this shape makes sense given the inputs.

7. The multi-head attention mechanism is defined as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where each head is computed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Given:

- $Q, K, V \in \mathbb{R}^{4\times8}$ (sequence length 4, embedding dimension 8)
- Number of heads $h = 2$
- $d_k = d_v = d_{\text{model}}/h = 4$

(a) What must be the dimensions of the projection matrices W_i^Q, W_i^K, W_i^V, and W^O?

(b) If the output of the concatenated heads is $C \in \mathbb{R}^{4\times8}$, what operations are needed to ensure the final output has the same dimensions as the input?

(c) Write the complete mathematical expression for one head of the multi-head attention, showing all matrix dimensions.

8. Consider the gradient flow through the attention mechanism. Let L be the loss function.

(a) Using the chain rule, express $\frac{\partial L}{\partial Q}$ in terms of:

- $\frac{\partial L}{\partial \text{Output}}$
- The attention weights matrix $A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$
- The value matrix V
- The key matrix K

(b) The softmax derivative for a vector $\mathbf{z}$ is:

$$\frac{\partial \text{softmax}(z_i)}{\partial z_j} = \text{softmax}(z_i)(\delta_{ij} - \text{softmax}(z_j))$$

where δ_{ij} is the Kronecker delta. How does this affect gradient computation in attention?

(c) Explain why the scaling factor $\frac{1}{\sqrt{d_k}}$ helps with gradient stability during training.