Equations for Quiz 2

1. **Noisy channel**
   Let x be the target and a the observed noisy signal.

   (a) Bayes rule:
   $$P(x \mid a) \propto P(a \mid x)\, P(x)$$

   (b) Decoding (MAP estimate):
   $$\hat{x} = \arg\max_x P(a \mid x)\, P(x)$$

2. Let x_i be the i-th token in the sequence. We can decompose the probability of the sequence as:
   $$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid x_1, \ldots, x_{i-1})$$

3. **Markov assumption** We can approximate the probability of generating x_i given prior words by assuming x_i only depends on the previous k tokens:
   $$P(x_i \mid x_1, \ldots, x_{i-1}) \approx P(x_i \mid x_{i-k}, \ldots, x_{i-1})$$

4. **RNN** Let h_i be the hidden state at time i, x_i the input embedding, W, U weight matrices, b a bias vector, and g an activation function (e.g., tanh). RNN hidden state update:
   $$h_i = R(h_{i-1}, x_i) = g(Wh_{i-1} + Ux_i + b)$$

5. Let h_{i-1} be the hidden state summarizing x_1 to x_{i-1}. Let V be the output weight matrix and b_o the output bias. RNN output prediction:
   $$O(h_i) = \text{softmax}(h_i V + b_o)$$

6. RNN-based language model:
   $$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid h_{i-1})$$

7. y_i^* is the gold (target) output token at time i. Seq2seq teacher forcing objective:
   $$\sum_{(x,y)} \sum_{i=1}^{n} \log P(y_i^* \mid x, y_1^*, \ldots, y_{i-1}^*)$$

8. Seq2seq (no teacher forcing, $\hat{y}_i$ is model prediction.
   $$\sum_{(x,y)} \sum_{i=1}^{n} \log P(y_i^* \mid x, \hat{y}_1, \ldots, \hat{y}_{i-1})$$

9. Attention Mechanism (Encoder-Decoder)
   Let $\bar{h}_i$ be the decoder state at step i, and h_j the encoder state at position j. W is a learned matrix. Let e_{ij} be unnormalized alignment scores.

(a) Additive attention score (Bahdanau):

$$e_{ij} = f(\bar{h}_i, h_j) = \tanh(W[\bar{h}_i; h_j])$$

(b) Dot-product attention score (Luong dot):

$$e_{ij} = \bar{h}_i \cdot h_j$$

(c) Bilinear attention score (Luong bilinear):

$$e_{ij} = \bar{h}_i^\top W h_j$$

(d) Attention weights:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j'} \exp(e_{ij'})}$$

10. Self-Attention and Transformers

Let $X \in \mathbb{R}^{n \times d_{\text{model}}}$ be the input sequence (n = number of tokens, d_{model} = hidden size).

Q, K, V are query, key, and value matrices obtained by linear projections of X. W^Q, W^K, W^V are learned projection matrices, and W^O is the output projection.

- Projection dimensions:

$$W^Q, W^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, \quad W^V \in \mathbb{R}^{d_{\text{model}} \times d_v}, \quad W^O \in \mathbb{R}^{h \cdot d_v \times d_{\text{model}}}.$$

W_i are feed-forward weight matrices, b_i bias terms, and Sublayer(x) denotes any sublayer (e.g., attention or FFN). h is number of attention heads.

Note: You might find illustrated transformer helpful.

(a) Scaled Dot-Product Attention

i. We obtain queries, keys, and values by linear projections:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V$$

ii. Raw scores:

$$S = \frac{QK^\top}{\sqrt{d_k}}$$

iii. Attention weights:

$$A = \text{softmax}(S)$$

iv. Output:

$$\text{Attn}(Q, K, V) = AV$$

(b) Multi-Head Attention

i. For each single head:

$$\text{head}_\ell = \text{Attn}(Q_l, K_l, V_l),$$

where

$$Q_\ell = XW_\ell^Q, \quad K_\ell = XW_\ell^K, \quad V_\ell = XW_\ell^V$$

ii. Concatenation + output projection:

$$\text{MultiHead}(X) = [\text{head}_1; \ldots; \text{head}_h]W^O$$

(c) Position-wise feed-forward network:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2$$

(d) Residual connection and layer normalization

$$\text{LayerNorm}(x + \text{Sublayer}(x))$$

11. Attention mask for Transformer decoder.
Let $q_i \in \mathbb{R}^{d_k}$ be the query vector at position i, and $k_j \in \mathbb{R}^{d_k}$ the key vector at position j. The unnormalized attention score e_{ij} is defined as:

$$e_{ij} = \begin{cases} q_i^\top k_j, & j < i \\ -\infty, & j \geq i \end{cases}$$