FIT3181/5215 Deep Learning

Advanced Sequential Models

Teaching team

Department of Data Science and AI
Faculty of Information Technology, Monash University
Email: **trunglm@monash.edu**

Question 1

- For an encoder-decoder model, which statements are correct?

A. Encoder tries to read from context vector to generate an output sequence.
B. Decoder tries to read from context vector to generate an output sequence.
C. Encoder tries to encode an input sequence to a context vector.
D. Decoder tries to encode an input sequence to a context vector.
E. Context vector summarizes an input sequence.
F. Context vector summarizes a target sequence.

Question 1

□ For an encoder-decoder model, which statements are correct?

A. Encoder tries to read from context vector to generate an output sequence.
B. Decoder tries to read from context vector to generate an output sequence. ✓
C. Encoder tries to encode an input sequence to a context vector. ✓
D. Decoder tries to encode an input sequence to a context vector.
E. Context vector summarizes an input sequence. ✓
F. Context vector summarizes a target sequence.

Question 2

□ In seq2seq for machine translation, which statements are correct?

A. Encoder is a feed-forward neural network and decoder is a feed-forward neural network.
B. Encoder is a convolutional neural network and decoder is a convolutional neural network.
C. Encoder is a recurrent neural network and decoder is a recurrent neural network.
D. Context vector could be the last hidden state of the decoder.
E. Context vector could be the last hidden state of encoder.
F. Context vector could be the first hidden state of encoder.

Question 2

□ In seq2seq for machine translation, which statements are correct?

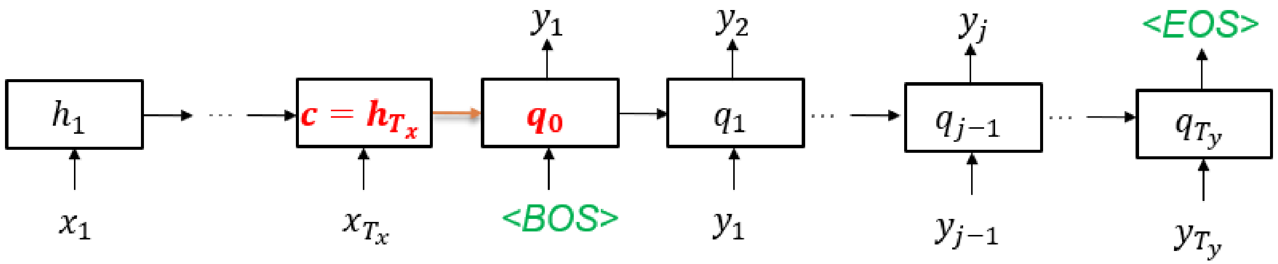
A. Encoder is a feed-forward neural network and decoder is a feed-forward neural network.
B. Encoder is a convolutional neural network and decoder is a convolutional neural network.
C. Encoder is a recurrent neural network and decoder is a recurrent neural network. ✓
D. Context vector could be the last hidden state of the decoder.
E. Context vector could be the last hidden state of encoder. ✓
F. Context vector could be the first hidden state of encoder.

Question 3

- In seq2seq for machine translation in the following figure, we derive the log-likelihood as follows:

$$P(\boldsymbol{y}|\boldsymbol{x},\theta) = P\left(\boldsymbol{y}_{1:T_y} \mid \boldsymbol{x}_{1:T_x},\theta\right) = P\left(\boldsymbol{y}_{1:T_y} \mid \boldsymbol{c},\theta\right) \overset{(1)}{\Rightarrow} P(\boldsymbol{y}_1 \mid \boldsymbol{c},\theta)P(\boldsymbol{y}_2 \mid \boldsymbol{y}_1,\boldsymbol{c},\theta)\dots P(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1},\boldsymbol{c},\theta)\dots P\left(\boldsymbol{y}_{T_y} \mid \boldsymbol{y}_{1:T_y-1},\boldsymbol{c},\theta\right)$$

$$= \prod_{j=1}^{T_y} P(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1},\boldsymbol{c},\theta) \overset{(2)}{\Rightarrow} \prod_{j=1}^{T_y} P(\boldsymbol{y}_j \mid \boldsymbol{q}_{j-1},\boldsymbol{c},\theta$$



Which statements are correct?

A. In the derivation (1), $\boldsymbol{c}$ is viewed as a summary of the sequence $\boldsymbol{x}_{1:T_x}$.

B. In the derivation (1), $\boldsymbol{c}$ is viewed as a summary of the sequence $\boldsymbol{y}_{1:T_y}$.

C. In the derivation (2), $\boldsymbol{q}_{j-1}$ is viewed as a summary of the sequence $\boldsymbol{y}_{1:j-1}$.

D. In the derivation (2), $\boldsymbol{q}_{j-1}$ is viewed as a summary of the sequence $\boldsymbol{y}_{1:T_y}$.

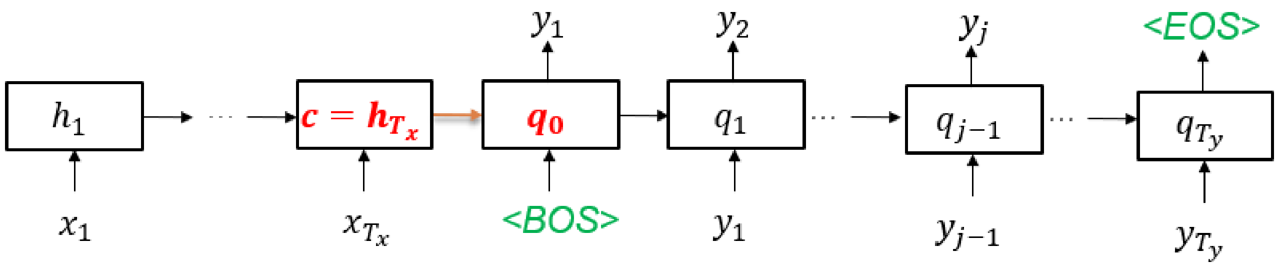
E. $P\left(\boldsymbol{y}_j \mid \boldsymbol{q}_{j-1},\boldsymbol{c},\theta\right)$ means that on top of $\boldsymbol{q}_{j-1}$, $\boldsymbol{c}$, we can build up some dense layers to predict $\boldsymbol{y}_j$.

Question 3

- In seq2seq for machine translation in the following figure, we derive the log-likelihood as follows:

$$P(\boldsymbol{y}|\boldsymbol{x},\theta) = P\left(\boldsymbol{y}_{1:T_y} \mid \boldsymbol{x}_{1:T_x},\theta\right) = P\left(\boldsymbol{y}_{1:T_y} \mid \boldsymbol{c},\theta\right) \overset{(1)}{\Rightarrow} P(\boldsymbol{y}_1 \mid \boldsymbol{c},\theta)P(\boldsymbol{y}_2 \mid \boldsymbol{y}_1,\boldsymbol{c},\theta)\dots P(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1},\boldsymbol{c},\theta)\dots P\left(\boldsymbol{y}_{T_y} \mid \boldsymbol{y}_{1:T_y-1},\boldsymbol{c},\theta\right)$$

$$= \prod_{j=1}^{T_y} P(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1},\boldsymbol{c},\theta) \overset{(2)}{\Rightarrow} \prod_{j=1}^{T_y} P(\boldsymbol{y}_j \mid \boldsymbol{q}_{j-1},\boldsymbol{c},\theta)$$



Which statements are correct?

A. In the derivation (1), c is viewed as a summary of the sequence $\boldsymbol{x}_{1:T_x}$. ✓

B. In the derivation (1), c is viewed as a summary of the sequence $\boldsymbol{y}_{1:T_y}$.

C. In the derivation (2), $\boldsymbol{q}_{j-1}$ is viewed as a summary of the sequence $\boldsymbol{y}_{1:j-1}$. ✓

D. In the derivation (2), $\boldsymbol{q}_{j-1}$ is viewed as a summary of the sequence $\boldsymbol{y}_{1:T_y}$.

E. $P(\boldsymbol{y}_j \mid \boldsymbol{q}_{j-1},\boldsymbol{c},\theta)$ means that on top of $\boldsymbol{q}_{j-1}$, $\boldsymbol{c}$, we can build up some dense layers to predict $\boldsymbol{y}_j$. ✓

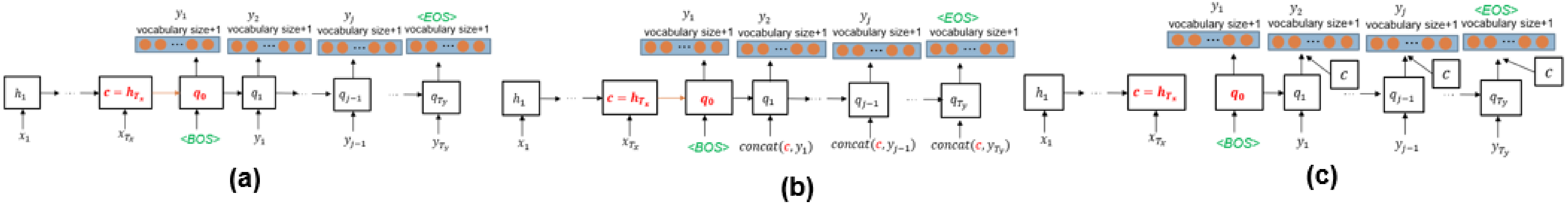
Question 4

- In seq2seq for machine translation, we derive as follows:

$$P(\boldsymbol{y}|\boldsymbol{x},\theta) = P\left(\boldsymbol{y}_{1:T_y} \mid \boldsymbol{x}_{1:T_x},\theta\right) = P\left(\boldsymbol{y}_{1:T_y} \mid \boldsymbol{c},\theta\right) \overset{(1)}{\Rightarrow} P(\boldsymbol{y}_1 \mid \boldsymbol{c},\theta)P(\boldsymbol{y}_2 \mid \boldsymbol{y}_1,\boldsymbol{c},\theta)\ldots P(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1},\boldsymbol{c},\theta)\ldots P\left(\boldsymbol{y}_{T_y} \mid \boldsymbol{y}_{1:T_y-1},\boldsymbol{c},\theta\right)$$

$$= \prod_{j=1}^{T_y} P(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1},\boldsymbol{c},\theta) \overset{(2)}{\Rightarrow} \prod_{j=1}^{T_y} P(\boldsymbol{y}_j \mid \boldsymbol{q}_{j-1},\boldsymbol{c},\theta)$$

We need to formulate $P(\boldsymbol{y}_j \mid \boldsymbol{q}_{j-1},\boldsymbol{c},\theta)$. Consider the diagrams (a), (b), (c). Which statements are correct?



A. Diagram (a) can be used to formulate the above conditional distribution.

B. Diagram (b) can be used to formulate the above conditional distribution.

C. Diagram (c) can be used to formulate the above conditional distribution.

D. None of (a), (b), (c) can be used to formulate the above conditional distribution.

E. Only (a) and (b) can be used to formulate the above conditional distribution.

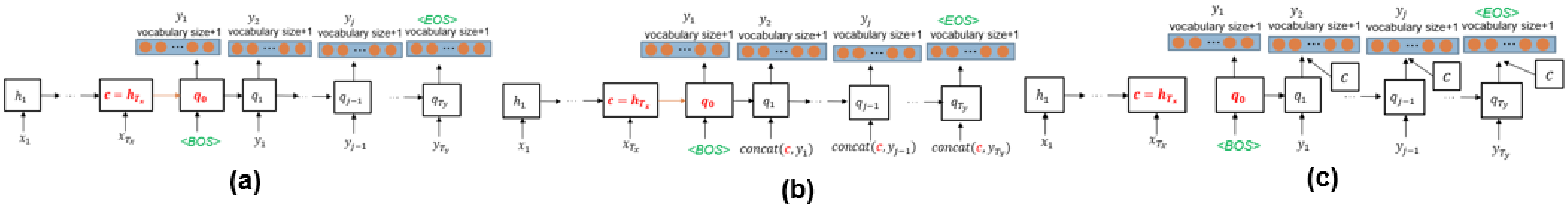
Question 4

- In seq2seq for machine translation, we derive as follows:

$$P(y|x,\theta) = P\left(y_{1:T_y} \mid x_{1:T_x},\theta\right) = P\left(y_{1:T_y} \mid c,\theta\right) \overset{(1)}{\Rightarrow} P(y_1 \mid c,\theta)P(y_2 \mid y_1, c,\theta)\ldots P(y_j \mid y_{1:j-1}, c,\theta)\ldots P\left(y_{T_y} \mid y_{1:T_y-1}, c,\theta\right)$$

$$= \prod_{j=1}^{T_y} P(y_j \mid y_{1:j-1}, c,\theta) \overset{(2)}{\Rightarrow} \prod_{j=1}^{T_y} P(y_j \mid q_{j-1}, c,\theta)$$

We need to formulate $P(y_j \mid q_{j-1}, c, \theta)$. Consider the diagrams (a), (b), (c). Which statements are correct?



A. Diagram (a) can be used to formulate the above conditional distribution. ✓

B. Diagram (b) can be used to formulate the above conditional distribution. ✓

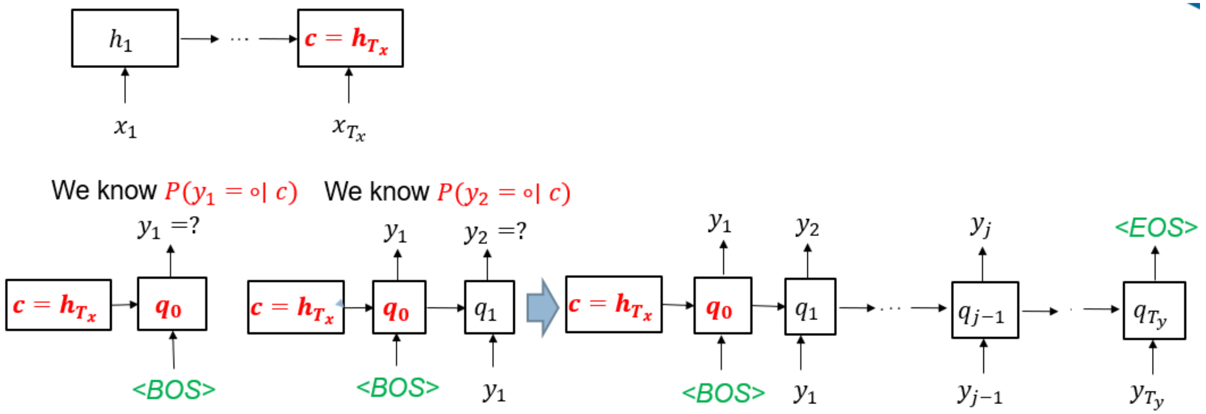
C. Diagram (c) can be used to formulate the above conditional distribution. ✓

D. None of (a), (b), (c) can be used to formulate the above conditional distribution.

E. Only (a) and (b) can be used to formulate the above conditional distribution.

Question 5

- In the decoding process of seq2seq for machine translation as in the following figure, which statements are correct?



A. In the phase 1, we feed the input sequence to the encoder to evaluate the context c as the last hidden state of the encoder.

B. In the phase 2, we feed EOS symbol the decoder and decode output sequence from this symbol.

C. In the phase 2, we feed BOS symbol the decoder and decode output sequence from this symbol.

D. In the phase 2, we initialize the first hidden state of decoder with the last item in the input sequence.

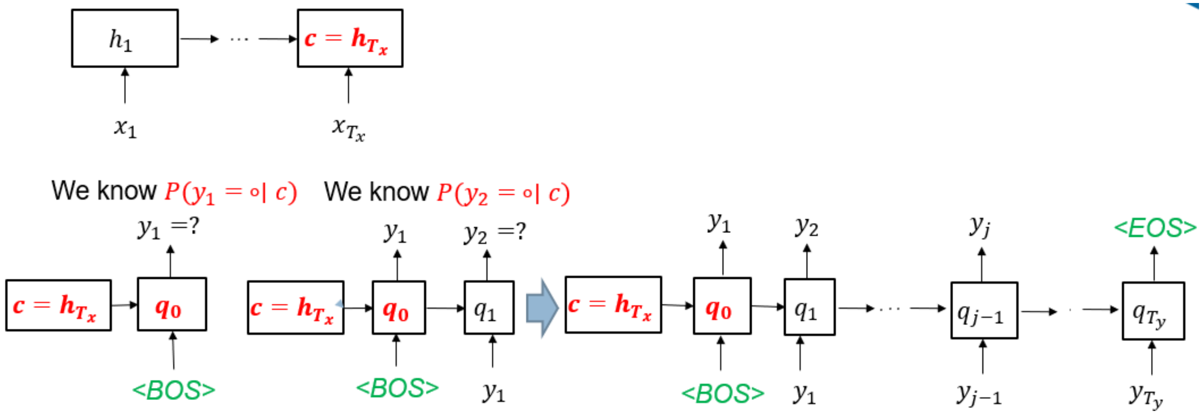
E. In the phase 2, we initialize the first hidden state of decoder with the last hidden state of the encoder.

F. In the phase 2, if we use the greedy strategy, at each timestep, we sample the next output item from the conditional distribution.

G. In the phase 2, if we use the greedy strategy, at each timestep, we choose the next output item that maximizes the conditional distribution.

Question 5

- In the decoding process of seq2seq for machine translation as in the following figure, which statements are correct?



A. In the phase 1, we feed the input sequence to the encoder to evaluate the context c as the last hidden state of the encoder. ✓

B. In the phase 2, we feed EOS symbol the decoder and decode output sequence from this symbol.

C. In the phase 2, we feed BOS symbol the decoder and decode output sequence from this symbol. ✓

D. In the phase 2, we initialize the first hidden state of decoder with the last item in the input sequence.

E. In the phase 2, we initialize the first hidden state of decoder with the last hidden state of the encoder. ✓

F. In the phase 2, if we use the greedy strategy, at each timestep, we sample the next output item from the conditional distribution.

G. In the phase 2, if we use the greedy strategy, at each timestep, we choose the next output item that maximizes the conditional distribution. ✓

Question 6

- What are the advantages of timely varied context comparing with fixed-length context?

A. Fixed-length context is possibly less powerful to capture long input sequences, while timely varied context can provide dynamic and timely adapted context for input sequences.
B. Fixed-length context is simpler and more compact than timely varied context.
C. Fixed-length context can summarize the input sequence, while timely varied context cannot.
D. Fixed-length context can summarize the input sequence more accurately than timely varied context can.
E. Timely varied context can focus on some input items or words that are more important to generate specific output items or words, while fixed-length context cannot.
F. Fixed-length context can focus on some input items or words that are more important to generate specific output items or words, while timely varied context cannot.

Question 6

□ What are the advantages of timely varied context comparing with fixed-length context?

A. Fixed-length context is possibly less powerful to capture long input sequences, while timely varied context can provide dynamic and timely adapted context for input sequences. ✓
B. Fixed-length context is simpler and more compact than timely varied context.
C. Fixed-length context can summarize the input sequence, while timely varied context cannot.
D. Fixed-length context can summarize the input sequence more accurately than timely varied context can.
E. Timely varied context can focus on some input items or words that are more important to generate specific output items or words, while fixed-length context cannot. ✓
F. Fixed-length context can focus on some input items or words that are more important to generate specific output items or words, while timely varied context cannot.

Question 7

☐ What are correct for the global attention?

A. In the global attention, the time varied context is computed based on encoder hidden states in a selective window.
B. In the global attention, the time varied context is computed based on all decoder hidden states.
C. In the global attention, the time varied context is computed based on decoder hidden states in a selective window.
D. In the global attention, the time varied context is computed based on all encoder hidden states.
E. In the global attention, the time varied context is a linear combination of all decoder hidden states.
F. In the global attention, the time varied context is a linear combination of all decoder hidden states.

Question 7

- What are correct for the global attention?

A. In the global attention, the time varied context is computed based on encoder hidden states in a selective window.

B. In the global attention, the time varied context is computed based on all decoder hidden states.

C. In the global attention, the time varied context is computed based on decoder hidden states in a selective window.

D. In the global attention, the time varied context is computed based on all encoder hidden states. ✓

E. In the global attention, the time varied context is a linear combination of all decoder hidden states.

F. In the global attention, the time varied context is a linear combination of all encoder hidden states. ✓

Question 8

- What are correct for the local attention?

A. In the local attention, the time varied context is computed based on all encoder hidden states in a selective window.
B. In the local attention, the time varied context is computed based on all decoder hidden states.
C. In the local attention, the time varied context is computed based on all decoder hidden states in a selective window.
D. In the local attention, the time varied context is computed based on all encoder hidden states.
E. In the local attention, the time varied context is a linear combination of all encoder hidden states in a selective window.
F. In the local attention, the time varied context is a linear combination of all encoder hidden states.

Question 8

- What are correct for the local attention?

A. In the local attention, the time varied context is computed based on all encoder hidden states in a selective window. ✓

B. In the local attention, the time varied context is computed based on all decoder hidden states.

C. In the local attention, the time varied context is computed based on all decoder hidden states in a selective window.

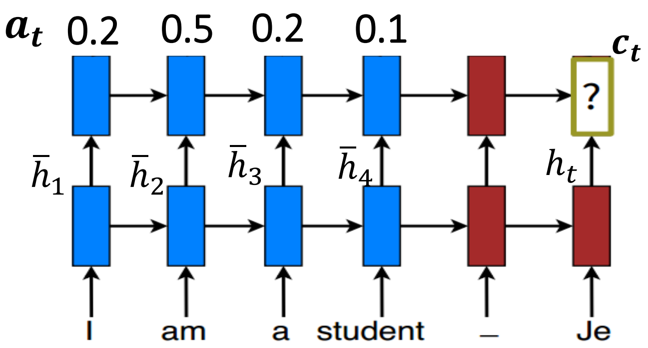
D. In the local attention, the time varied context is computed based on all encoder hidden states.

E. In the local attention, the time varied context is a linear combination of all encoder hidden states in a selective window. ✓

F. In the local attention, the time varied context is a linear combination of all encoder hidden states.

Question 9

- Consider the below seq2seq model. We apply the global attention to compute the context vector c_t. What are correct?



A. The second word is more important to the generation of the current output word.

B. The fourth word is more important to the generation of the current output word.

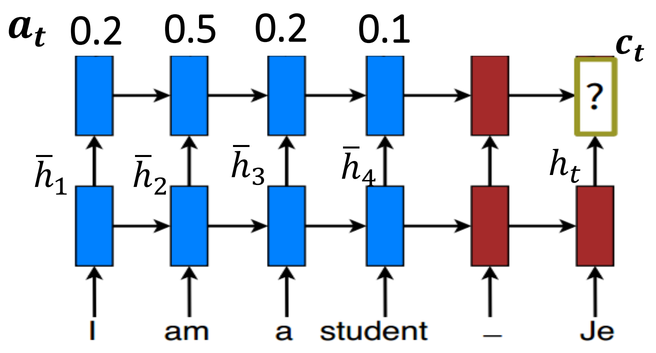
C. $c_t = 0.2\bar{h}_1 + 0.5\bar{h}_2 + 0.2\bar{h}_3 + 0.1\bar{h}_4$

D. $c_t = h_t$.

E. $c_t = 0.1\bar{h}_1 + 0.2\bar{h}_2 + 0.5\bar{h}_3 + 0.2\bar{h}_4$

Question 9

- Consider the below seq2seq model. We apply the global attention to compute the context vector c_t. What are correct?



A. The second word is more important to the generation of the current output word. ✓

B. The fourth word is more important to the generation of the current output word.

C. $c_t = 0.2\bar{h}_1 + 0.5\bar{h}_2 + 0.2\bar{h}_3 + 0.1\bar{h}_4$ ✓

D. $c_t = h_t$.

E. $c_t = 0.1\bar{h}_1 + 0.2\bar{h}_2 + 0.5\bar{h}_3 + 0.2\bar{h}_4$

Question 10

- In Transformers, what are correct about the Positional Encoding?

A. It helps capture the position of a sentence in a mini-batch.
B. It helps capture the position of a word/token in a sentence.
C. It produces the embeddings for words/tokens in a sentence.
D. It is added to the embeddings of words/tokens in a sentence.
E. It is used as the main signal to input to transformers.

Question 10

- In Transformers, what are correct about the Positional Encoding?

A. It helps capture the position of a sentence in a mini-batch.
B. It helps capture the position of a word/token in a sentence. ✓
C. It produces the embeddings for words/tokens in a sentence.
D. It is added to the embeddings of words/tokens in a sentence. ✓
E. It is used as the main signal to input to transformers.

Question 11

☐ In Transformers, what are correct about the Layer Norm?

A. It normalizes the input tensor across the batch size dimension.
B. It normalizes the input tensor across the embedding size dimension (i.e., the dimension of d_model).
C. It has no parameters.
D. It has the scaling and shifting parameters γ and β.
E. It is more effective than Batch Norm for sequential data.
F. It is less effective than Batch Norm for sequential data.

Question 11

☐ In Transformers, what are correct about the Layer Norm?

A. It normalizes the input tensor across the batch size dimension.
B. It normalizes the input tensor across the embedding size dimension (i.e., the dimension of d_model). ✓
C. It has no parameters.
D. It has the scaling and shifting parameters γ and β. ✓
E. It is more effective than Batch Norm for sequential data. ✓
F. It is less effective than Batch Norm for sequential data.

Question 12

- Assume that we have a sequence of token embeddings $x_1, \dots, x_L$ (L is the sequence length) is inputted to a Self-Attention layer to obtain another sequence of token embedding $z_1, \dots, z_L$. What are correct?

A. The token embedding z_i is only dependent on its previous token embedding x_i.

B. The token embedding z_i is mainly dependent on its previous token embedding x_i, but other x_j $(j \neq i)$ also contributes to the computation of z_i.

C. More similar x_j is to x_i, more contribution it is to the the computation of z_i.

D. More similar x_j is to x_i, less contribution it is to the the computation of z_i.

Question 12

- Assume that we have a sequence of token embeddings $x_1, \dots, x_L$ (L is the sequence length) is inputted to a Self-Attention layer to obtain another sequence of token embedding $z_1, \dots, z_L$. What are correct?

A. The token embedding z_i is only dependent on its previous token embedding x_i.

B. The token embedding z_i is mainly dependent on its previous token embedding x_i, but other x_j ($j \neq i$) also contributes to the computation of z_i. ✓

C. More similar x_j is to x_i, more contribution it is to the the computation of z_i. ✓

D. More similar x_j is to x_i, less contribution it is to the the computation of z_i.

Question 13

- Assume that we input to a Self-Attention layer a matrix $X = \begin{bmatrix} x_1 \\ \cdots \\ x_L \end{bmatrix}$ $(L = seq\_len)$ that contains the token/word embeddings of a sentence. What are correct about the Self-Attention layer?

A. We use three weight matrices W_Q, W_K, W_V to compute Q, K, V respectively.

B. We rely on Q, V to compute the attention scores to store in a matrix B that has shape [L,L].

C. We rely on Q, K to compute the attention scores to store in a matrix B that has shape [L,L].

D. Q, K can be considered as two other views of X.

E. We apply the softmax function to the attention scores B to gain the attention probabilities A that has shape [L,L].

F. We multiply B and V to obtain the new token/word embeddings $Z = BV$.

G. We multiply A and V to obtain the new token/word embeddings $Z = AV$.

Question 13

- Assume that we input to a Self-Attention layer a matrix $X = \begin{bmatrix} x_1 \\ \cdots \\ x_L \end{bmatrix}$ $(L = seq\_len)$ that contains the token/word embeddings of a sentence. What are correct about the Self-Attention layer?

A. We use three weight matrices W_Q, W_K, W_V to compute Q, K, V respectively. ✓

B. We rely on Q, V to compute the attention scores to store in a matrix B that has shape [L,L].

C. We rely on Q, K to compute the attention scores to store in a matrix B that has shape [L,L]. ✓

D. Q, K can be considered as two other views of X. ✓

E. We apply the softmax function to the attention scores B to gain the attention probabilities A that has shape [L,L]. ✓

F. We multiply attention scores B and V to obtain the new token/word embeddings $Z = BV$.

G. We multiply attention probs A and V to obtain the new token/word embeddings $Z = AV$. ✓

Question 14

- What are correct about the multi-head Self-Attention?

A. Each head has its own W_Q, W_K, W_V.

B. The weight matrices W_Q, W_K, W_V are shared across the heads.

C. We perform each head independently.

D. The outputs of the heads are conditionally dependent.

E. We concatenate the outputs of each head and use this concatenation as the output of the multi-head Self-Attention.

F. We concatenate the outputs of each head and input this concatenation to one more linear layer W_o to gain the output of multi-head Self-Attention.

Question 14

- What are correct about the multi-head Self-Attention?

A. Each head has its own W_Q, W_K, W_V. ✓

B. The weight matrices W_Q, W_K, W_V are shared across the heads.

C. We perform each head independently. ✓

D. The outputs of the heads are conditionally dependent.

E. We concatenate the outputs of each head and use this concatenation as the output of the multi-head Self-Attention.

F. We concatenate the outputs of each head and input this concatenation to one more linear layer W_o to gain the output of multi-head Self-Attention. ✓

Question 15

- What are correct about the Cross-Attention?

A. We use the Cross-Attention to inject the encoder output to the decoder layers.
B. The Cross-Attention computation only depends on the current decoder input.
C. For the Cross-Attention, the decoder input is used to compute Q, whereas the encoder output is used to compute K, V.
D. For the Cross-Attention, the decoder input is used to compute K, V, whereas the encoder output is used to compute Q.
E. The Cross-Attention is involved in the computation of encoder output.
F. The Cross-Attention is involved in the computation of decoder output.

Question 15

- What are correct about the Cross-Attention?

A. We use the Cross-Attention to inject the encoder output to the decoder layers. ✓

B. The Cross-Attention computation only depends on the current decoder input.

C. For the Cross-Attention, the decoder input is used to compute Q, whereas the encoder output is used to compute K, V. ✓

D. For the Cross-Attention, the decoder input is used to compute K, V, whereas the encoder output is used to compute Q.

E. The Cross-Attention is involved in the computation of encoder output.

F. The Cross-Attention is involved in the computation of decoder output. ✓

Thanks for your attention!