FIT3181/5215 Deep Learning

Vision Transformers and Model Fine-Tuning Techniques

Teaching team

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

Question 1

- What are correct about the self-attention?

A. With self-attention, we can compute the token embeddings in parallel.

B. With self-attention, we need to compute the token embeddings sequentially.

C. For self-attention, the queries (Q), keys (K), and values (V) computed based on source sequences

D. For self-attention, the queries (Q) are computed based on target sequences, while keys (K), and values (V) are computed based on source sequences

E. Self-attention can capture **intra-**sequence dependencies between source sequences

F. Self-attention can capture inter-sequence dependencies between source sequences and target sequences

Question 1

- What are correct about the self-attention?

A. With self-attention, we can compute the token embeddings in parallel. **[x]**

B. With self-attention, we need to compute the token embeddings sequentially.

C. For self-attention, the queries (Q), keys (K), and values (V) computed based on source sequences **[x]**

D. For self-attention, the queries (Q) are computed based on target sequences, while keys (K), and values (V) are computed based on source sequences

E. Self-attention can capture **intra-**sequence dependencies between source sequences **[x]**

F. Self-attention can capture inter-sequence dependencies between source sequences and target sequences

Question 2

- What are correct about the cross-attention?

A. With self-attention, we can compute the token embeddings in parallel.

B. With self-attention, we need to compute the token embeddings sequentially.

C. For self-attention, the queries (Q), keys (K), and values (V) computed based on source sequences

D. For self-attention, the queries (Q) are computed based on target sequences, while keys (K), and values (V) are computed based on source sequences

E. Self-attention can capture **intra-**sequence dependencies between source sequences

F. Self-attention can capture inter-sequence dependencies between source sequences and target sequences

Question 2

- What are correct about the self-attention?

A. With self-attention, we can compute the token embeddings in parallel. **[x]**

B. With self-attention, we need to compute the token embeddings sequentially.

C. For self-attention, the queries (Q), keys (K), and values (V) computed based on source sequences

D. For self-attention, the queries (Q) are computed based on target sequences, while keys (K), and values (V) are computed based on source sequences **[x]**

E. Self-attention can capture **intra-**sequence dependencies between source sequences

F. Self-attention can capture inter-sequence dependencies between source sequences and target sequences **[x]**

Question 3

- What are the drawbacks of CNNs?

A. CNNs cannot capture the global information of images.
B. CNN can capture the local information of images.
C. CNNs cannot capture the spatial relationship of local objects inside images
D. CNNs are locality sensitivity
E. CNNs combine local patterns to learn broader local patterns

Question 3

- What are the drawbacks of CNNs?

A. CNNs cannot capture the global information of images. **[x]**
B. CNN can capture the local information of images.
C. CNNs cannot capture the spatial relationship of local objects inside images **[x]**
D. CNNs are locality sensitivity **[x]**
E. CNNs combine local patterns to learn broader local patterns

Question 4

- What are correct about Vision Transformers (ViTs)?

A. For ViTs, a token or visual word is a patch of an image.

B. For ViTs, a token or visual word is a pixel of an image.

C. We input the patch tokens directly to the transformer block

D. We apply a linear projection to the flattened patches to transform them to token embeddings.

E. We inject the class token to the token embeddings of the patches and keep the class token fixed during training

F. We inject the class token to the token embeddings of the patches and learn the class token fixed during training

G. On top of the class token at the final layer, we build up the MLP head to make predictions.

Question 4

- What are correct about Vision Transformers (ViTs)?

A. For ViTs, a token or visual word is a patch of an image. **[x]**

B. For ViTs, a token or visual word is a pixel of an image.

C. We input the patch tokens directly to the transformer block

D. We apply a linear projection to the flattened patches to transform them to token embeddings. **[x]**

E. We inject the class token to the token embeddings of the patches and keep the class token fixed during training

F. We inject the class token to the token embeddings of the patches and learn the class token fixed during training **[x]**

G. On top of the class token at the final layer, we build up the MLP head to make predictions. **[x]**

Question 5

- What is the reason that Vision Transformers (ViTs) can capture the global information of images?

A. This is because the point-wise FFN of the Encoder blocks.
B. This is because the Add & Layer norm operations of the Encoder blocks.
C. This is because the Multi-head Self-attention layers of the Encoder blocks.
D. The global information is captured in the class token at the final layer because this summarizes the token embeddings at the input layer.
E. The global information is captured in the class token at the input layer

Question 5

- What is the reason that Vision Transformers (ViTs) can capture the global information of images?

A. This is because the point-wise FFN of the Encoder blocks.
B. This is because the Add & Layer norm operations of the Encoder blocks.
C. This is because the Multi-head Self-attention layers of the Encoder blocks. **[x]**
D. The global information is captured in the class token at the final layer because this summarizes the token embeddings at the input layer. **[x]**
E. The global information is captured in the class token at the input layer

Question 6

- What are correct about Vision Transformers (ViTs)?

A. ViTs can naturally capture the global information of images.

B. ViTs can be trained directly on small-scaled datasets.

C. ViTs can find the long-term dependencies among image patches.

D. We need massive datasets to train ViTs.

E. ViTs are more robust to patch permutation and occlusion than CNNs

Question 6

- What are correct about Vision Transformers (ViTs)?

A. ViTs can naturally capture the global information of images. **[x]**
B. ViTs can be trained directly on small-scaled datasets.
C. ViTs can find the long-term dependencies among image patches. **[x]**
D. We need massive datasets to train ViTs. **[x]**
E. ViTs are less robust to patch permutation and occlusion than CNNs

Question 7

□ What are correct about Swin Transformers?

A. Swin Transformers employ smaller patches of [3,4,4].
B. Swin Transformers employ patches of [3,16,16] similar to ViTs.
C. Swin Transformers apply a linear projection to flattened patches to gain [C, H/4, W/4].
D. Swin Transformers apply a linear projection to flattened patches to gain [C, H/16, W/16].
E. For Swin Transformers, the input shape [C, H/4, W/4] is kept fixed across the layers.
F. For Swin Transformers, we apply patch merging to down-sample the input shape by two while doubling the depth.

Question 7

- What are correct about Swin Transformers?

A. Swin Transformers employ smaller patches of [3,4,4]. **[x]**

B. Swin Transformers employ patches of [3,16,16] similar to ViTs.

C. Swin Transformers apply a linear projection to flattened patches to gain [C, H/4, W/4]. **[x]**

D. Swin Transformers apply a linear projection to flattened patches to gain [C, H/16, W/16].

E. For Swin Transformers, the input shape [C, H/4, W/4] is kept fixed across the layers.

F. For Swin Transformers, we apply patch merging to down-sample the input shape by two while doubling the depth. **[x]**

Question 8

- What are correct about Patch Merging in Swin Transformers?

A. We merge 2x2 neighbourhood patches, concatenate their embeddings, and then apply a linear projection.

B. We apply a linear projection directly to token embeddings.

C. If we input the patch merging [C, H/4, W/4], we gain [2C, H/4, W/4]

D. If we input the patch merging [C, H/4, W/4], we gain [C, H/4, W/4]

E. If we input the patch merging [C, H/4, W/4], we gain [2C, H/8, W/8]

Question 8

- What are correct about Patch Merging in Swin Transformers?

A. We merge 2x2 neighbourhood patches, concatenate their embeddings, and then apply a linear projection. **[x]**

B. We apply a linear projection directly to token embeddings.

C. If we input the patch merging [C, H/4, W/4], we gain [2C, H/4, W/4]

D. If we input the patch merging [C, H/4, W/4], we gain [C, H/4, W/4]

E. If we input the patch merging [C, H/4, W/4], we gain [2C, H/8, W/8] **[x]**

Question 9

- What are correct about Window Self-Attention in Swin Transformers?

A. We apply the Self-Attention to all token embeddings.

B. We divide all token embeddings into many local windows and then apply the Self-Attention to each local windows independently.

C. The output shape of Window Self-Attention is different from the input shape.

D. The output shape of Window Self-Attention is the same as the input shape.

Question 9

- What are correct about Window Self-Attention in Swin Transformers?

A. We apply the Self-Attention to all token embeddings.

B. We divide all token embeddings into many local windows and then apply the Self-Attention to each local windows independently. **[x]**

C. The output shape of Window Self-Attention is different from the input shape.

D. The output shape of Window Self-Attention is the same as the input shape. **[x]**

Question 10

□ What are correct about Window Self-Attention in Swin Transformers?

A. The Window Self-Attention can speed up the standard Self-Attention
B. The Window Self-Attention is slower than the standard Self-Attention The output shape of Window Self-Attention is different from the input shape.
C. The Window Self-Attention only allows a token to interact with the ones in the same local window.
D. The Window Self-Attention allows a token to interact with the ones in the different local windows.
E. The Window Self-Attention enables the interaction across local windows.

Question 10

- What are correct about Window Self-Attention in Swin Transformers?

A. The Window Self-Attention can speed up the standard Self-Attention **[x]**

B. The Window Self-Attention is slower than the standard Self-Attention The output shape of Window Self-Attention is different from the input shape.

C. The Window Self-Attention only allows a token to interact with the ones in the same local window. **[x]**

D. The Window Self-Attention allows a token to interact with the ones in the different local windows.

E. The Window Self-Attention enables the interaction across local windows.

Question 11

- What are correct about Shifted Window Self-Attention in Swin Transformers?

A. The Shifted Window Self-Attention only allows a token to interact with the ones in the same local window.

B. The Shifted Window Self-Attention allows a token to interact with the ones in the different local windows.

C. The Shifted Window Self-Attention enables the interaction across local windows.

D. The Shifted Window Self-Attention shifts a local window to right and bottom to become a new local window

E. The output shape of Shifted Window Self-Attention is different from the input shape.

F. The output shape of Shifted Window Self-Attention is the same as the input shape.

Question 11

- What are correct about Shifted Window Self-Attention in Swin Transformers?

A. The Shifted Window Self-Attention only allows a token to interact with the ones in the same local window.

B. The Shifted Window Self-Attention allows a token to interact with the ones in the different local windows. **[x]**

C. The Shifted Window Self-Attention enables the interaction across local windows. **[x]**

D. The Shifted Window Self-Attention shifts a local window to right and bottom to become a new local window **[x]**

E. The output shape of Shifted Window Self-Attention is different from the input shape.

F. The output shape of Shifted Window Self-Attention is the same as the input shape. **[x]**

Question 12

- What is the principle of the model fine-tuning for ViTs with additional components?

A. We insert additional components to pretrained ViTs that favour the original computation of ViTs and then fine-tune the additional components

B. We insert additional components to pretrained ViTs that favour the original computation of ViTs and then freeze the additional components

C. We insert additional components to pretrained ViTs that favour the original computation of ViTs and then consider the additional components as variables to optimize in optimizers

D. We insert additional components to pretrained ViTs that require us to significantly modify the original computation of ViTs and then fine-tune the additional components

Question 12

- What is the principle of the model fine-tuning for ViTs with additional components?

A. We insert additional components to pretrained ViTs that favour the original computation of ViTs and then fine-tune the additional components **[x]**

B. We insert additional components to pretrained ViTs that favour the original computation of ViTs and then freeze the additional components

C. We insert additional components to pretrained ViTs that favour the original computation of ViTs and then consider the additional components as variables to optimize in optimizers **[x]**

D. We insert additional components to pretrained ViTs that require us to significantly modify the original computation of ViTs and then fine-tune the additional components

Question 13

- What are correct about the model fine-tuning for ViTs with prompt-tuning?

A. We insert learnable prompts to token embeddings of ViTs and then fine-tune these prompts

B. We insert learnable prompts to pointwise networks of ViTs and then fine-tune these prompts

C. We insert learnable prompts to the key, query, and value matrices of ViTs and then fine-tune these prompts

Question 13

- What are correct about the model fine-tuning for ViTs with prompt-tuning?

A. We insert learnable prompts to token embeddings of ViTs and then fine-tune these prompts **[x]**

B. We insert learnable prompts to pointwise networks of ViTs and then fine-tune these prompts

C. We insert learnable prompts to the key, query, and value matrices of ViTs and then fine-tune these prompts

Question 14

- What are correct about the model fine-tuning for ViTs with adapters?

A. We insert adapters to token embeddings of ViTs and then fine-tune these adapters
B. We insert adapters to pointwise networks of ViTs and then fine-tune these adapters
C. We insert adapters to the key, query, and value matrices of ViTs and then fine-tune these adapters

Question 14

- What are correct about the model fine-tuning for ViTs with adapters?

A. We insert adapters to token embeddings of ViTs and then fine-tune these adapters

B. We insert adapters to pointwise networks of ViTs and then fine-tune these adapters **[x]**

C. We insert adapters to the key, query, and value matrices of ViTs and then fine-tune these adapters

Question 15

- What are correct about the model fine-tuning for ViTs with LoRA?

A. We insert low-ranked matrices to token embeddings of ViTs and then fine-tune these low-ranked matrices

B. We insert low-ranked matrices to pointwise networks of ViTs and then fine-tune these low-ranked matrices

C. We insert low-ranked matrices to the key, query, and value matrices of ViTs and then fine-tune these low-ranked matrices

Question 15

- What are correct about the model fine-tuning for ViTs with LoRA?

A. We insert low-ranked matrices to token embeddings of ViTs and then fine-tune these low-ranked matrices

B. We insert low-ranked matrices to pointwise networks of ViTs and then fine-tune these low-ranked matrices

C. We insert low-ranked matrices to the key, query, and value matrices of ViTs and then fine-tune these low-ranked matrices **[x]**

Thanks for your attention!