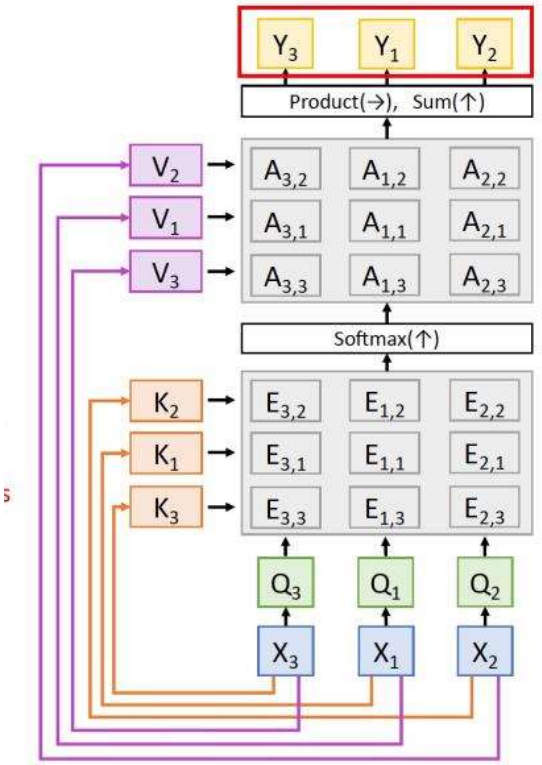
1. As shown in figure 1, assuming that two self-attention layer share the same parameter, is the feature $X1$ equal to $X2$? Please illustrate how to make your point. (15 points)

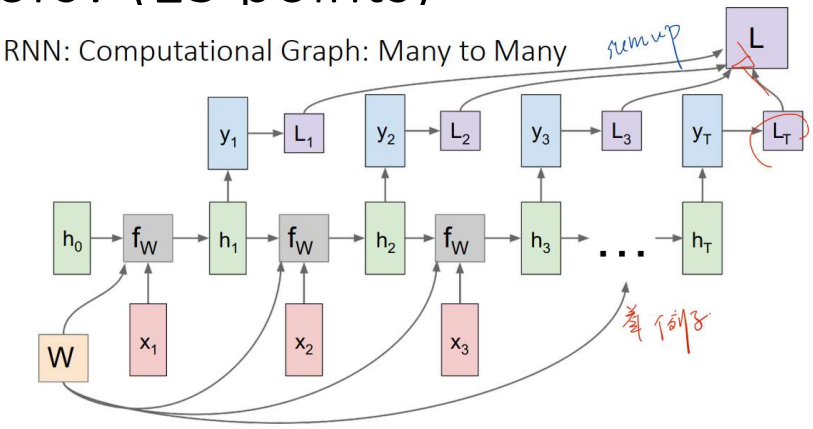
- Yes, $X1$ equals to $X2$

- Because permuting the input vectors Q will also permute key vectors K and value vectors V. Then they take element wise products in self-attention layer, so the output features vector Y will still be the same, but just permuted. Since there is permute back operation in our case, the final output X1 will be equal to X2.

- Because self-attention layer is permutation equivariant: f(s(x)) = s(f(x)), s() is the self-attention layer, and f() is the permutation operation.



2. What is RNN_____(full spells)? If computational graph in RNN is “many to many”, do you need to use different parameter sets for different input vectors? (15 points)



- RNN is Recurrent Neural Networks
- The parameter sets or weight matrix (W) for different input vectors should be the same at every time-step.
- Because if the parameters change with input vectors, the total number of parameters would grow with the length of the inputs. Large number of parameters would require massive computational resources, and prone to overfit to training data.
- Moreover, parameter sharing reflects the fact that we are doing the same task for different input vectors in the sequence.

3. Please make comparisons among R-CNN, Fast R-CNN, Faster R-CNN and explain what do “fast” and “faster” mean. (15 points)

- ANS: They are all object detection networks

1. R-CNN first get RoI from proposal method and forward each region through ConvNet independently
2. Fast R-CNN: Run whole image through ConvNet and crop on feature map based on RoI Alignment, then pass the cropped features to object classification and box offset
3. Faster R-CNN: Insert Region Proposal Network (RPN) to predict proposals from features, and then use RoI pooling before classification and bounding box regression

- “Fast” and “Faster” indicates that the whole running time of the object detection network is lesser during the training and testing process.

4. As shown in Figure 2, object detectors often output many overlapping detections. How to solve this problem? (15 points)

- ANS: Using Non-Max Suppression (NMS), steps as below:
  1. Select next highest-scoring box
  2. Eliminate lower-scoring boxes with IoU (Intersection over Union) > threshold
  3. If any boxes remain, Go back to step 1
- Detailed Steps:
  1. Select the blue box which has the highest-scoring
  2. Remove orange box whose IoU with blue box is bigger than threshold
  3. Select the purple box which has the next highest-scoring
  4. Remove yellow box whose IoU with purple box is bigger than threshold
  5. Only blue and purple boxes are remained and they are not overlapping

5. Please illustrate inputs and outputs for the task of semantic segmentation with specific dimension i.e. $H \times W \times C$. Upsampling is the key part in fully convolution network for semantic segmentation. How to do the upsampling? (15 poins)

- Inputs are images and semantic categories.
- Output a image, and each pixel of the new image is labeled with a semantic category (pixel wise prediction).

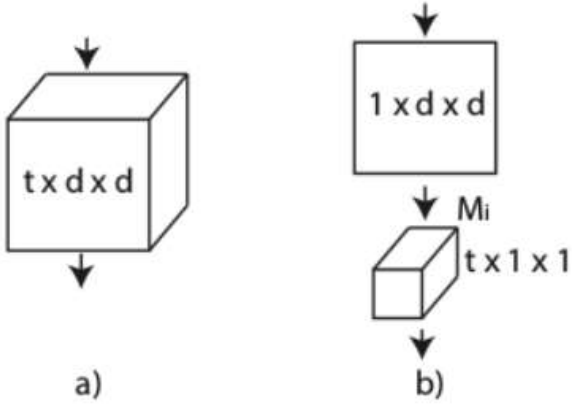
- "Upsample" is the reversed operations of pooling. For fully convolution network, we can use strided transpose convolution to achieve that. Transpose convolution take a single value from the low-resolution feature map and multiply all of the weights in our filter by this value, projecting those weighted values into the output feature map. Overlapping values are simply added together. If the resulted output is bigger than the desired output, extra pixels are trimmed.

6. Please compare the 3D convolution and R(2+1)D convolution for spatiotemporal modeling and list key differences. You can choose to describe in words or draw diagrams. (15 points)

ANS:

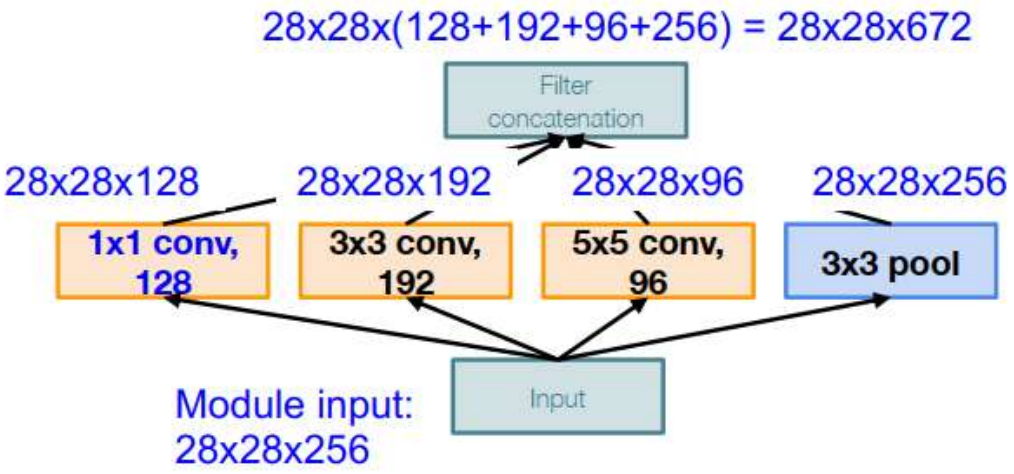
(a) 3D convolution use a kernel of size txdxd, where d is the spatial width and height, and t is the temporal extent.

(b) indicates R(2+1)D convolution. It consists of a spatial 2D convolution and followed by a temporal 1D convolution. Mi is the numbers of 2D spatial filters.



- 3D convolution kernel retains the temporal information with spatial information which can be transmitted together between 3d conv layers. While R(2+1)D decompose 3D kernel into two distinct spatial and temporal kernels, which allows the network to locally learn spatial and spatiotemporal features in distinct layers.

7. Please draw the diagram of the basic inception block with specific feature sizes. (10 points)



- This is the basic Naïve inception block with a input feature size of 28X28X256
- Conv Ops: [1x1 conv, 128] 28x28x128x1x1x256 [3x3 conv, 192] 28x28x192x3x3x256 [5x5 conv, 96] 28x28x96x5x5x256 Total: 854M ops