Faster R-CNN:

Visual Encoder: Faster R-CNN object detector working Encoder in image captioning. To compute the GFLOPS for faster r-cnn we use the standard size (VGA image resolution) of object detection. The Faster R-CNN typically consists of:

1. **Feature Extractor Backbone of Faster R-CNN**
2. **RPN (Region Proposal Network)**
3. **Classification and Regression Layer of Proposal**

1. Feature Extractor Backbone: In the image captioning encoder, ResNet101 is typically employed within Faster R-CNN, utilizing only the convolutional layers of ResNet101. As a result, the feature extractor involves a computational cost of **96.1** GFLOPs and 42.5 Million Parameters.

2. **RPN (Region Proposal Network)**: RPN uses three convolution layers, which are:

conv): Conv2d(1024, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
(objectness_logits): Conv2d(512, 12, kernel_size=(1, 1), stride=(1, 1))
(anchor_deltas): Conv2d(512, 48, kernel_size=(1, 1), stride=(1, 1))

So, GFLOPs is 11.398 GFLOPs and 4.749 million

3. **Classification and Regression Layer of Proposal:** In Faster R-CNN, 300 proposals from RPN were used for classification and regression. Each classification and regression 0.032 GFLOPs, which are:

(box_predictor): FastRCNNOutputLayers(
(cls_score): Linear(in_features=2048, out_features=1601, bias=True)
(bbox_pred): Linear(in_features=2048, out_features=6400, bias=True)
)

So, 300 proposal GFLOPs is 9.831 GFLOPs, and the parameter 16.386 million (considers only one proposal)