# Image Captioning

Faculty:-
Prof.C Krishna Mohan

Mentor:-
Prudviraj Jeripothula

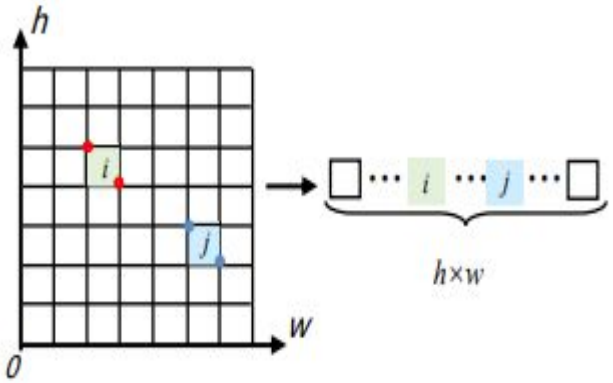
Presented By:-
Subhasree Balija

# Contents

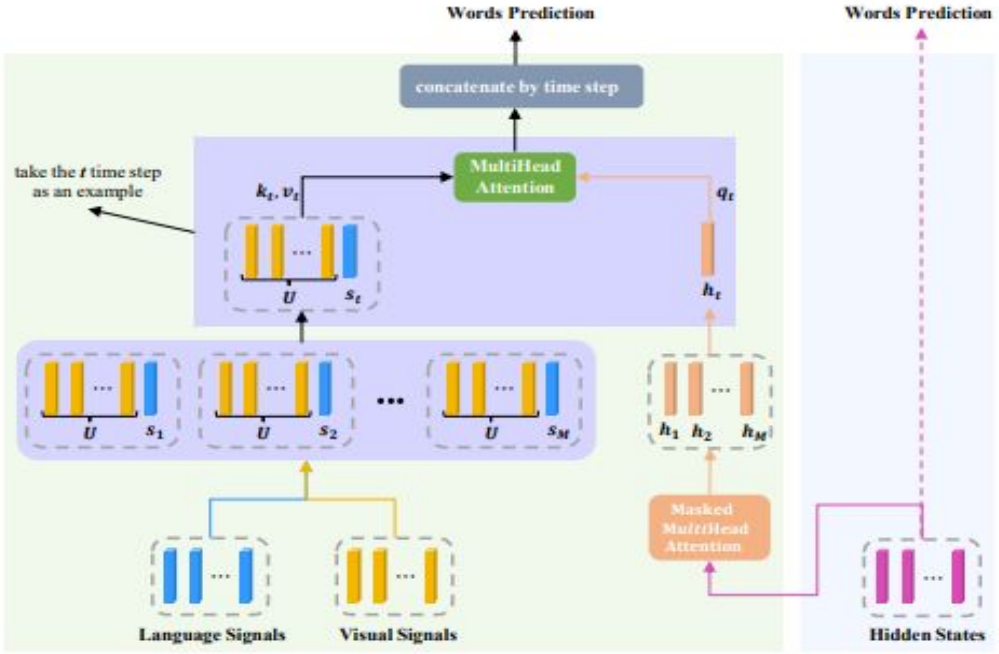
- Summary of Prev Presentation
- Image Captioning Results
- Novelty
- Conclusion
- References

# Captioning with Adaptive Attention on Visual and Non-Visual words

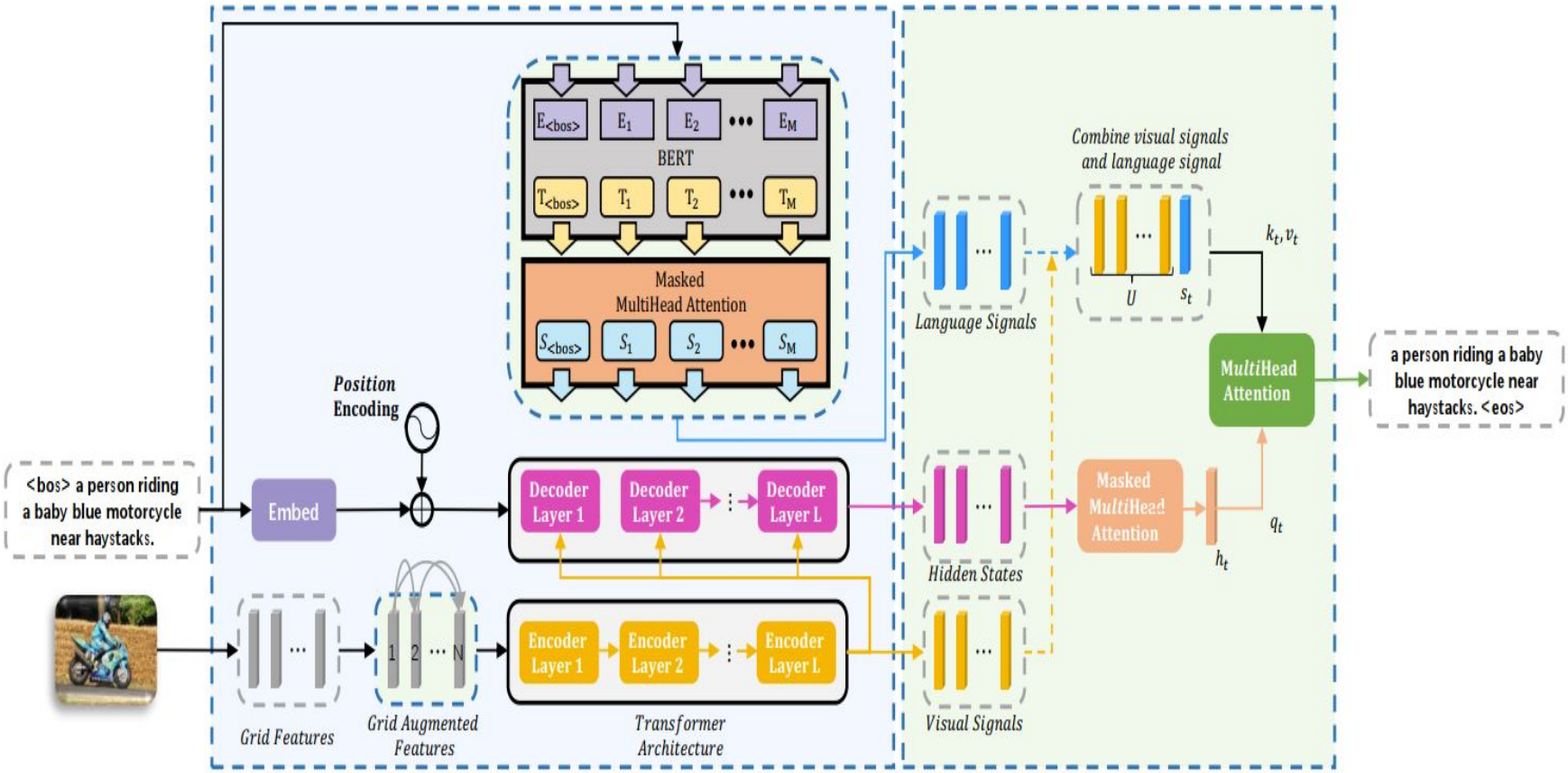
## Grid-Augmented Module




## Adaptive Attention Module



# Relationship Sensitive Transformer(RSTNet)



# Image Captioning Results

Training Details:

Datasets - MSCOCO 2014

Standard RESNet-50  as Visual backbone, Imagenet dataset

Input :224 × 224 image ,Output: 7 × 7 grid of 2048-dimensional features

Use Linear Projection 7 x 7 x 2048 -> 7 x 7 x H

Transformer as Linguistic decoder

GELU as Activation function

Tokenize captions with Sentence piece using BPE algorithm

SGD with momentum 0.9 , weight decay 10^-4

Learning Rate of RESNet 0.2,Transformer 0.001

# Training Results

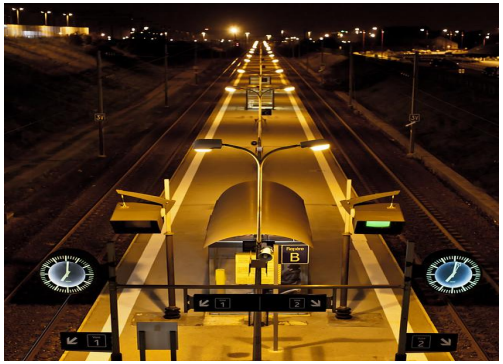
| epoch | Train loss | Train cap acc |
|---|---|---|
| 0 | 15.176218 | 0.058669 |
| 25 | 3.111872 | 0.114059 |
| 50 | 2.467931 | 0.123626 |
| 100 | 1.024006 | 0.234425 |
| 150 | 0.804063 | 0.241819 |

| epoch | Val loss | Val cap acc |
|---|---|---|
| 0 | 6.619026 | 0.072032 |
| 25 | 3.249872 | 0.114138 |
| 50 | 0.119163 | 0.119163 |
| 100 | 0.935401 | 0.238409 |
| 150 | 0.759156 | 0.244353 |

Total 150 epochs

# Qualitative Results

Man standing next to truck on dirt road with trees



An empty boarding walk is lit up in yellow at night.

A sheep standing in the middle of a snowyic flower garden.

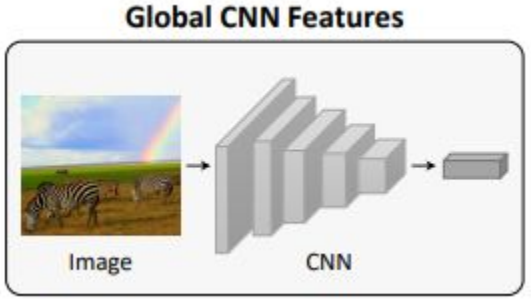
A skate boarder leaning over a very high ramp.

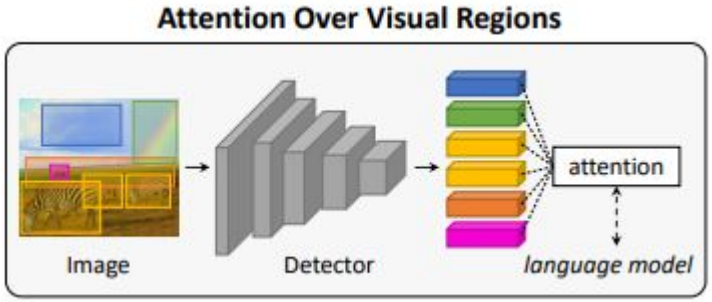
A woman skier on a snowy plain with another skier in the background.

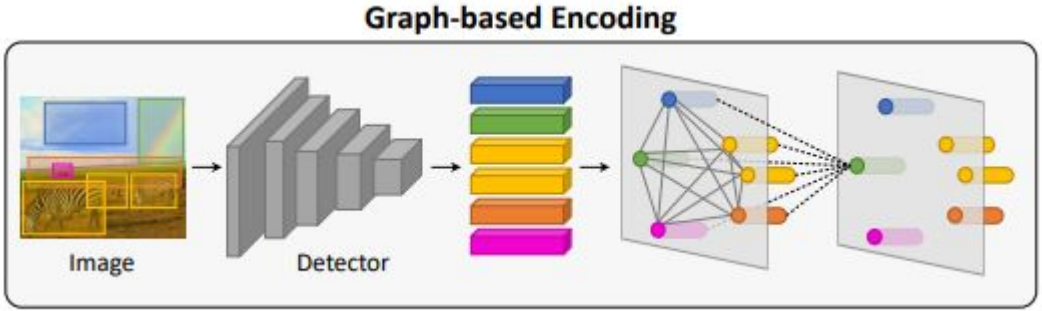
# Novelty

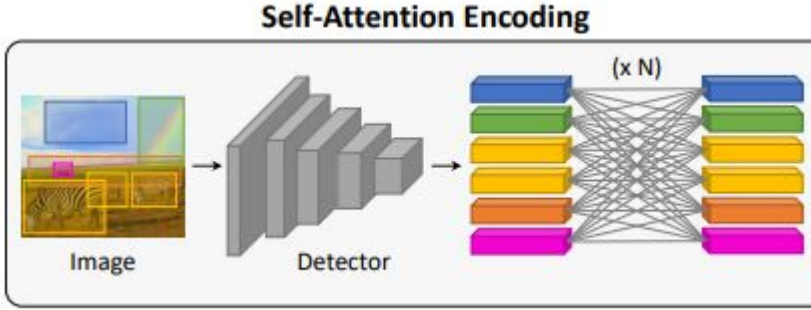
Visual Encoder + Language Decoder

- Non-Attentive CNN’S
- Additive Attention based CNN’s
- Graph-based Attention
- Self-Attention



# Scene Graph Captioning(SGC)

Structured representation of a scene

Express Objects ,Attributes and Relationships

Higher level of understanding and reasoning about visual scenes

- Scene graph generation

  Object, Relationship and Caption regions

- Region Captioning

  language description of  Scene graph

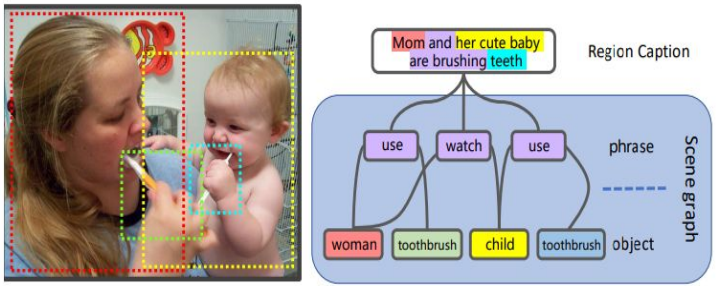
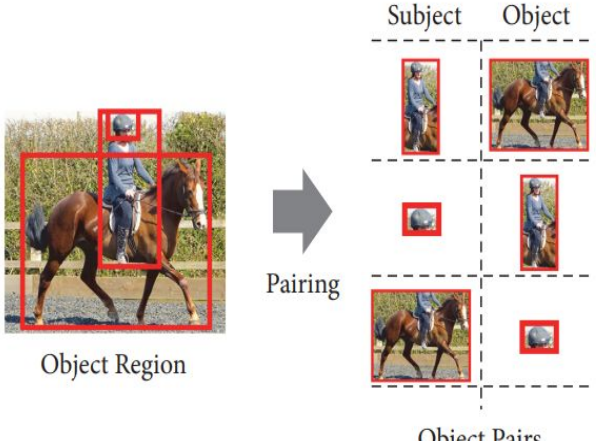
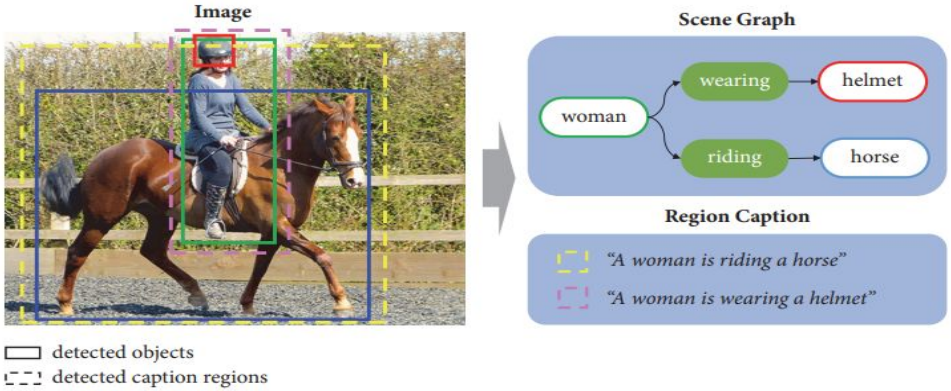



Figure 1. Image with annotations of different semantic levels: objects, phrases and region captions. Scene graph is generated using all objects and their relationships in the image.

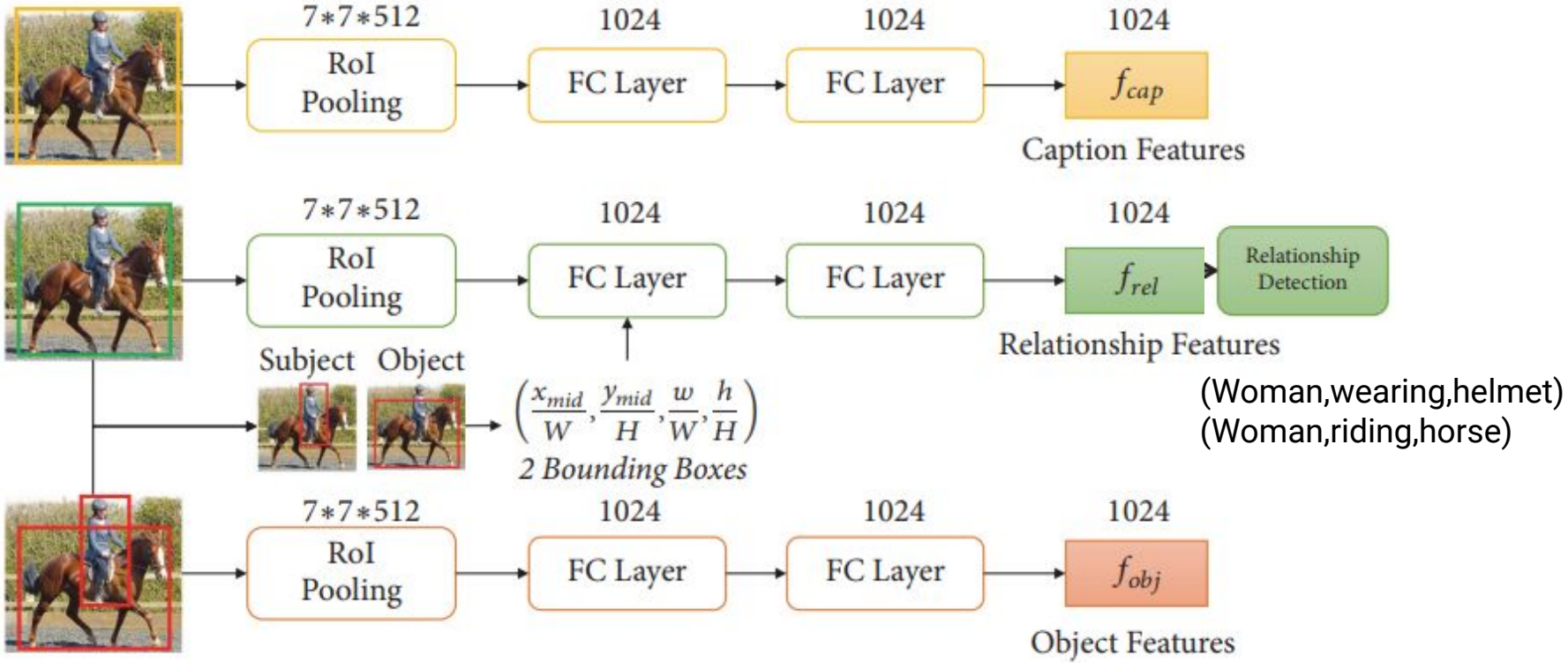
# Scene Graph Generation(SGG)

- Object region proposals- Region Proposal Network (RPN)
- Relationship/phrase region proposals: N object proposals to N(N −1) object pairs
- Caption region proposals:  RPN trained with ground truth captions



# SGG


7*7*512
RoI Pooling
1024
FC Layer
1024
FC Layer
1024
$f_{cap}$
Caption Features
7*7*512
RoI Pooling
1024
FC Layer
1024
FC Layer
1024
$f_{rel}$
Relationship Detection
Relationship Features
Subject
Object
$\left(\frac{x_{mid}}{W}, \frac{y_{mid}}{H}, \frac{w}{W}, \frac{h}{H}\right)$
2 Bounding Boxes
(Woman,wearing,helmet)
(Woman,riding,horse)
7*7*512
RoI Pooling
1024
FC Layer
1024
FC Layer
1024
$f_{obj}$
Object Features

# Dynamical Graph Construction

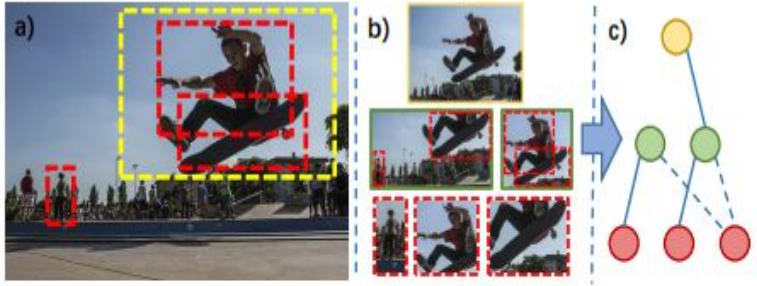
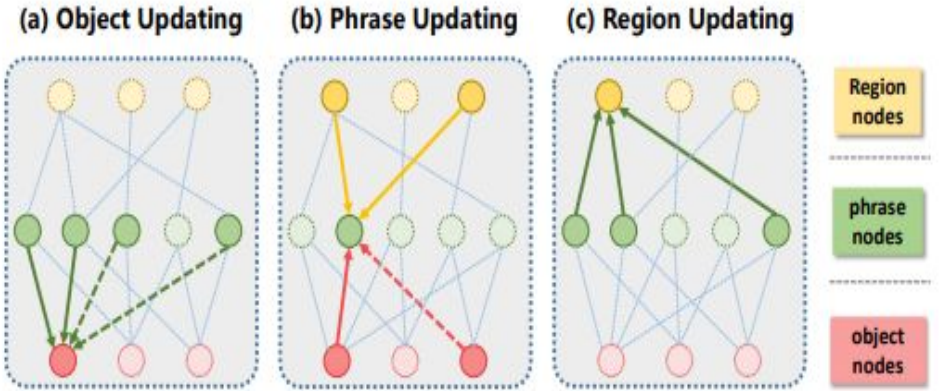
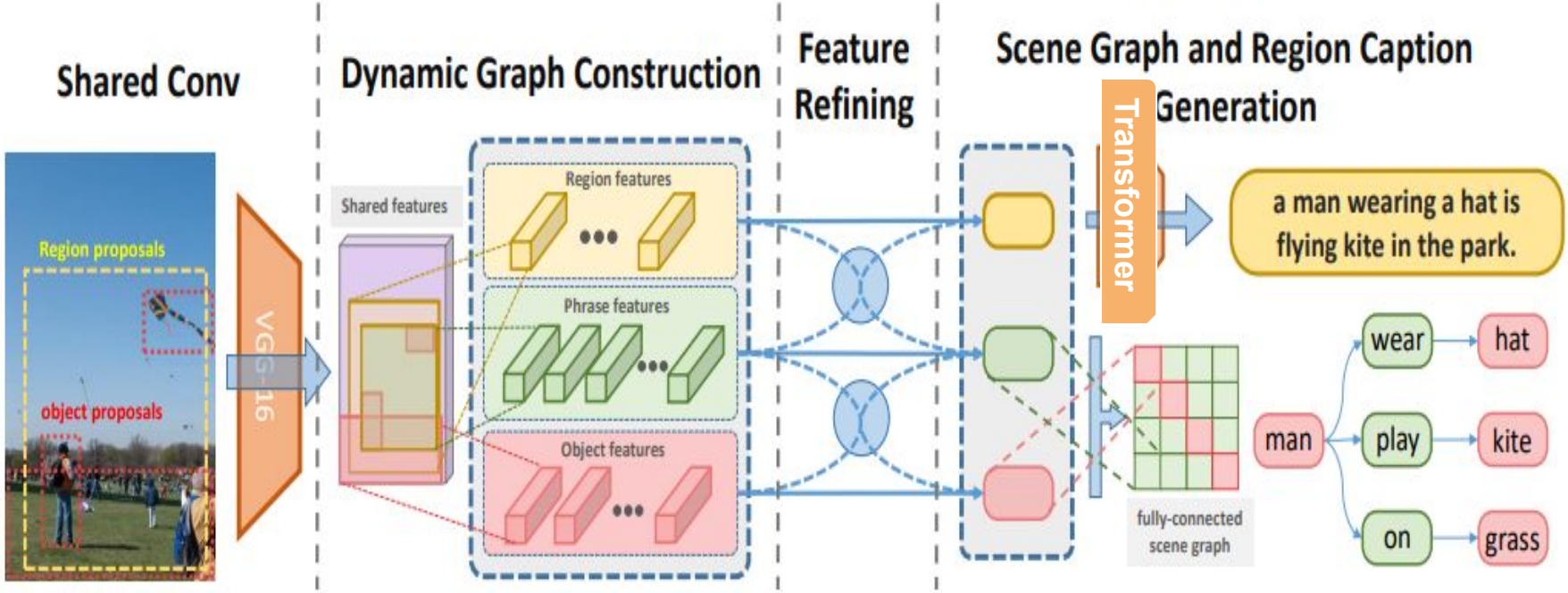



Figure 3. Dynamical graph construction. (a) the input image. (b) object(bottom), phrase(middle) and caption region(top) proposals. (c) The graph modeling connections between proposals. Some of the phrase boxes are omitted.

# Feature Refining

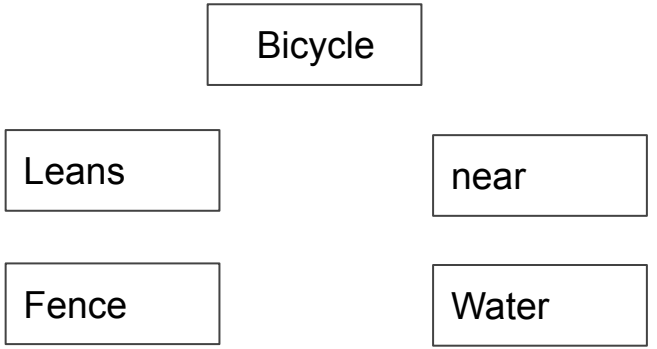


# Model


Shared Conv
Dynamic Graph Construction
Feature Refining
Scene Graph and Region Caption Generation
Region proposals
object proposals
VGG-16
Shared features
Region features
Phrase features
Object features
Transformer
a man wearing a hat is flying kite in the park.
fully-connected scene graph
man
wear
hat
play
kite
on
grass

# Background Features

Scene graph completely neglects Background features




GT:A bicycle leaning against a fence in a flooded street

SGC-A bicycle leans against a fence near water

# Language Decoder

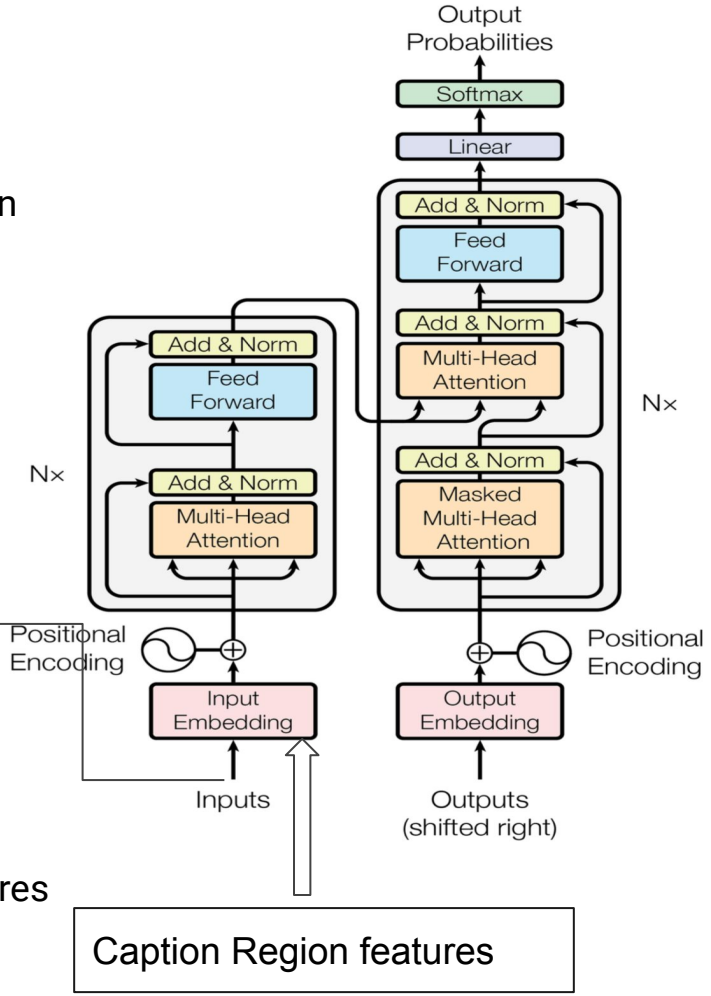
Inputs - Background features +One hot encodings of  Caption Region features

Output-Caption

Resnet

Extracting Background Features

Caption Region features



# Conclusion

Scene graphs + Background features + Transformer

Provide Powerful representations for the semantic features of a scene

Extend to 2D/3D scene understanding,VQA,human-object interaction (HOI),Image Generation

Generation of SG is more Complex and Time Consuming

**Future Works:**

SGG With Prior Knowledge:

Difficult to get all relationships from the SGG training data

Introduction of prior knowledge can enhance the detection and recognition of visual relationships

# References


- T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” ICLR, 2017
- X. Yang, K. Tang, H. Zhang, and J. Cai, “Auto-Encoding Scene Graphs for Image Captioning,” in CVPR, 2019.
- Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, “Scene Graph Generation from Objects, Phrases and Region Captions,”
- T. Yao, Y. Pan, Y. Li, and T. Mei, “Exploring Visual Relationship for Image Captioning,” in ECCV, 2018
- S. Herdade, A. Kappeler, K. Boakye, and J. Soares, “Image captioning: Transforming objects into words,” in Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc., 2019.
- S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NIPS, 2015.

THANK YOU