



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

**Guide:-**

**Dr C Krishna Mohan**

**Mentor:-**

**Prudviraj Jeripothula**

**Presented By:-**

**B Subha Sree  
(CS19B1005)**

# Contents



- Introduction
- Motivation
- Challenges
- Problem Statement
- Existing Methods
- Transformers and BERT
- State of Art Experiments and Results
- Conclusion
- References

# Introduction:-

**Image Captioning** refers to the process of generating textual description from an image - based on the objects and actions in the image. It uses both Natural language processing and Computer vision to generate the captions.

It involves three parts:

- Perceiving the visual space
- Grounding to world knowledge in the language space
- Generation of textual sentence



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."

Image Captioning

# Motivation

- Assistance for visually impaired people
- People who are not visually impaired but works in surveillance fields
- Self driving vehicles
- Accelerates the closed captioning for digital content production
- Extend to Visual Question Answering, Video Captioning
- Human-Robot interaction
- Search engines



Surveillance fields



Video: A person is standing next to the sink washing dishes.

RF+Floormap: A person is cooking food on the stove.

GT: A person is cooking with a black pan and spatula on the stove.

# Challenges

## 1. Various Captions

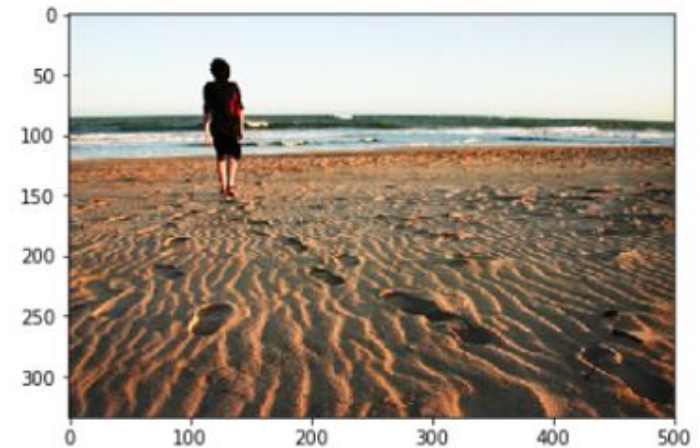
- A white dog in a grassy area
- White dog with brown spots
- A dog on grass and some pink flowers



## 2. MisCaptioning



Greedy: man in black shirt is skateboarding down ramp



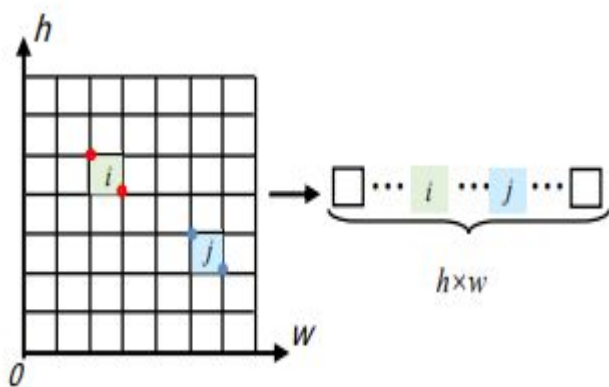
Greedy: a boy is walking on the beach with the ocean .

# Problem statement



Relationship-Sensitive Transformer (RSTNet vanilla transformer model)

## Grid-Augmented module



## Adaptive Attention module



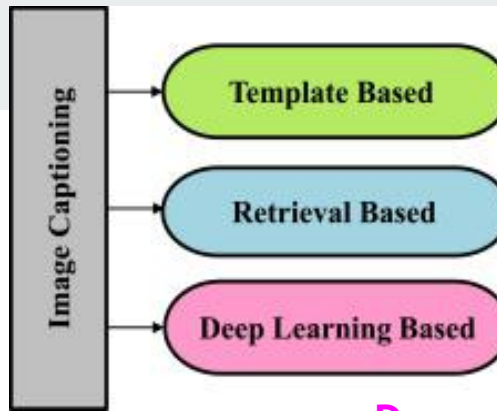
**man** -Visual word  
(Visual signal)

with-Non visual word  
(Language signal)

Caption: A **man** hitting a tennis  
ball **with** a racquet.



# Existing Methods



## Traditional based

### Template based Image Captioning:-

Predefined sentence structure to generate final image caption sentence

More rigid and lack diversity

The predicted nouns, verbs, and scenes are applied to fill in the syntactic structure to compose a description sentence

Cannot generate novel captions for a given image

### Retrieval based Image Captioning:-

Re-use description sentences available from the searched tagged images

Existing human-written phrases from a caption database by measuring the visual similarity

Cannot generate novel captions for a given image

They both need a predefined rigid sentence structure.

## Deep learning based

Image is first encoded into a fixed-length embedding vector by an encoder

Output from the last hidden state of the CNN(Encoder) is given to the first time step of the decoder

To generate final textual description

Encoders are based on convolutional neural networks (CNNs)

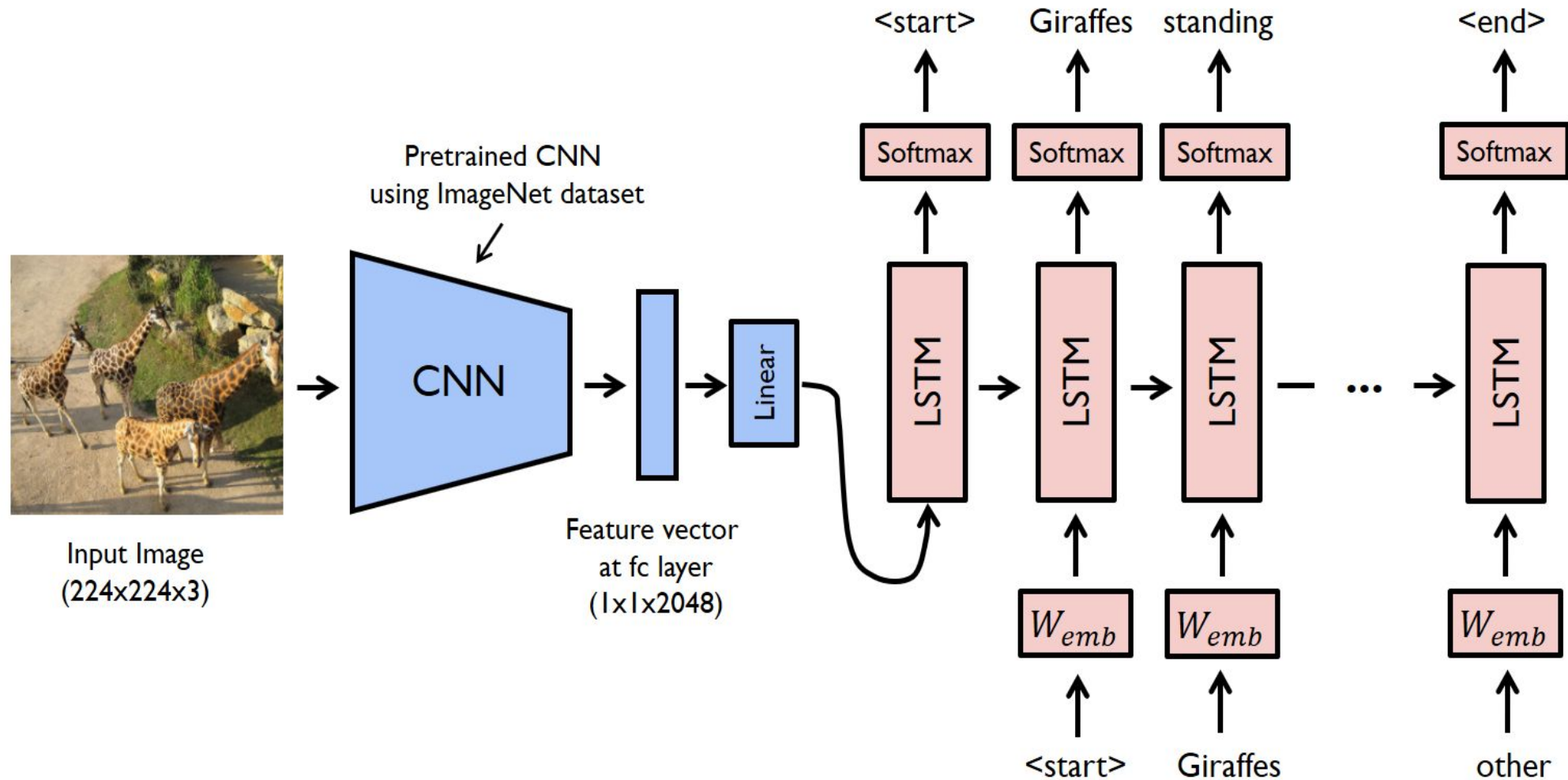
They are effective for object detection and recognition in area of image processing.

Decoders are models used in NLP such as recurrent neural networks (RNNs) like LSTM & Transformers

Can generate novel image captions without a predefined rigid sentence template

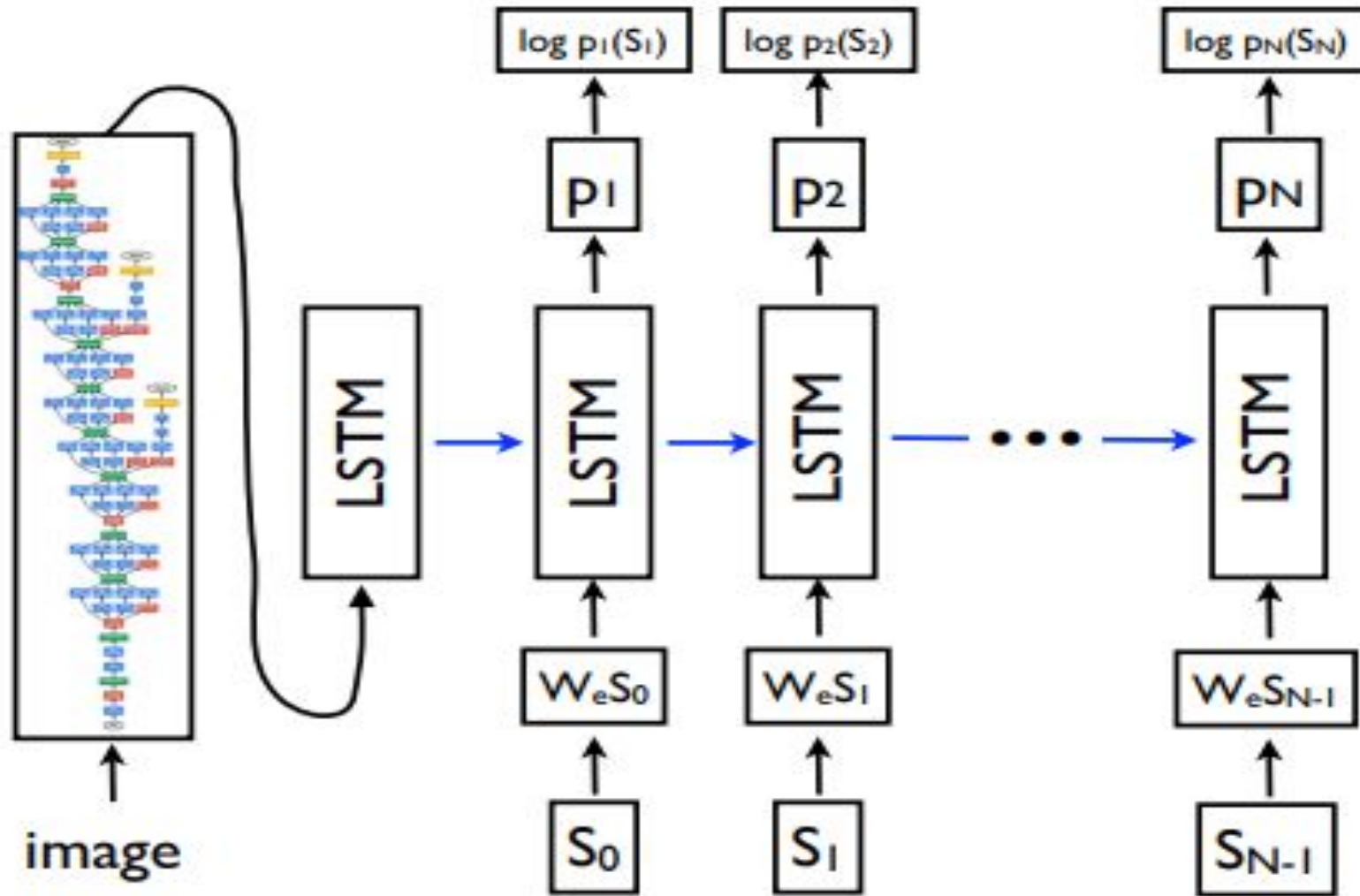
Attention based mechanisms

# Existing Encoder-Decoder framework (2015)



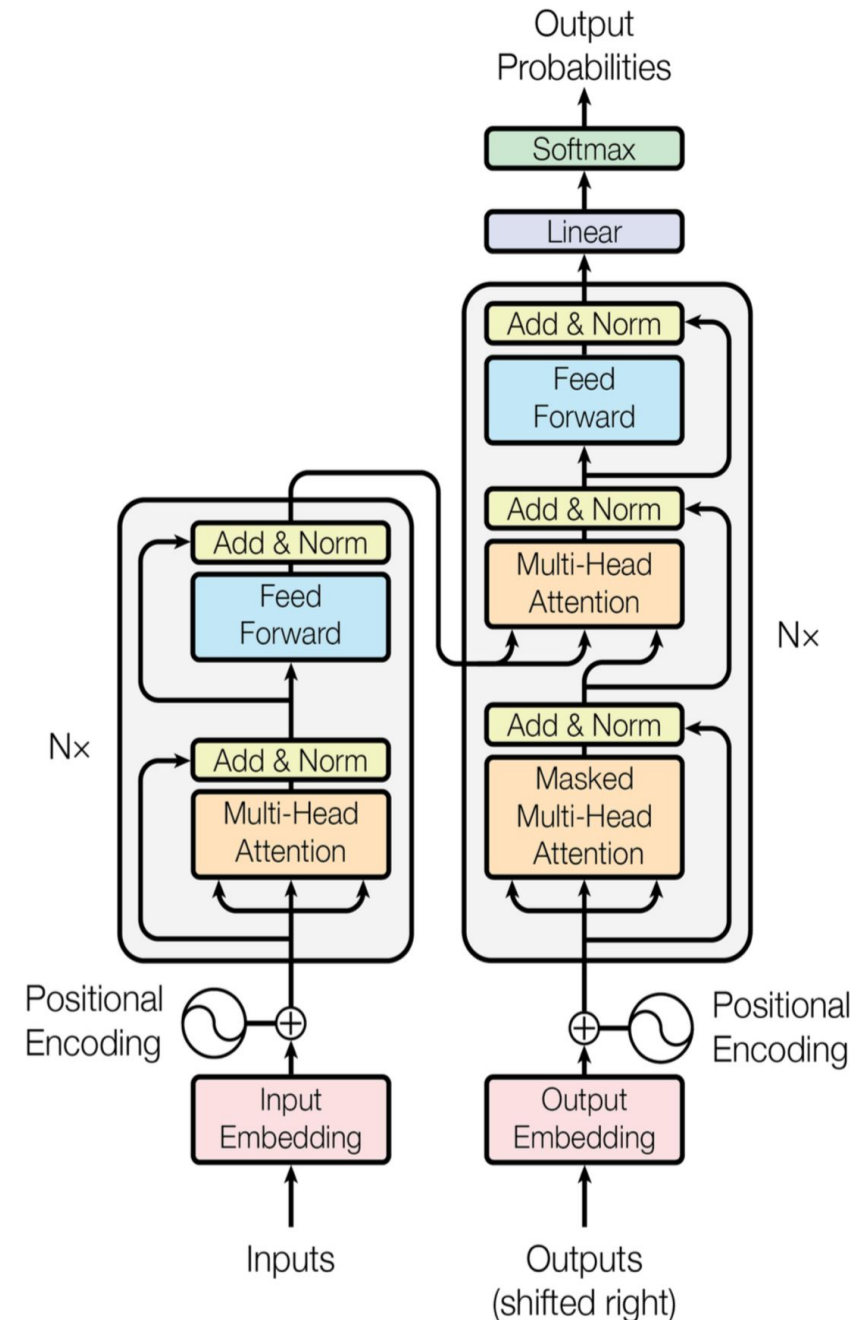


# Encoder-Decoder framework (2015)



# Transformers

- Attention based Encoder-Decoder Architecture
- Machine learning Translation, Conversational chatbots
- Faster than LSTM in learning
- Use attention mechanism
- Input Embedding
- Positional Encoding
- Multi-head attention and Masked multi head attention

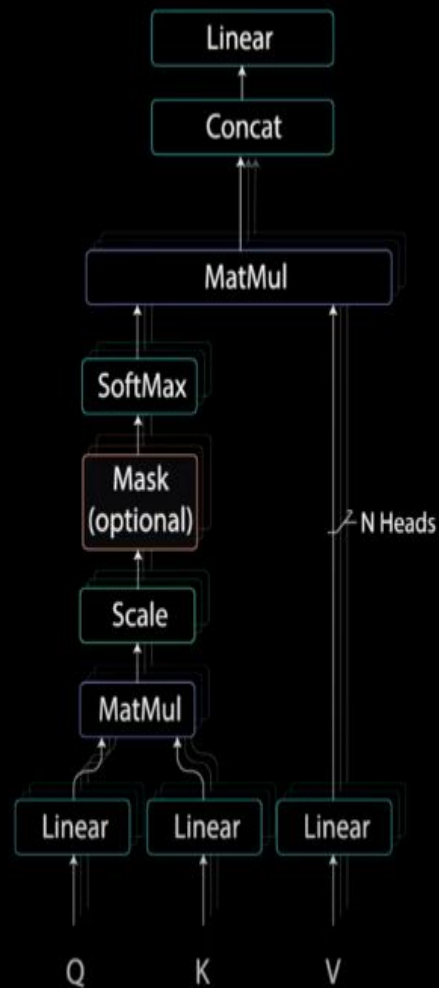


# Multi-Head Attention

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

where  $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$



$$\begin{aligned} X = \{x_1, x_2, x_3, x_4\} &\times \begin{matrix} W_Q \\ \text{Green Grid} \end{matrix} \Rightarrow Q = XW_Q \\ &\times \begin{matrix} W_K \\ \text{Red Grid} \end{matrix} \Rightarrow K = XW_K \\ &\times \begin{matrix} W_V \\ \text{Yellow Grid} \end{matrix} \Rightarrow V = XW_V \end{aligned}$$

# Masked Multi-head Attention

Softmax(

0.7	-inf	-inf	-inf
0.1	0.6	-inf	-inf
0.1	0.3	0.6	-inf
0.1	0.3	0.3	0.3

) =

	<start>	I	am	fine
<start>	1	0	0	0
I	0.37	0.62	0	0
am	0.26	0.31	0.43	0
fine	0.21	0.26	0.26	0.26

# BERT(Bi-directional Encoder Representation from Transformers)

Stack of Transformer Encoders (BERT BASE-12, BERT LARGE -24)

Language features of given sequence

Question Answering, Sentimental Analysis, Search Engine

Bidirectional trained (2500M Words from wiki & 800M Words from books )

Why BERT?(instead of Transformers)

Memory information and hidden information are highly coupled in transformer decoder, resulting in a serious language bias

- PRE-TRAINING (To Understand the language)
- FINE-TUNING (For a Specific task)

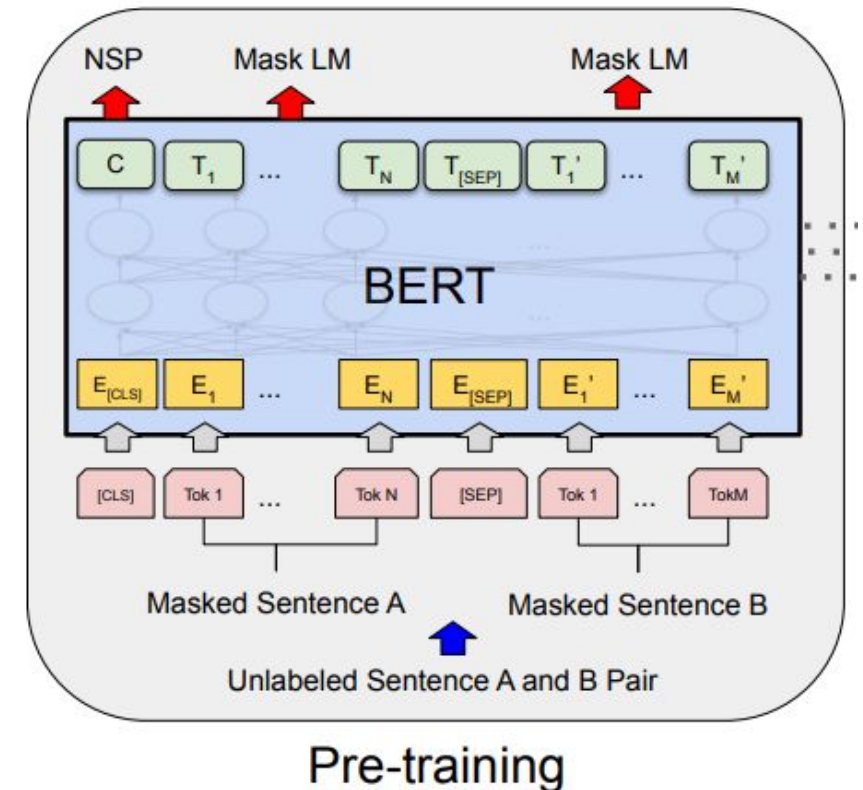
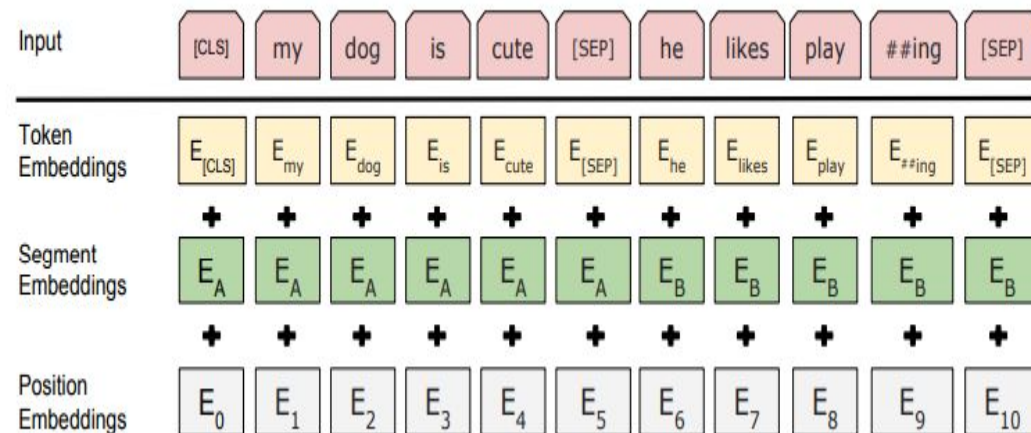
# Pre-training BERT

## Task #1: Masked Language Model(MLM):

- Of 15% of the token positions at random
- [MASK] token 80% of the time
- a random token 10% of the time
- the unchanged token 10% of the time

## Task #2: Next Sentence Prediction (NSP):

Ti will be used to predict the original token with cross entropy loss





# BERT-Based Language Model in RSTNET

Pre-trained BERT model can be fine-tuned with just one additional output layer

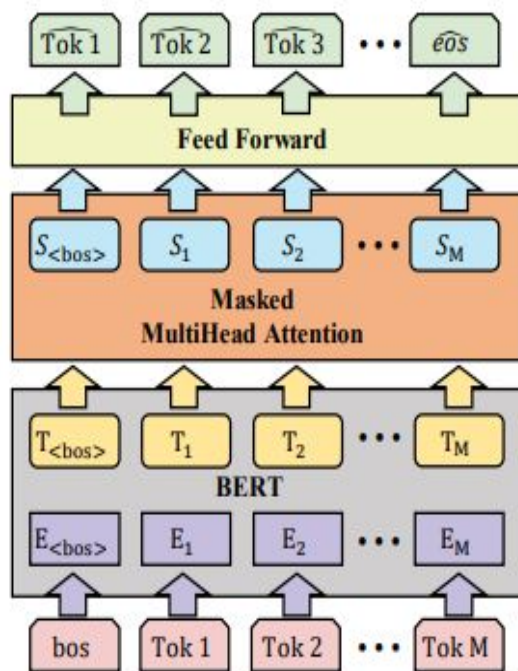


Figure 3. The architecture of our BERT-Based Language Model. The pre-trained BERT model is used to extract language features, and the Masked Multi-Head Attention prevents the word prediction of current step from the interference of the later step.

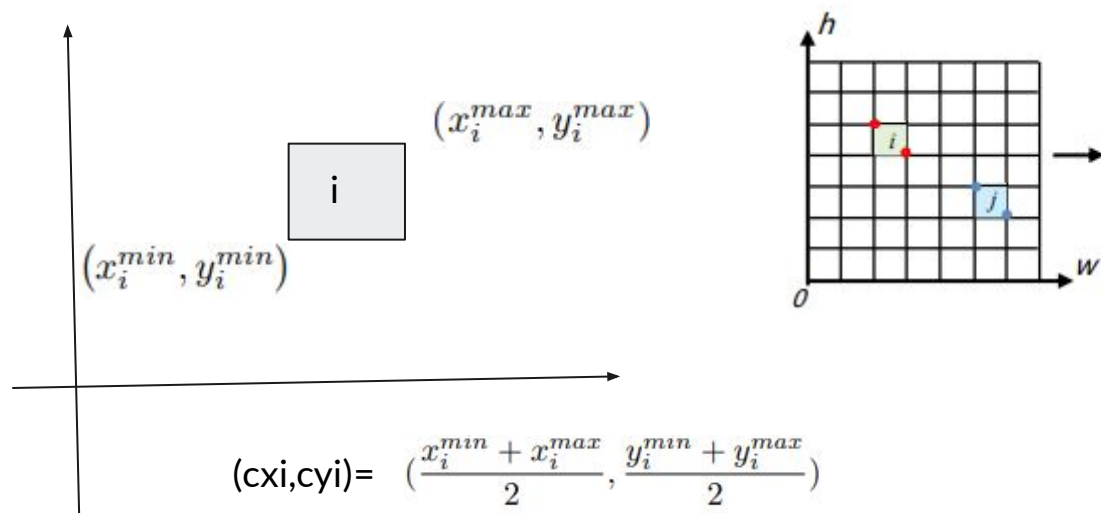
$$\begin{aligned} lf &= BERT(W), \\ S &= MaskedAttentionModule(FF1(lf) + pos), \\ \hat{W} &= log\_softmax(FF2(S)), \end{aligned}$$

$$s_t \leftarrow BBLM(W_{<t}), s_t \in \mathbb{R}^{d_{model}}$$

S is used as the representation of language features in RSTNet

# Grid Augmented module

2D relative positions of each grid  $\{(x_i^{min}, y_i^{min}), (x_i^{max}, y_i^{max})\}$  of grid  $i$



$$w_i = (x_i^{max} - x_i^{min}) + 1,$$

$$h_i = (y_i^{max} - y_i^{min}) + 1.$$

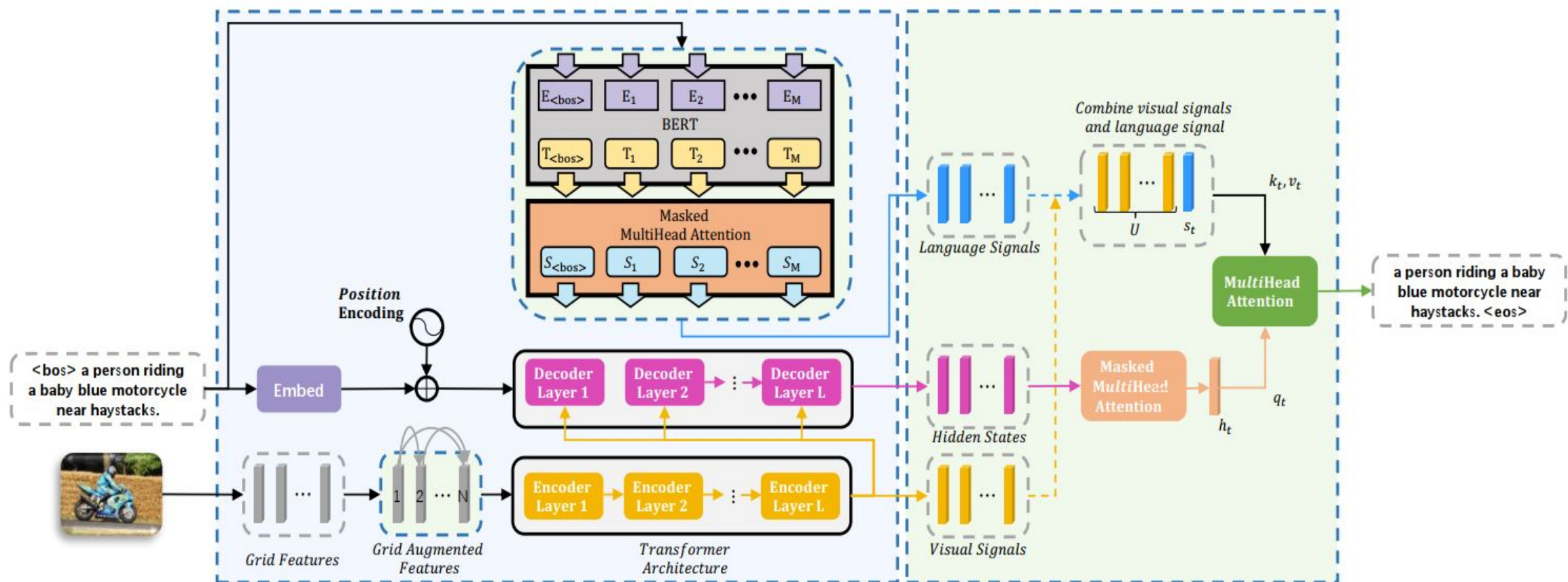
$$r_{ij} = \begin{pmatrix} \log(\frac{|cx_i - cx_j|}{w_i}) \\ \log(\frac{|cy_i - cy_j|}{h_i}) \\ \log(\frac{w_i}{w_j}) \\ \log(\frac{h_i}{h_j}) \end{pmatrix}, \quad (4)$$

$$G_{ij} = FC(r_{ij}), \quad (5)$$

$$\lambda_{ij}^g = ReLU(w_g^T G_{ij}), \quad (6)$$

where  $r \in \mathbb{R}^{N \times N \times 4}$  is the relative geometry relationship between grids,  $FC$  is a fully-connected layer with activation function,  $G \in \mathbb{R}^{N \times N \times d_g}$  is a high-dimensional representation of  $r$ ,  $w_g$  is a weight parameter to be learned,  $\lambda^g \in \mathbb{R}^{N \times N}$  is the relative geometry feature, and  $N = h \times w$ . The ReLU function acts as a zero trimming operation, which makes sure that we only consider the relations between grids with geometric relationships.

# RSTNET





# Adaptive Attention Module

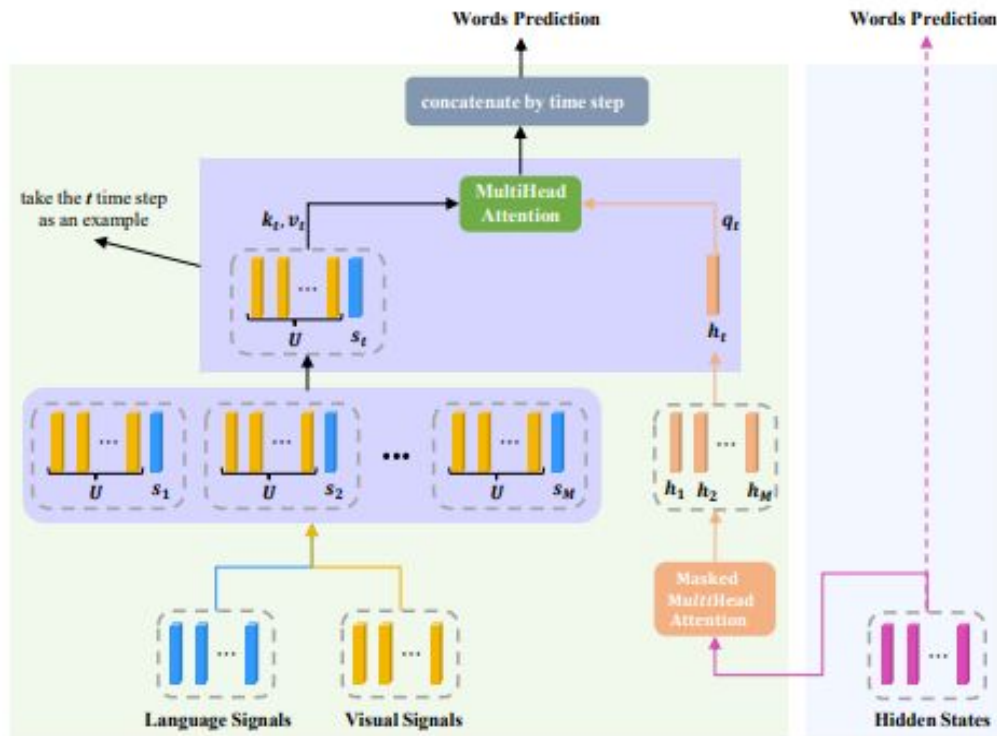


Figure 4. Illustration of our adaptive attention module. This module ensures that our model reconsiders the effect of language context before word prediction at each time step.

$$Q = UW_q, K = UW_k, V = UW_v,$$

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad \longrightarrow \quad Z_{aug} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \lambda^g\right)V,$$

$$U \leftarrow U + Z.$$

where  $U \in \mathbb{R}^{N \times d_{model}}$  is the packed visual feature vectors passing in transformer encoder layer,  $W_q, W_k, W_v$  are matrices of learnable weights, and  $d_k$  is a scaling factor.

$$q_{i,t} = h_t W_i^Q, k_{i,t} = [U; s_t] W_i^K, v_{i,t} = [U; s_t] W_i^V, \quad (17)$$

$$head_{i,t} = \text{softmax}(q_{i,t} k_{i,t}^T) v_{i,t}, \quad (18)$$

$$head_i = \text{Concate}(head_{i,1}, \dots, head_{i,M}), \quad (19)$$



















































$$att = \text{Concate}(head_1, \dots, head_h) W^O, \quad (20)$$

$$h_t = \text{Decoder}(U, W_{<t}),$$

$$W_{<t} = (w_0, w_2, \dots, w_{t-1})^T,$$

where  $q_{i,t}$  is the query vector for the  $t$ -th word word in head  $i$  of multi-head attention,  $k_{i,t}, v_{i,t}$  are the key matrix and value matrix for the  $t$  time step word in head  $i$  of multi-head attention respectively,  $head_{i,t}$  is the attention result for the  $t$ -th word word in head  $i$ ,  $head_i$  is the attention result for the word sequence in head  $i$ ,  $att$  is the attention result of multi-head attention for sequence generation.

# Visualness ( $\gamma$ )

Type	High Visualness					Low Visualness				
Word	person	boy	cat	car	flower	the	a	of	to	that
Image										
										
										
										
										

$$\alpha_{i,t} = \text{softmax}(q_{i,t}k_{i,t}^T), \alpha_{i,t} \in \mathbb{R}^{n+1}, \quad (21)$$

$$\beta_{i,t} = \alpha_{i,t}[-1], \beta_{i,t} \in \mathbb{R}, \quad (22)$$

$$\beta_t = \text{average}(\beta_{1,t}, \dots, \beta_{h,t}), \beta_t \in \mathbb{R}, \quad (23)$$

$$\gamma_t = 1 - \beta_t, \quad (24)$$

where  $\alpha_{i,t}$  is the softmax distribution of attention for the  $t$ -th word in head  $i$ ,  $\beta_{i,t}$  is language signal weight for the  $t$ -th word in head  $i$ ,  $\beta_t$  is average pooling of language signal weight over all head of multi-head attention for the  $t$ -th word.

# Training Details

Optimize using cross-entropy loss

$$L_{XE}(\theta) = - \sum_{t=1}^T \log(p_{\theta}(w_t^* | w_{1:t-1}^*)), \quad (25)$$

where  $\theta$  is the parameters of our model,  $w_{1:T}^*$  is the target ground truth sequence.

$p_{\theta}$  is prediction of the next word

Self-Critical Sequence Training(SCST)

$$L_{RL}(\theta) = -E_{w_{1:T} \sim p_{\theta}}[r(w_{1:T})],$$

where the reward  $r(\cdot)$  is the CIDEr-D score.

$w_{1:T}$  is the original caption

To minimize the negative expected reward



# Experiments and Results

Datasets used: **MSCOCO** ( 82,783 training images, 40,504 validation images, 40,775 testing images, 5DC)

Grid size =  $7 \times 7$  , Dimension of image features = 2048.

Dmodel of the transformer = 512, Number of heads = 8, Inner dimension of FFN module = 2048.

Dropout probability is 0.1

## Evaluation Metrics:

Captioning Metrics: BLEU , METEOR, ROUGR , CIDEr and SPICE

Adam optimizer

Learning rate self-critical sequence training =  $5 \times 10^{-6}$

Cross Entropy Optimization:

Cider value drops for 5 consecutive epochs, turn to SCST.

Cider value drops for 5 consecutive epochs in SCST, Training process stops

# Ablative Analysis

Table 1. Ablation study on ResNext101 grid features

GA module	AA module	B@1	B@4	M	R	C	S
X	X	80.9	38.9	29.0	58.5	131.2	22.7
X	✓	80.9	39.0	29.2	58.6	132.6	22.8
✓	X	80.9	39.0	29.2	58.7	132.1	22.8
✓	✓	<b>81.1</b>	<b>39.3</b>	<b>29.4</b>	<b>58.8</b>	<b>133.3</b>	<b>23.0</b>

Table 2. Ablation study on ResNext152 grid features

GA module	AA module	B@1	B@4	M	R	C	S
X	X	81.2	39.4	29.4	59.0	133.2	23.1
X	✓	81.0	39.2	29.6	58.9	134.3	23.3
✓	X	81.6	39.6	29.6	59.2	134.2	23.2
✓	✓	<b>81.8</b>	<b>40.1</b>	<b>29.8</b>	<b>59.5</b>	<b>135.6</b>	<b>23.3</b>

\*ResNext 101 and ResNext152 are different variants of RSTNet

# Quantitative Analysis

Table 5. Leaderboard of the published state-of-the-art image captioning models on the COCO online testing server, where B@N, M, R and C are short for BLEU@N, METEOR, ROUGE-L and CIDEr scores. All values are reported as percentage.

Model	B@1		B@2		B@3		B@4		METEOR		ROUGE-L		CIDEr-D	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST [31]	78.1	93.7	61.9	86.0	47.9	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-Down [2]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet [16]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	37.1	122.9	125.1
GCN-LSTM [43]	80.8	95.9	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
SGAE [42]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
ETA [20]	81.2	95.0	65.5	89.0	50.9	80.4	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
AoANet [13]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
$M^2$ Transformer [6]	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
X-Transformer(ResNet-101) [28]	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
X-Transformer(SENet-154) [28]	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
RSTNet(ResNext101)	81.7	<b>96.2</b>	66.5	<b>90.9</b>	51.8	<b>82.7</b>	39.7	72.5	29.3	38.7	59.2	74.2	<b>130.1</b>	<b>132.4</b>
RSTNet(ResNext152)	<b>82.1</b>	<b>96.4</b>	<b>67.0</b>	<b>91.3</b>	52.2	<b>83.0</b>	40.0	<b>73.1</b>	<b>29.6</b>	39.1	<b>59.5</b>	74.6	<b>131.9</b>	<b>134.0</b>

# Qualitative Analysis



**RSTNet** : a bus stop on the side of a street.

**Base Transformer** : a street sign on the side of a street.

**GT1**: A bus stop sign on a city street.

**GT2**: A blue bus stop sign near a highway.

**GT3**: A large blue bus sign sitting on the side of a road.



**RSTNet** : A small bird is perched on a bird feeder.

**Base Transformer** : A small bird sitting on a piece of bread.

**GT1**: A small bird is perched on an empty bird feeder.

**GT2**: A picture of a bird on a rustic looking feeder.

**GT3**: A small bird perched on the edge of a bird feeder.



**RSTNet** : a woman with a curly hair is brushing her teeth .

**Base Transformer** : a woman is smiling while holding a yellow flower.

**GT1**: A woman brushing her teeth in front of a bathroom mirror.

**GT2**: The women with curly hair is brushing her teeth.

**GT3**: A girl with blonde curly hair brushing her teeth.



**RSTNet** : a black and white cat is sleeping on a bed

**Base Transformer** : a cat curled up sleeping on a bed

**GT1**: A black and white cat sleeping on top of a bed.

**GT2**: A white and black cat is sleeping on a bed.

**GT3**: A black and white colored cat sleeping on a bed spread.



**RSTNet** : a person walking in the rain with an umbrella.

**Base Transformer** : a person walking down a city street with an umbrella.

**GT1**: A person walking in the rain on the sidewalk..

**GT2**: A person walking through the rain with an umbrella..

**GT3**: A person walking in the rain while holding an umbrella.

# Conclusion

We developed an RSTNET( Relationship Sensitive Transformer-based) model

- Grid augmented module
  - Relative spatial geometry features to compensate for the loss of spatial information
- Adaptive attention module
  - Measure the contribution of visual signals and language signal for word
- Visualness attribute
  - High Visualness:Best possible visual representation of a word
  - Low Visualness:Random visual representation of a word

**Results on MSCOCO Data set and Performance of RSTNet**

# References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering
- Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik LearnedMiller, and Xinlei Chen. In defense of grid features for visual question answering
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator



**THANK YOU**