# Selected Topics from Computer Science (Deep Learning)

## CS F 441

## Research Paper  Presentation


**BITS** Pilani
Hyderabad Campus

# Research Paper Details

**Name of the paper:**
Location-aware Graph Convolutional Networks for Video Question Answering

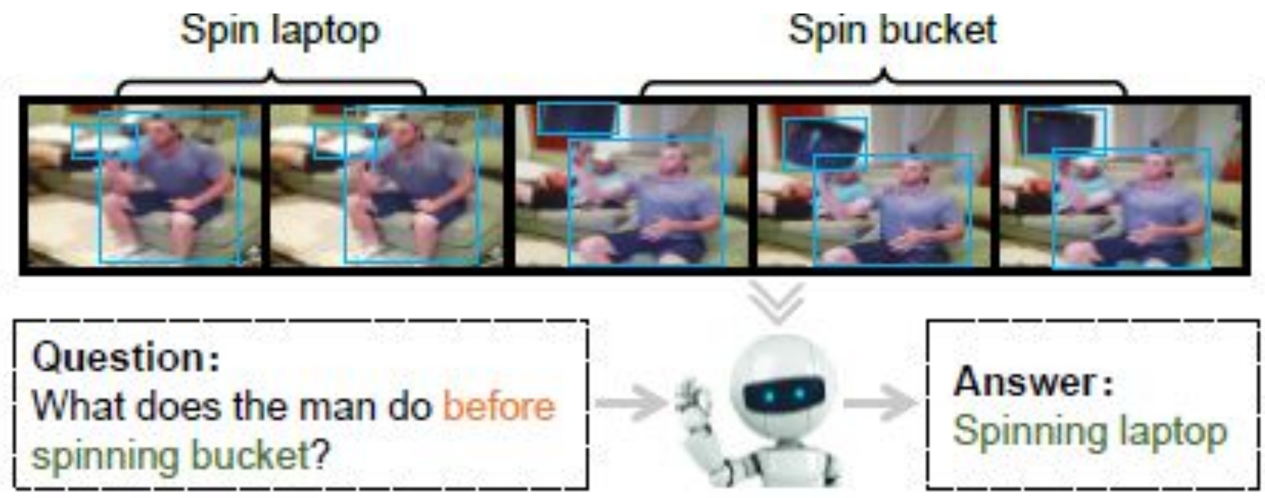
**Authors:**
Deng Huang, Peihao Chen, Runhao Zeng , Qing Du, Mingkui Tan, Chuang Gan

**Published in:**
Association for the Advancement of Artificial Intelligence, 2020. (Core 2020, A*)

# Introduction to the research problem

**What is Video Question Answering ?**

1. It is a task where a bot is required to answer questions after watching a video.
2. Video QA is difficult:
   a. Visual content is complex, more than 1000 frames
   b. Contain strong but irrelevant background content
   c. it is dynamic, “temporal” component is introduced
3. Require recognizing the actions by understanding the interaction between the objects



# Goal

**Final Goal**

Efficient mechanism of video question answering

01 To learn object interaction by excluding irrelevant background content

02 Modeling object relations with **GCNs**

03 Incorporate object location information into graph

04 Achieving state-of-the-art performance on TGIF-QA & Youtube2Text-QA

# Related work

| Title | Jang et al, 2017, TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering | Beyond RNNs: Positional Self-Attention with Co-Attention for Video Question Answering | Videos as Space-Time Region Graphs |
|---|---|---|---|
| What problem is addressed | ST-VQA model captures visual-textual association between a video and QA sentences using two dual-layer LSTMs, one for each input. | PSAC can exploit the global dependencies of question and temporal information in the video | Represent videos as graphs to mode temporal shape dynamics and functional relationships between objects |
| Feature extraction | 1. frame-level (ResNet-152)<br>2. sequence-level (C3D) | frame features (CNN) | frame features - ResNet or Inflated 3d convet (TxHxWxd)<br>Object feautres - RPN (Nxd) |
| spatio-temporal properties | dual layer LSTM as attention mechanism | Positional self attention (fram features + sinusoidal functions to encode temporal features) | Using ConvNet and GCN |
| Graph based reasoning | ✗ | ✗ | ✓ |
| Question encoding | Glove, LSTM | Q(w) = word embeddings + conv(character embeddings)<br>Q(w) = Uses highway network<br>Q(o) = positional self attention(Q(w)) | NA |
| Visual question interaction | LSTM, trains answer decoder on softmax loss | video to question attention<br>question to video attention<br>co-attention of video and question | NA |
| Comments | No object features used<br>different attention mechanisms (Spatial and Temporal) shows the effectiveness of temporal attention mechanism, achieving the best performance | No object features used<br>Replaced LSTM with positional self attention | Uses graphs to model interaction and sequence of objects in a video |

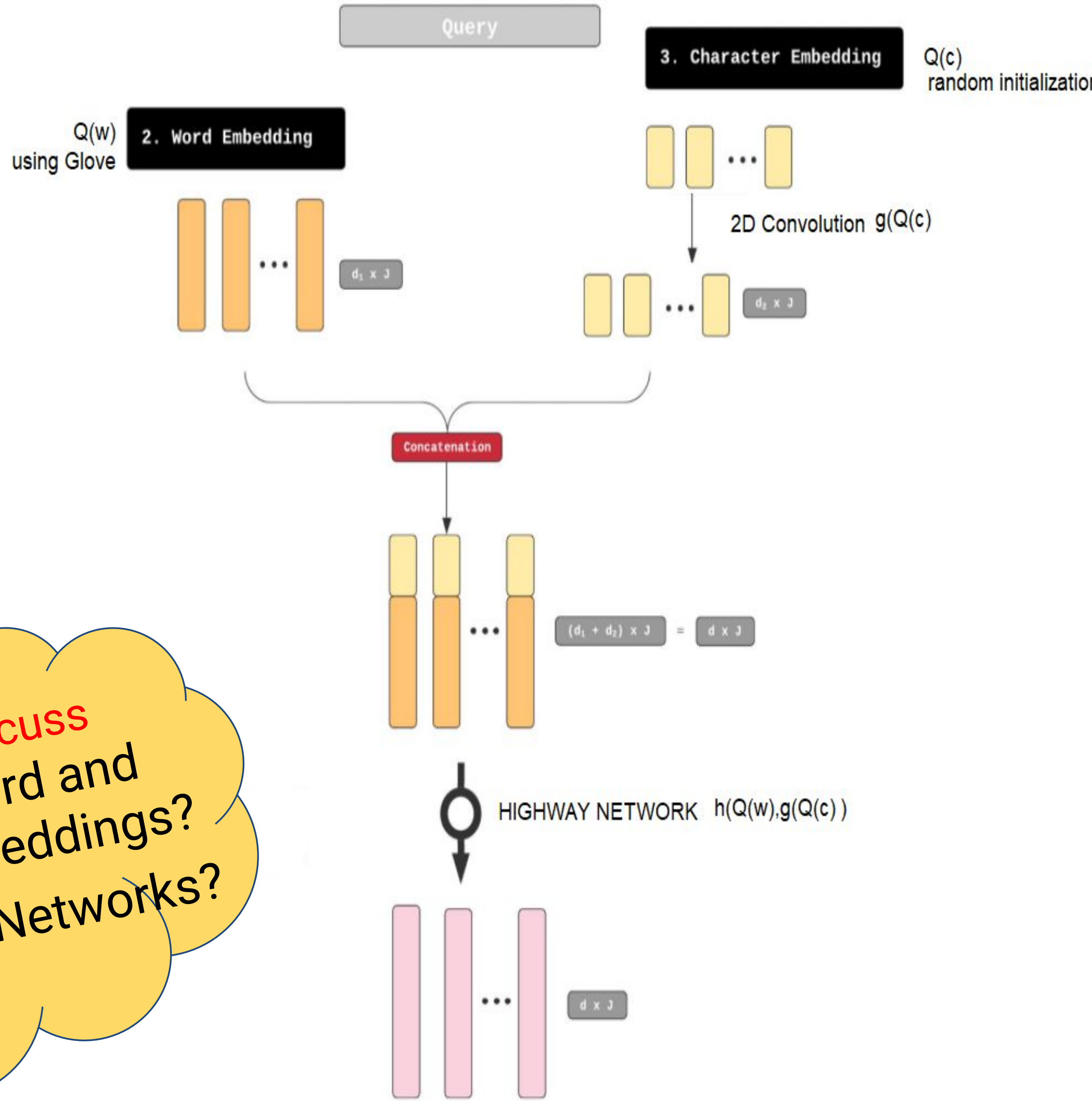
# Proposed model

# Question Encoder

QUESTION ENCODER

VISUAL ENCODER

VISUAL QUESTION INTERACTION




Points to discuss

Why both word and character embeddings?

Why Highway Networks?

# Question Encoder

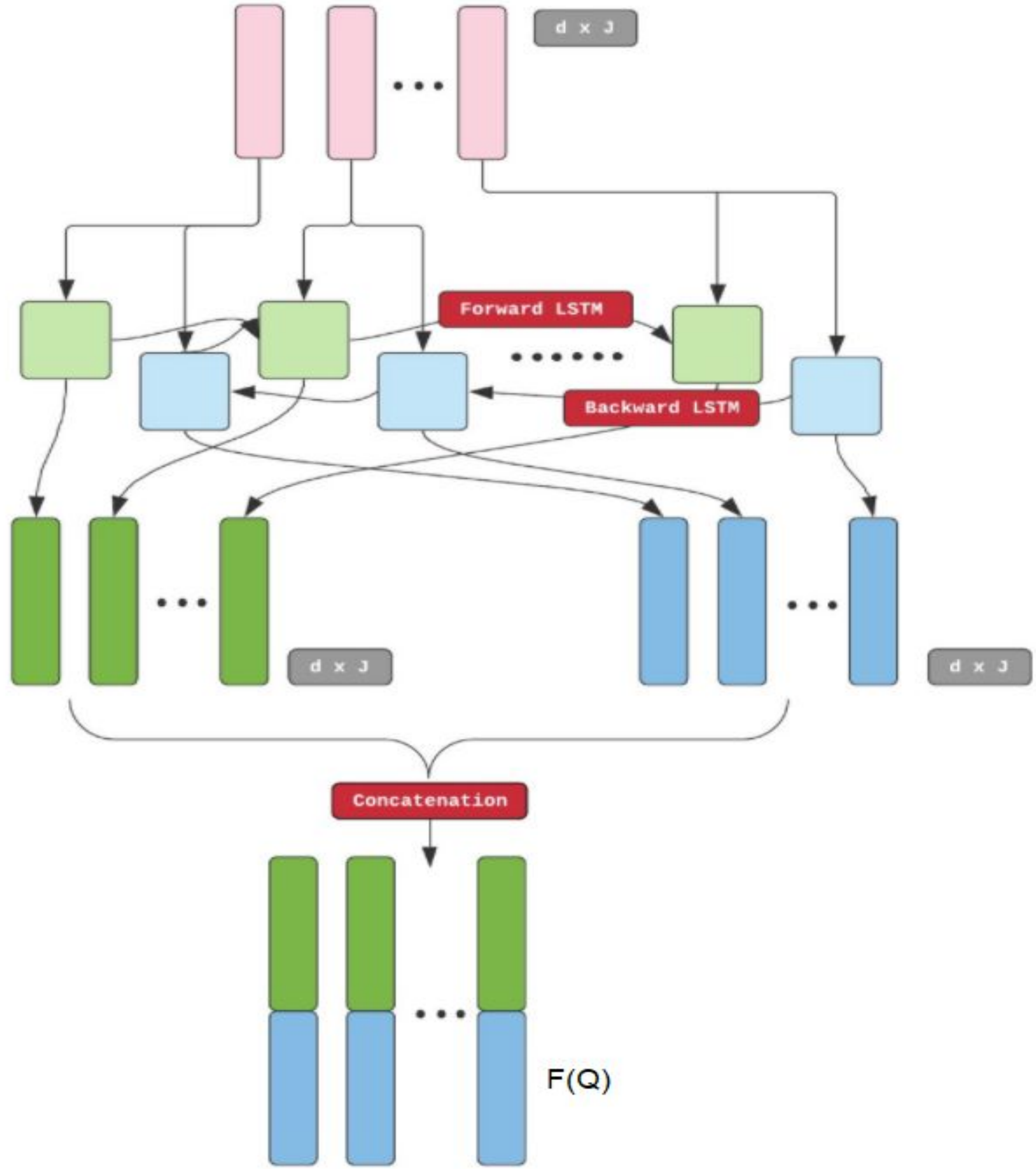
QUESTION ENCODER

VISUAL ENCODER

VISUAL QUESTION INTERACTION

Points to discuss

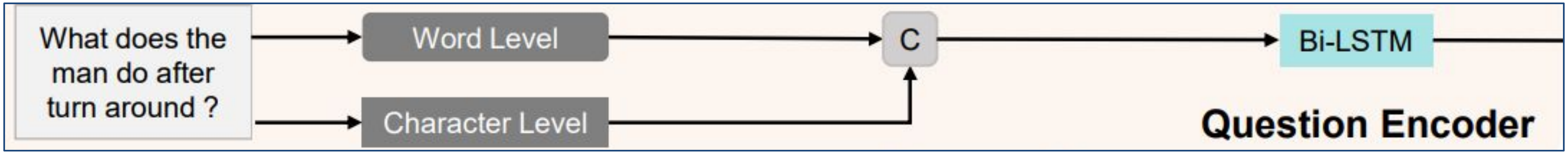
Why Bi-LSTM?



# Proposed model

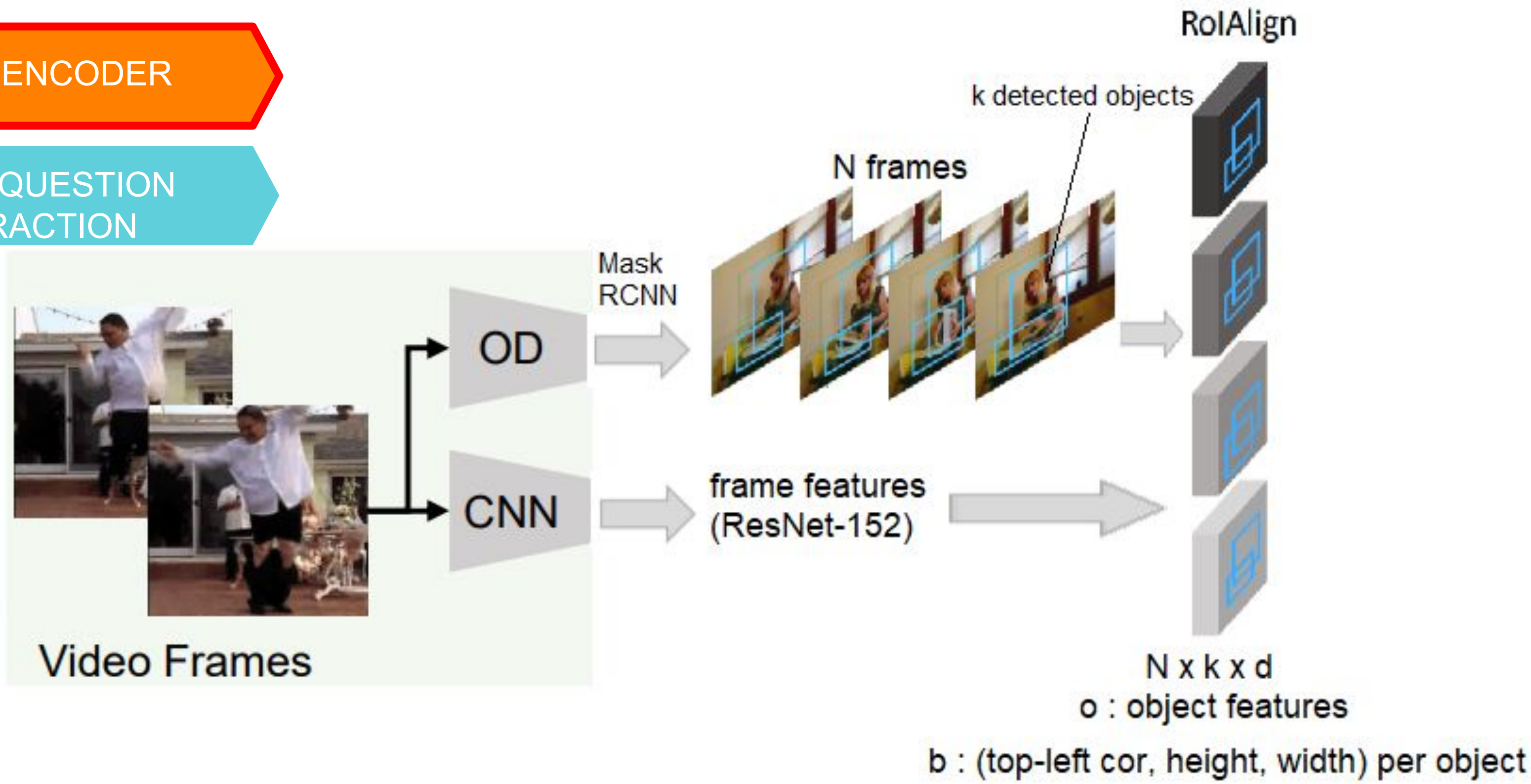
The model uses two streams of data and is divided into three sub parts:

1. **Question Encoder**
   1. The first stream is input to the question encoder to model question for video QA.
   2. $\mathbf{Q}^w$, word embedding is obtained by initializing the function with a pre-trained 300 dimension GloVe
   3. $\mathbf{Q}^c$, character embedding function is initialized randomly
   4. Given $\mathbf{Q}^w$, $\mathbf{Q}^c$ question embeddings are given to a highway network *h(.,.)* $\mathbf{Q} = h(\mathbf{Q}^w, g(\mathbf{Q}^c)),$ where g(.) consist of a 2d convolutional layer
   5. **Bi-LSTM** is used to encode the question **Q,** to obtain question feature $\mathbf{F}^{\mathbf{Q}}$

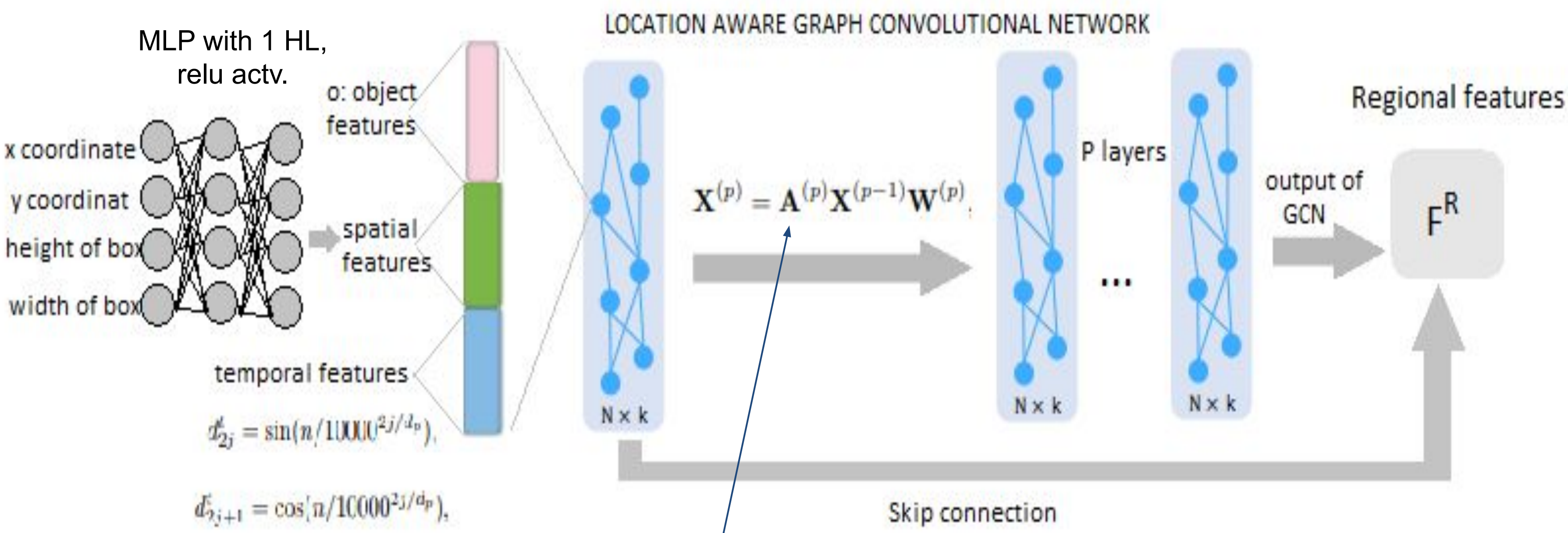


# Visual Encoder - Object features

- QUESTION ENCODER
- **VISUAL ENCODER**
- VISUAL QUESTION INTERACTION

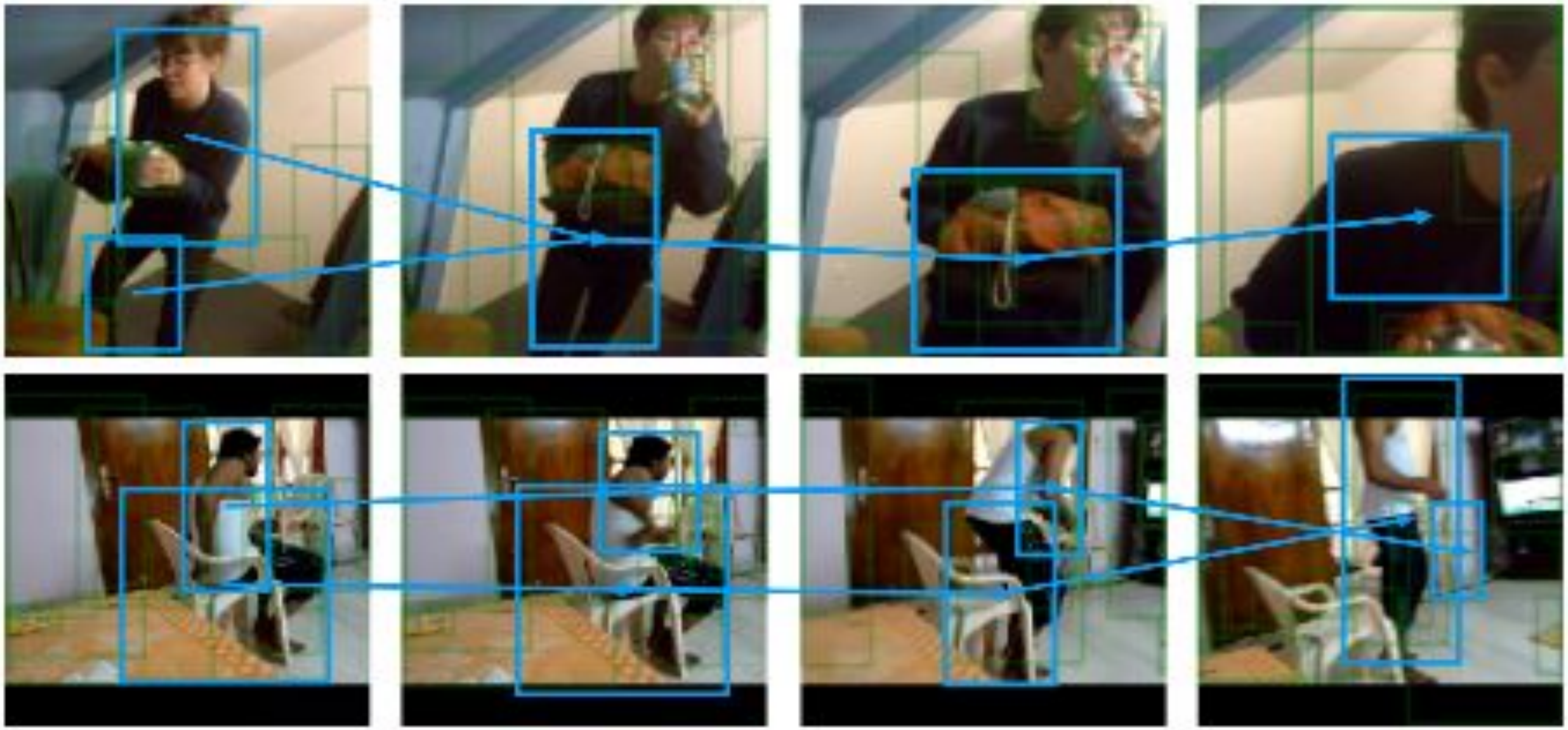


# Visual Encoder - LGCN




- Same object in different states in different video frames
- highly correlated for recognizing the actions
- based on 2 projection matrices - t to t+1 and t+1 to t

# GCN : A SPATIO -TEMPORAL INTUITION



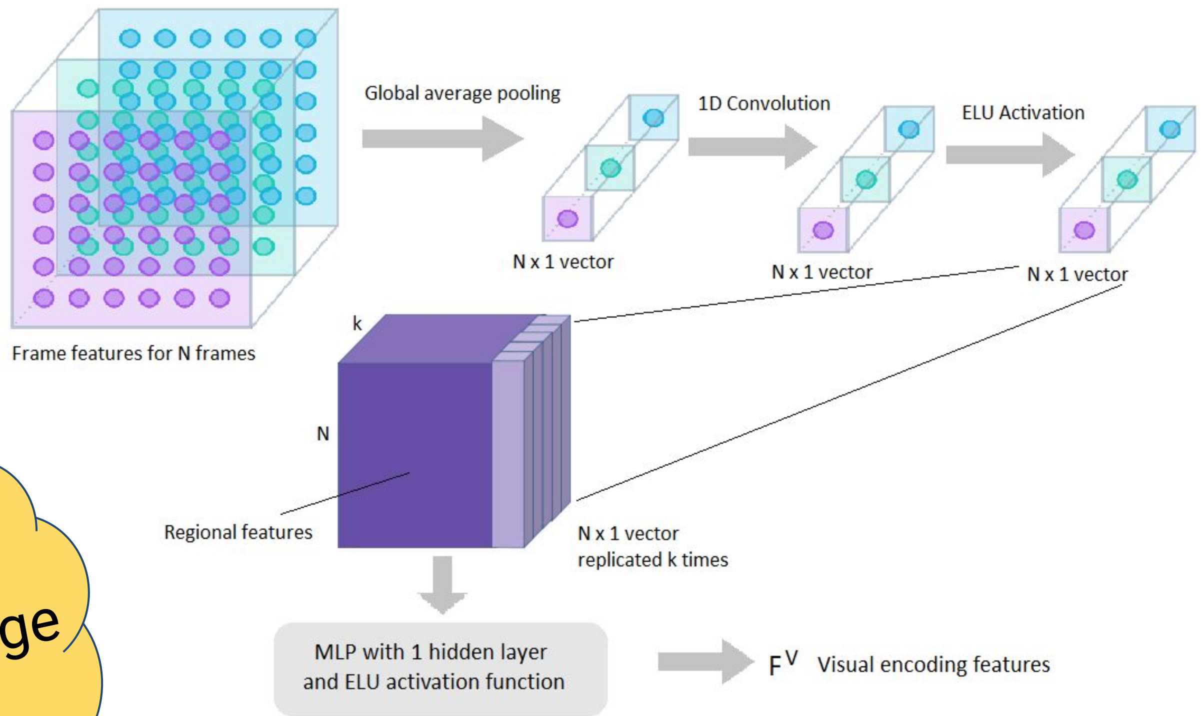
Spatial-Temporal Graph plotted across the neighbouring frames for Graph convolutional Networks. **Highly overlapping object proposals across neighboring frames are linked by directed edge**. some example trajectories with blue boxes and the direction shows the arrow of time.

# Visual Encoder - Combining LGCN and Global features

QUESTION ENCODER

VISUAL ENCODER

VISUAL QUESTION INTERACTION




Points to discuss

Why Global average pooling?

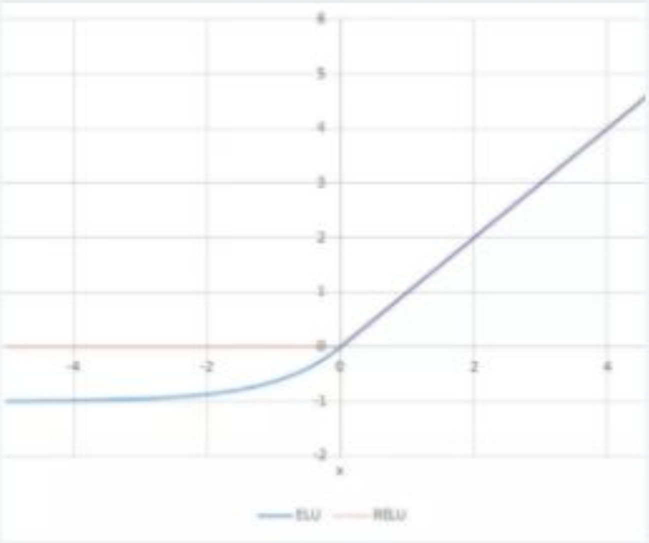
Use of ELU...

# Points to Discuss

- Global average pooling
- Why ELU is used?

Function

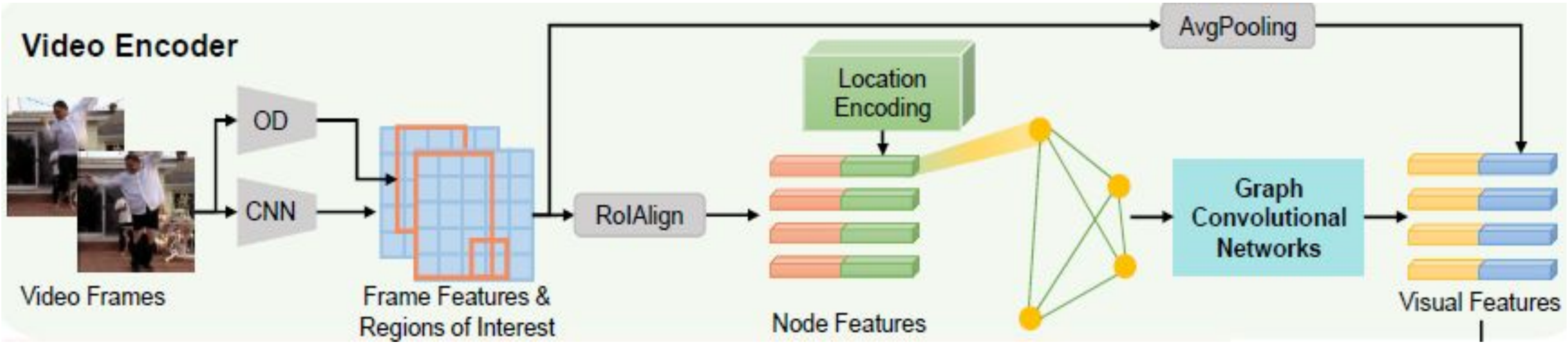
$$R(z) = \begin{Bmatrix} z & z > 0 \\ \alpha.(e^z - 1) & z <= 0 \end{Bmatrix}$$



# Visual Encoder - Object features

The second stream(N frames) is input to the visual encoder to model video contents via object interaction for video QA.

1. **Frame features** - feature extractor algorithm
2. **K bounding boxes** - Mask RCNN
3. **Object features o** - RoIAlign followed by *1 FC with ELU activation function,* further given to location aware GCN (output as regional features $F^R$)
4. **Context Information**
    a. **Global Features $F^G$** - global average pooling on the frame features
    b. *1D convolutional layer and an ELU activation function* to merge the information from neighbor frames.
    c. replicate K times
5. **Visual features $F^V$** - *MLP with 1 hidden and ELU activation function* (input $F^R$, $F^G$ )

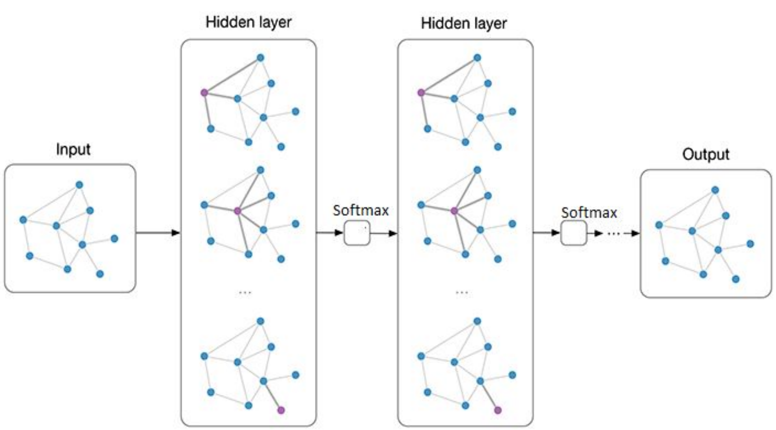


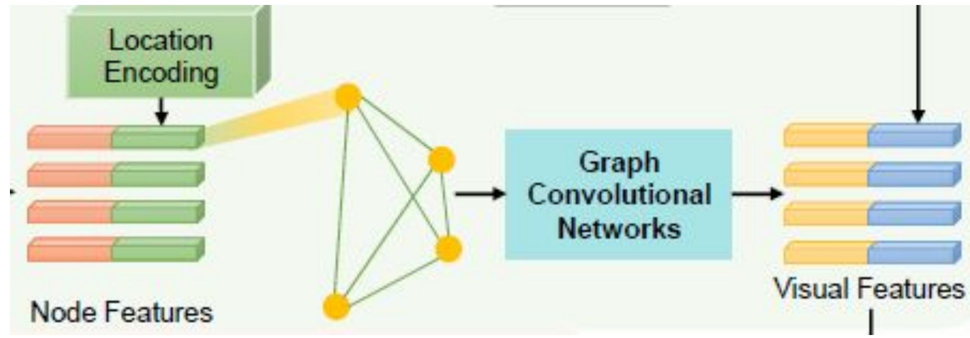
# Visual Encoder - L-GCN

<u>Location encoding</u>

1. Spatial features $d^s$ is calculated based on inputs
    a. top-left coordinates
    b. height
    c. weight;

    using MLP with 2 FC and ReLU activation.
1. Temporal features $d^t$ are represented as sinusoidal values
2. $v = [o; d^s; d^t]$




<u>GCN - regional features</u>

v is taken as input to GCN with P layer graph convolutions. For a layer p and hidden layer X, formula can be given as

$$\mathbf{X}^{(p)} = \mathbf{A}^{(p)}\mathbf{X}^{(p-1)}\mathbf{W}^{(p)}$$

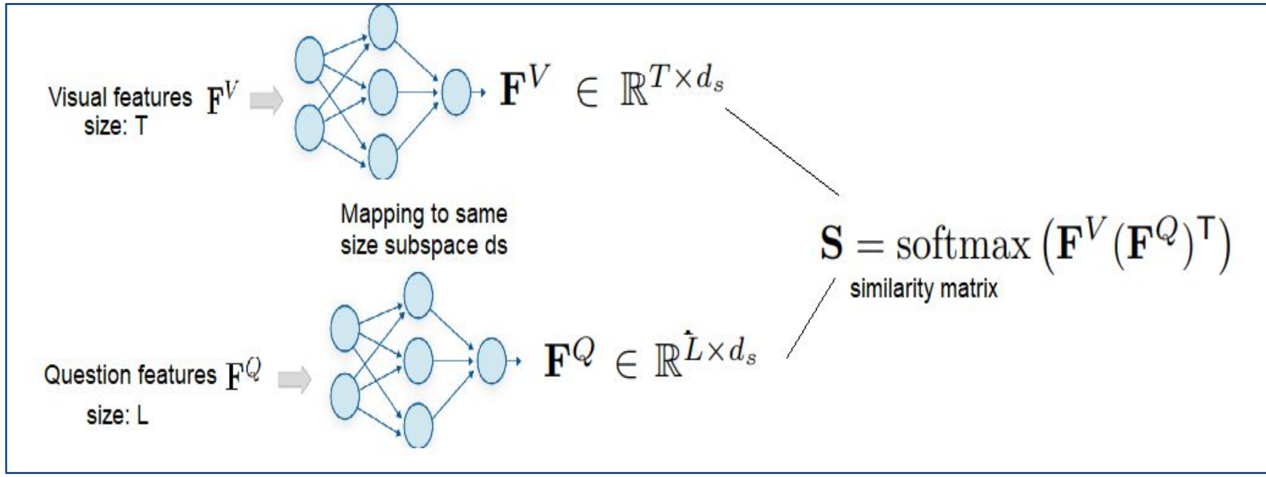
where A is the adjacency matrix

$$\mathbf{A}^{(p)} = \mathrm{softmax}\left(\mathbf{X}^{(p-1)}\mathbf{W}_1 \cdot (\mathbf{X}^{(p-1)}\mathbf{W}_2)^{\mathsf{T}}\right)$$

Final output is summation of $P^{th}$ layer and input termed as $F^R$

# Visual question interaction

- QUESTION ENCODER
- VISUAL ENCODER
- **VISUAL QUESTION INTERACTION**

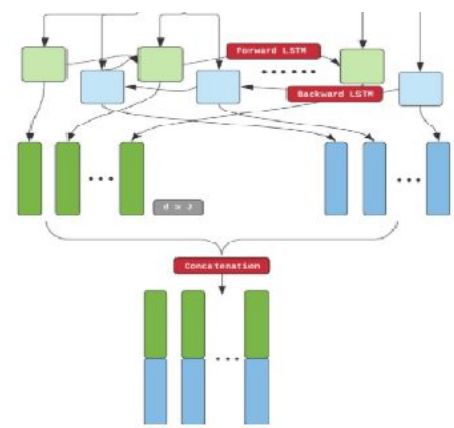



weighted question features

$$\widetilde{\mathbf{F}}^Q = \mathbf{S} \times \mathbf{F}^Q$$

cross modality

$$\mathbf{F}^C = [\mathbf{F}^V, \widetilde{\mathbf{F}}^Q, \mathbf{F}^V \odot \widetilde{\mathbf{F}}^Q]$$




Bi-LSTM to yeild final representation

# Answer Reasoning

QUESTION ENCODER

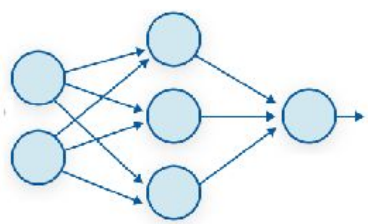
VISUAL ENCODER

VISUAL QUESTION INTERACTION

To predict answers for Multiple choice and Open-ended questions

U choices → $\mathbf{F}^A$ → $[\mathbf{F}^V, \widetilde{\mathbf{F}}^Q, \widetilde{\mathbf{F}}^A, \mathbf{F}^V \odot \widetilde{\mathbf{F}}^Q, \mathbf{F}^V \odot \widetilde{\mathbf{F}}^A]$

answer feature vector for each independent choice U, in total



$$A = \{a_1, \dots, a_U\}$$

predict scores for each answer from U

process scores using softmax loss

$$L_M = -\sum_{i=1}^{U} y_i \log\left(\frac{e^{a_i}}{\sum_{j=1}^{U} e^{a_j}}\right)$$

cross-entropy

The choice with the highest score is taken as the prediction

**Multiple-choice question**: there exist U choices and the model is required to choose the correct one

**Open-ended question**: model is required to choose a correct word as answer from the predefined answer set of C candidate words in tota innovate achieve lead

# Answer Reasoning

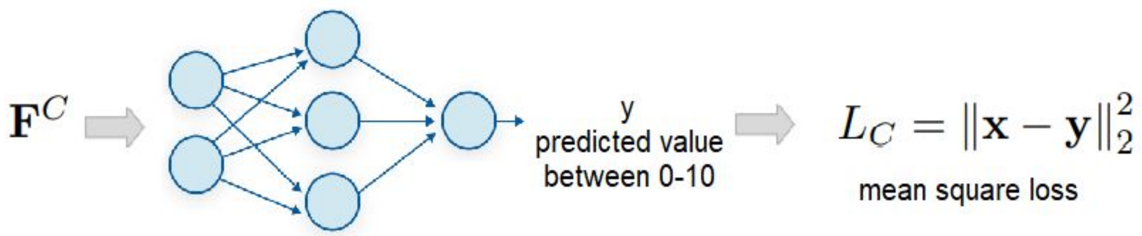
QUESTION ENCODER

VISUAL ENCODER

VISUAL QUESTION INTERACTION

To predict answers for counting questions

**Counting question**: The model is required to predict a number ranging from 0 to 1



# Visual-question Interaction Module

Combine FV and FQ to predict answer, attention mechanism is used to learn a cross modality representation

1. Similarity matrix $\mathbf{S} = \text{softmax}\left(\mathbf{F}^V (\mathbf{F}^Q)^\top\right).$
2. Weighted question Features - dot product of $F^Q$ and S
3. Cross modality representation $\mathbf{F}^C = [\mathbf{F}^V, \widetilde{\mathbf{F}}^Q, \mathbf{F}^V \odot \widetilde{\mathbf{F}}^Q]$
4. Final answer representation - *Bi-LSTM,* max pooling layer across dimension T.

*How to predict answers for different question types given cross modality features $F^C$*

| Multiple Choice Questions | Open-Ended questions | Counting Questions |
|---|---|---|
| • exist U choices with independent answer features $F^A$ and each is interacted with visual features<br>• weighted answer feature $\widetilde{\mathbf{F}}^A$<br>• $\mathbf{F}^C = [\mathbf{F}^V, \widetilde{\mathbf{F}}^Q, \widetilde{\mathbf{F}}^A, \mathbf{F}^V \odot \widetilde{\mathbf{F}}^Q, \mathbf{F}^V \odot \widetilde{\mathbf{F}}^A]$<br>• Predict scores using **1 FC layer on U with softmax function.** | • Choose a correct word as answer from the predefined answer set of C candidate words in total.<br>• Predict the scores using **1 FC layer together with a softmax layer.** | • Predict a number ranging from 0 to 10.<br>• **FC layer upon $F^C$ to predict the number.**<br>• y is ground truth<br>• Prediction is rounded to the nearest integer between 0 to 10. |

# Experiment and Implementation Details

Dataset used in experiment:

1) TGIF-QA (Jang et al. 2017)
2) Youtube2Text-QA (Ye et al. 2017)
3) MSVD-QA (Xu et al. 2017)

Implementation Details:

1) Evaluation metrics: MSE and accuracy.
2) Training: All words are in small caps. Each word is transformed to a 300-dimension vector with a pre-trained GloVe model. Mask R-CNN is used as an object detector. Number of layers in GCN is 2. Adam optimizer is used to train the network with learning rate 1e-4. Batch sizes are 64 for MCQs and 128 for open ended tasks.

Table 1: Statistics of three video QA datasets. #MC denotes the number of options for multiple-choice questions.

| Dataset | Vocab. size | #Video | #Question | Answer size | #MC | Feature type | #Sampled frame |
|---|---|---|---|---|---|---|---|
| TGIF-QA | 8,000 | 71,741 | 165,165 | 1,746 | 5 | ResNet-152 | 35 |
| Youtube2Text-QA | 6,500 | 1,970 | 99,429 | 1,000 | 4 | ResNet-101+C3D | 40 |
| MSVD-QA | 4,000 | 1,970 | 50,505 | 1,000 | NA | VGG+C3D | 20 |

# Experiment and Implementation Details

Table 2: Comparisons with state-of-the-arts on TGIF-QA dataset. R, C and F denote features extracted by ResNet, C3D and Optical Flow, respectively.

| Model | Action | Trans. | FrameQA | Count (MSE) |
|---|---|---|---|---|
| ST-VQA(R+C) | 60.8 | 67.1 | 49.3 | 4.28 |
| Co-Mem(R+F) | 68.2 | 74.3 | 51.5 | 4.10 |
| PSAC(R) | 70.4 | 76.9 | 55.7 | 4.27 |
| HME(R+C) | 73.9 | 77.8 | 53.8 | 4.02 |
| Ours(R) | **74.3** | **81.1** | **56.3** | **3.95** |

Table 3: Comparisons with state-of-the-art methods on Youtube2Text-QA.

| Task | Method | What | Who | Other | All |
|---|---|---|---|---|---|
| Multiple-Choice | r-ANL | 63.3 | 36.4 | 84.5 | 52.0 |
| | HME | 83.1 | 77.8 | **86.6** | 80.8 |
| | Ours | **86.0** | **81.5** | 80.6 | **83.9** |
| Open-Ended | r-ANL | 21.6 | 29.4 | **80.4** | 26.2 |
| | HME | **29.2** | 28.7 | 77.3 | 30.1 |
| | Ours | 24.5 | **53.2** | 70.4 | **38.0** |

Table 4: Comparisons with state-of-the-arts on MSVD-QA.

| Model | ST-VQA | Co-Mem | AMU | HME | Ours |
|---|---|---|---|---|---|
| Acc | 31.3 | 31.7 | 32.0 | 33.7 | **34.3** |

# Ablation Study

**Impact of each component:**

Models are divided into different categories:

1) Baseline – Uses only global frame features to generate visual features
2) OF – Include object features
3) GCNs – Includes Graph Convolution Networks
4) Loc – Includes location features
5) FC – GCN replaced by 2 fully connected layers
6) LSTM – GCN replaced by 2-layer LSTM
7) Loc_T – Location features with temporal location information only
8) Loc_S – Location features with spatial location information only

Table 6: Ablation study on #GCNs layers on TGIF-QA.

| #GCNs layers | Action | Trans. | FrameQA | Count |
|---|---|---|---|---|
| 1 | 74.24 | 81.02 | 55.97 | 4.16 |
| 2 | **74.32** | 81.13 | **56.32** | **3.95** |
| 3 | **74.32** | **81.58** | 56.23 | 4.16 |
| 4 | 73.97 | 80.86 | 56.01 | 4.10 |

Table 5: Performance comparisons of different variants on TGIF-QA. "**OF**" and "**Loc**" denote object and location features, respectively.

| Model | Action | Trans. | FrameQA | Count |
|---|---|---|---|---|
| **baseline** | 70.58 | 79.59 | 55.37 | 4.33 |
| **baseline+OF** | 72.82 | 80.10 | 55.79 | 4.24 |
| **baseline+OF+GCNs** | 74.10 | 80.39 | 56.10 | 4.15 |
| **baseline+OF+GCNs+Loc** | **74.32** | **81.13** | **56.32** | **3.95** |
| **baseline+OF+FC** | 72.96 | 80.18 | 55.94 | 4.22 |
| **baseline+OF+LSTM** | 72.65 | 80.07 | 55.49 | 4.25 |
| **baseline+OF+GCNs+Loc_T** | 73.75 | 80.97 | 55.54 | 4.17 |
| **baseline+OF+GCNs+Loc_S** | 73.58 | 80.89 | 56.07 | 4.12 |

# Conclusion

In this paper, a location-aware graph is proposed to model the relationships between detected objects for video QA task. Compared with existing spatial-temporal attention mechanism, L-GCN is able to explicitly get rid of the influences from irrelevant background content. Moreover, the network is aware of the spatial and temporal location of events, which is important for predicting correct answer. This method outperforms state-of-the-art techniques on three benchmark datasets.

Comparison with state-of-the-art results:

1) Results on TGIF-QA : Compared with models like ST-VQA, Co-Mem, PSAC and HME. L-GCN outperforms HME, ST-VQA and Co-Mem by a large margin.
2) Results on Youtube2Text-QA : Compared with models like HME and r-ANL. The L-GCN performs with an overall better accuracy.
3) Results on MSVD-QA : Compared with models like ST-VQA, Co-Mem, AMU and HME. L-GCN performs well in overall accuracy

**<u>Reason for better performance:</u>**

The L-GCN performs better because it is leveraging an object graph to capture the object-object interaction and perform reasoning. This gives it a much higher accuracy than other models.

***Thank You!!***