Sequence to Sequence
Video to Text

Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue
Raymond Mooney, Trevor Darrell, Kate Saenko

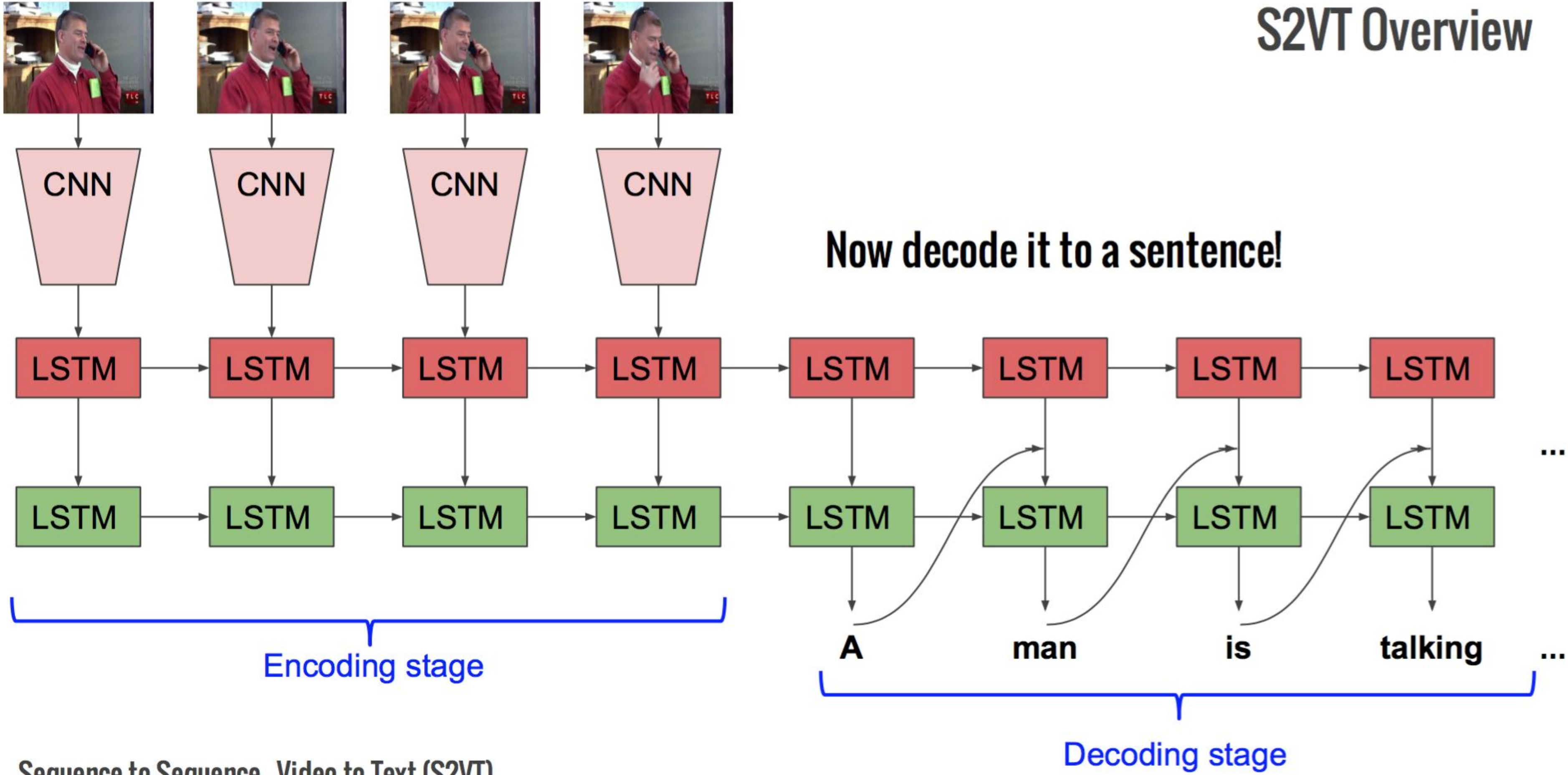
Outline

- Objective
- Experimental Setup
- Current model.
- A Simple Extension.
- How is information distributed within the video ?
- Does model capture temporal information ?
- Conclusions & Future Work

Objective

Generate video descriptions.

S2VT Overview



Sequence to Sequence - Video to Text (S2VT)
S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko

Experimental Setup

Code: Forked from author’s github account

Frame Sampling: 1 in 10 (unless otherwise mentioned)

Network Architecture: VGG CNN + 2 layer LSTM

Dataset : MSVD Youtube dataset (Avg Length 10.2 s, #sentences per video = 41)

Vocabulary : MSVD + MPII-MD + MVAD

Performance Metric: METEOR

Evaluation Tool: coco_evaluation

Forward Model

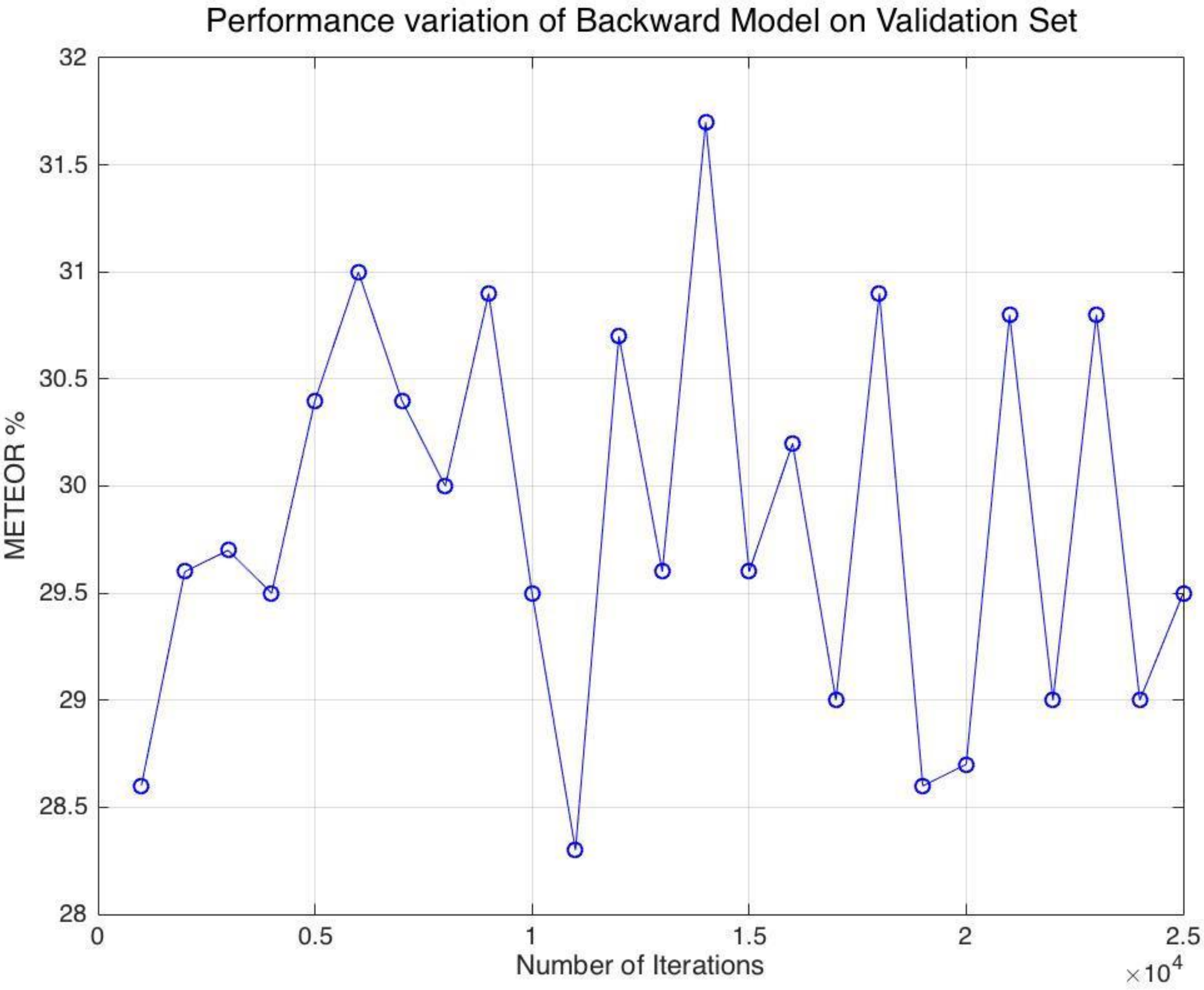
- Able to learn abstract attributes like young etc to reasonable extent.
- Able to capture main content of video in most cases.

PROBLEMS:

- Long sentences repeat words multiple times leading to lower quality sentences
    - The boys are playing with a **group of** a **group of** a **group of** people is sitting on a **group of** a **group of** people are watching a gym
    - A woman is cutting a **piece** of a **piece** of a **pair** of a **pair** of a **pair.**
    - A man is cutting a **large** of a **large large large large** floor.

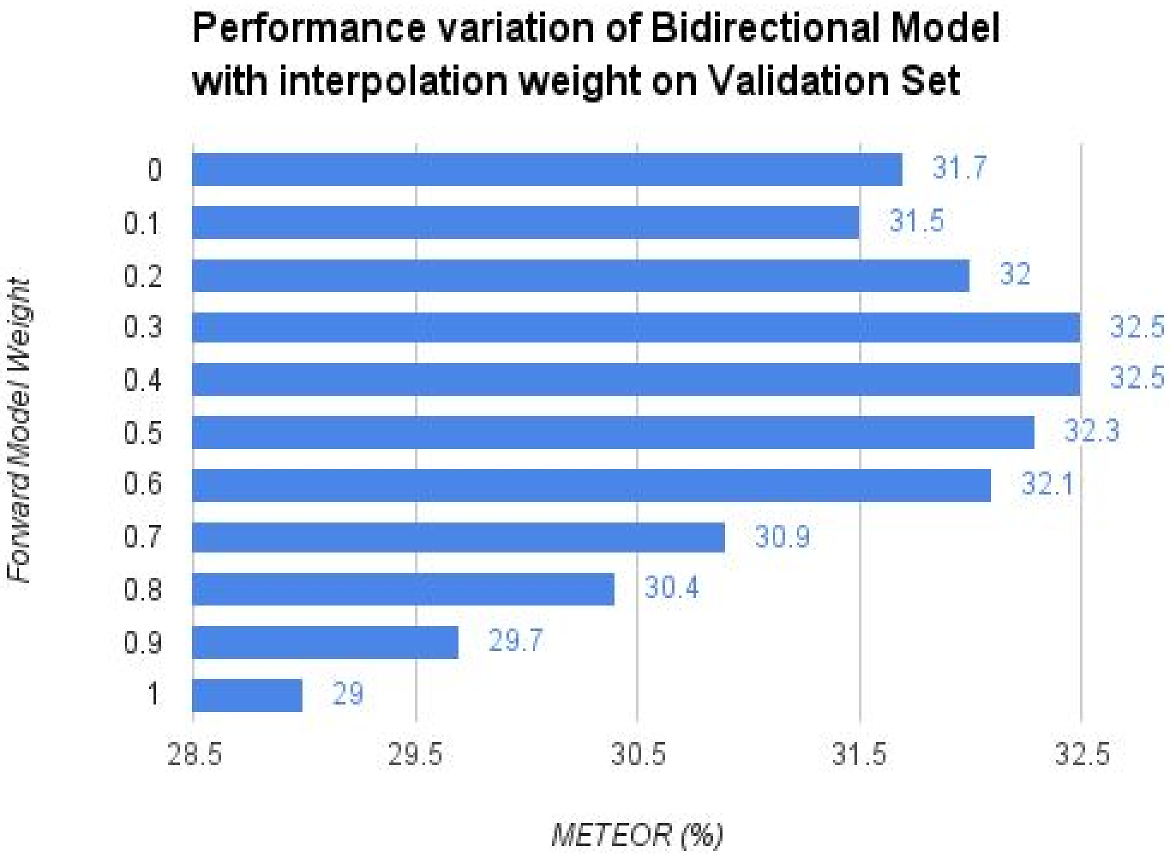
Backward Model

- Process frames in reverse order !!
- Seems to perform better than forward model on validation set but almost similar performance on test set.
- How to choose best backward model ?


Performance variation of Backward Model on Validation Set
METEOR %
32
31.5
31
30.5
30
29.5
29
28.5
28
0
0.5
1
1.5
2
2.5
Number of Iterations
×10^4

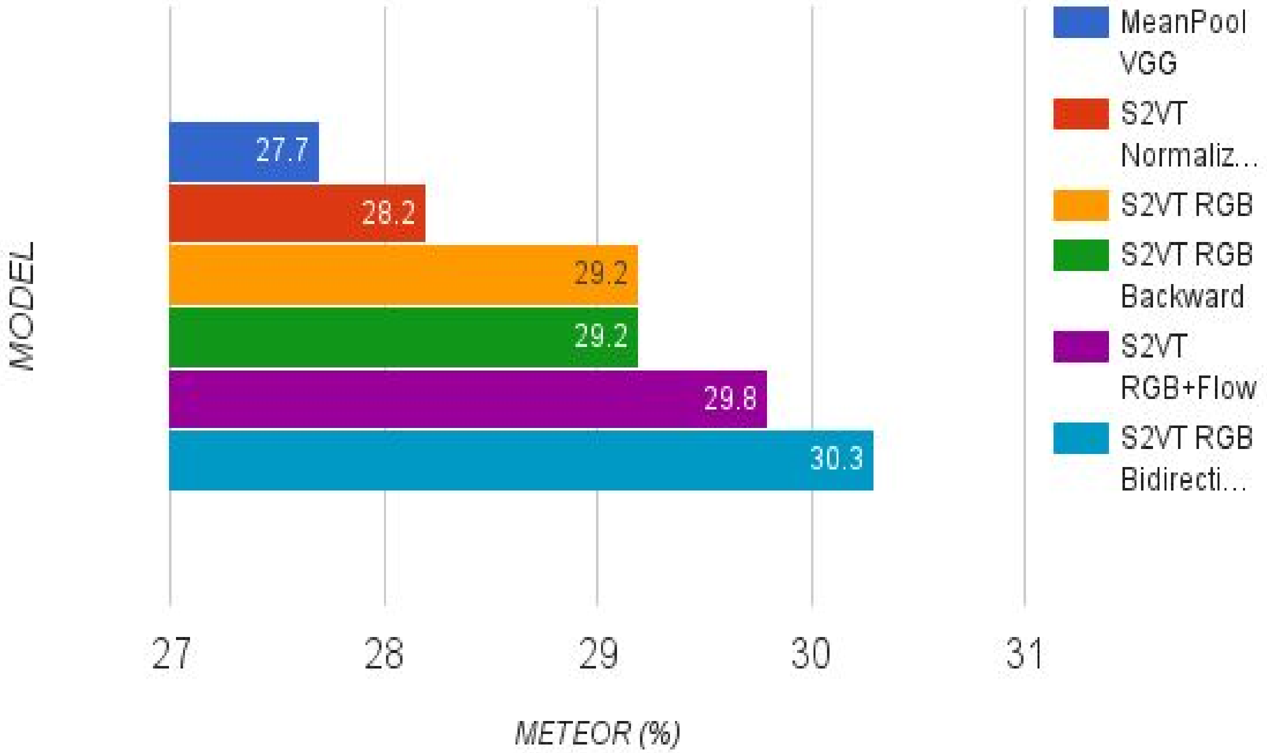
Bidirectional Model

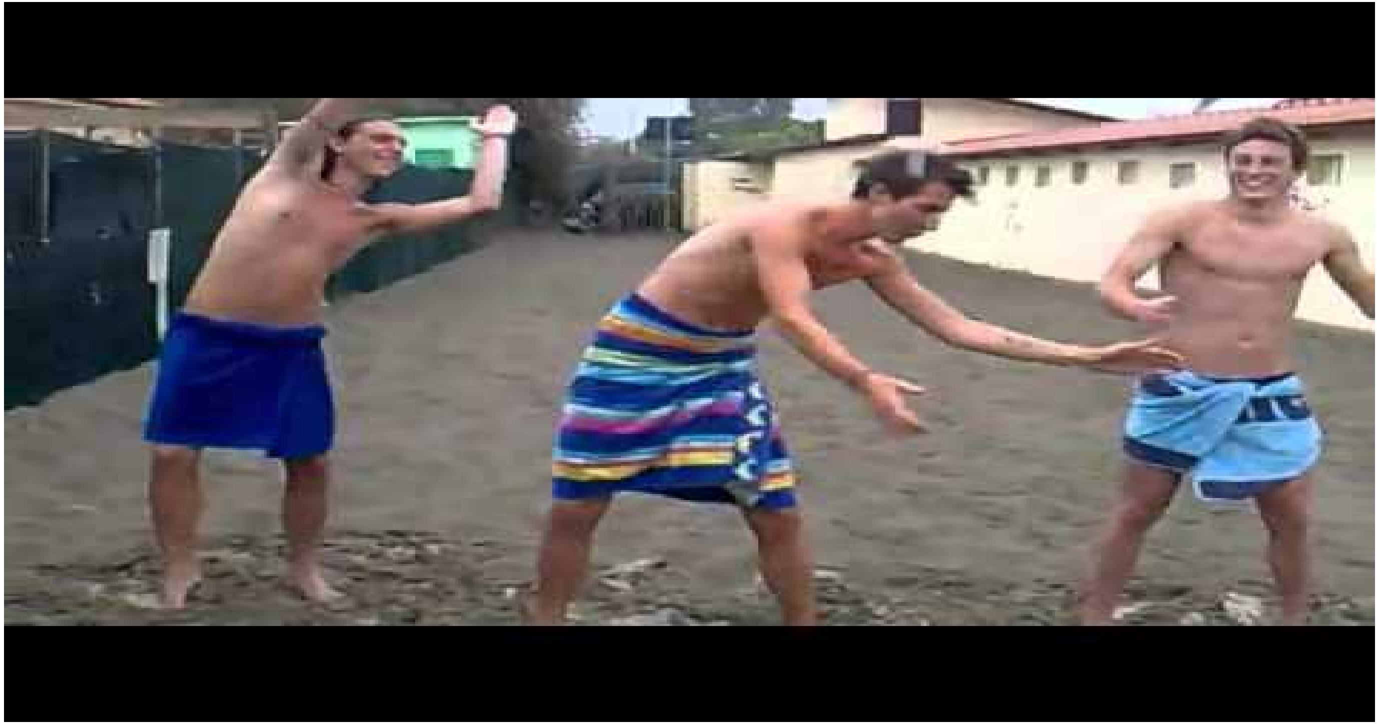
- Motivated from Bidirectional N gram models used for Language Modelling in NLP
- Combine forward and backward models.
    - How do we select forward and backward model ?
    - Combining strategy ?
    - How are weights selected ?



Performance variation of Bidirectional Model with interpolation weight on Validation Set

Performance Comparison of all models





Your description ??

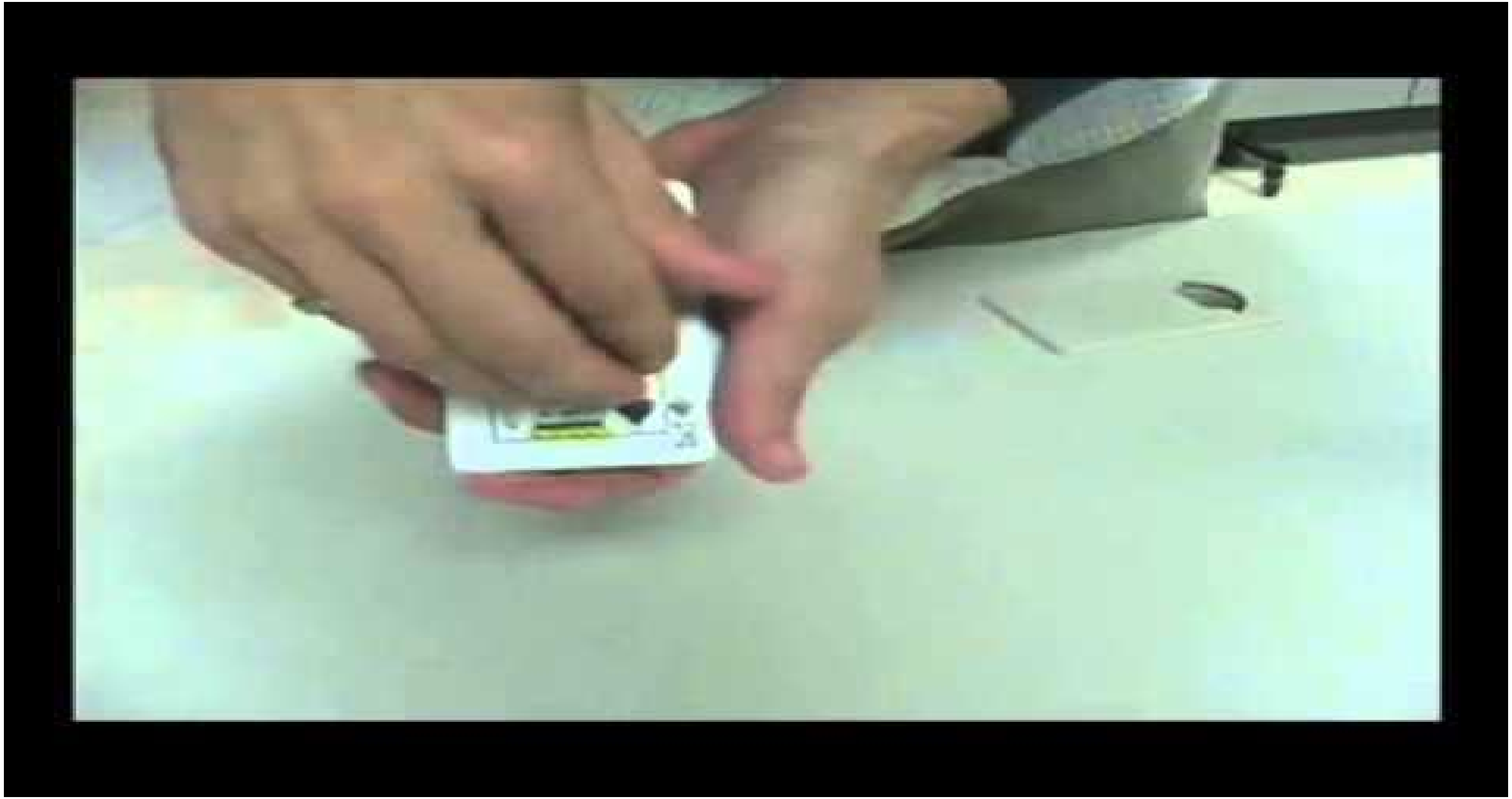
FORWARD:
The boys are playing with a **group of a group of a group of** people is sitting on a **group of a group of** people are watching a gym !!

BACKWARD: Two boys are dancing.

BIDIRECTIONAL: The boys are playing.

LABEL: Three men are dancing in beach towels.

This eg shows utility of Bidirectional Model.



Your description ??

FORWARD: A man is using a piece of a sharp.

BACKWARD: A person is cutting a piece of a brush.

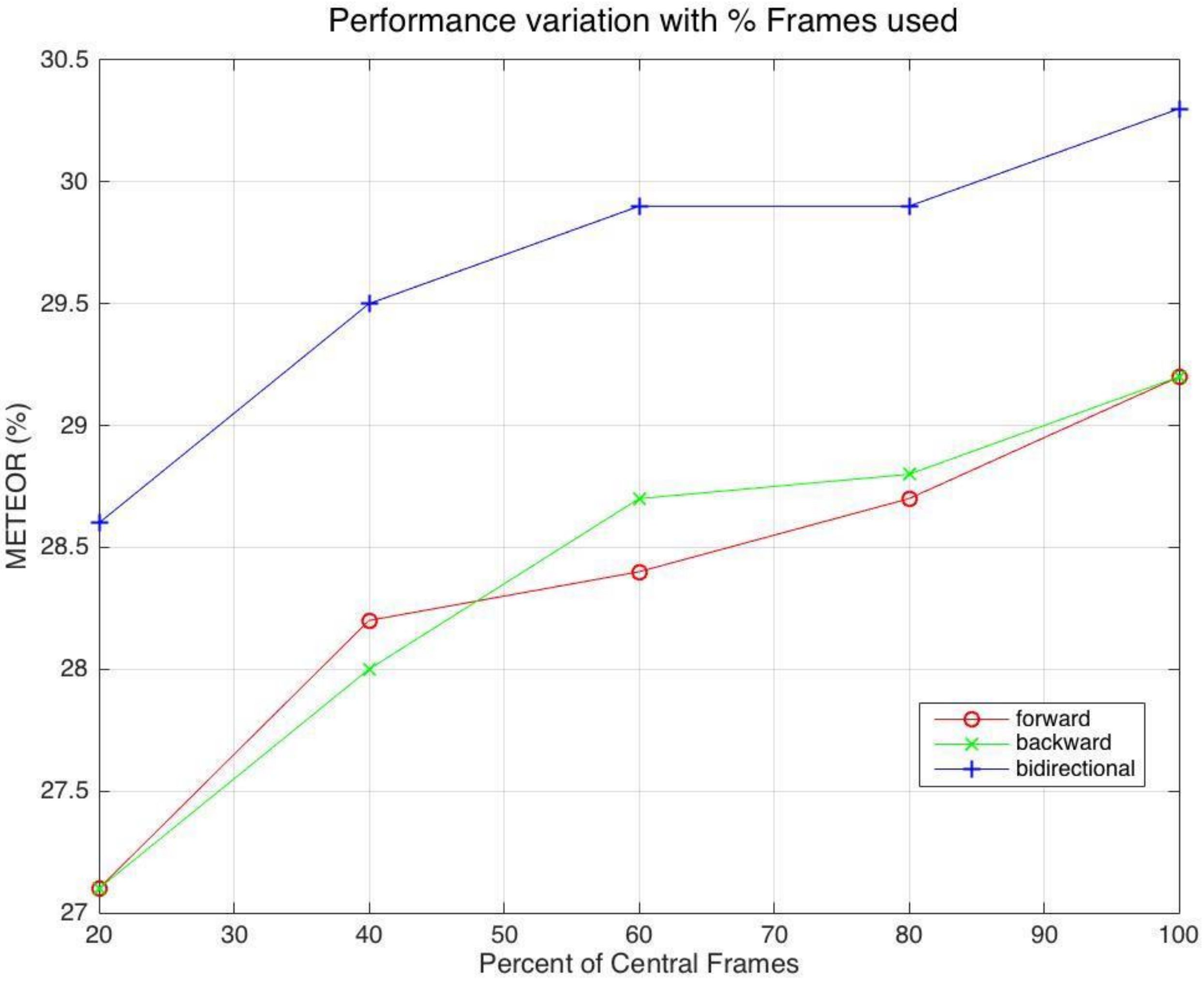
BIDIRECTIONAL: A man is cutting a piece of a brush.

LABEL: A person is performing some card tricks.

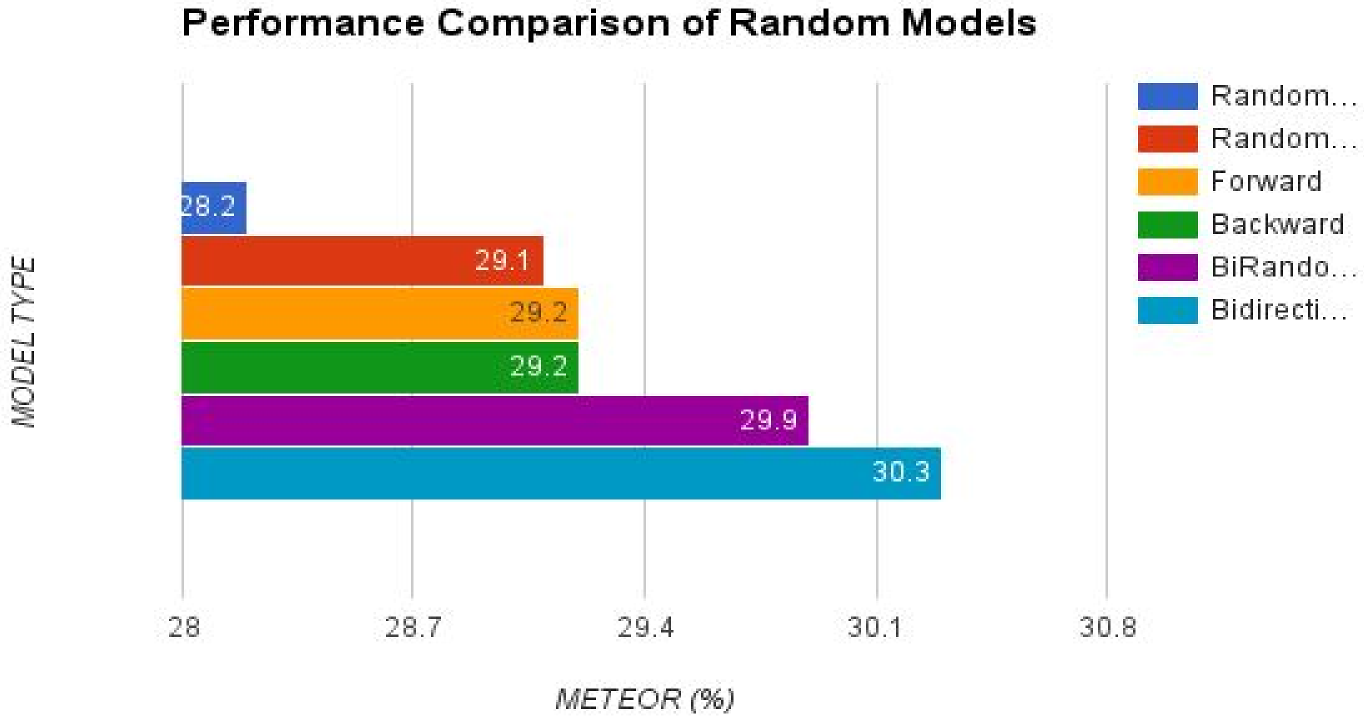
All Fail :(

How is information distributed within video ?

Conjecture: Central part of video contains more relevant information than frames at beginning and end for most videos


Performance variation with % Frames used
METEOR (%)
Percent of Central Frames
30.5
30
29.5
29
28.5
28
27.5
27
20
30
40
50
60
70
80
90
100
forward
backward
bidirectional

Does Model Capture Temporal Information ?



Conclusions

- Bidirectional model is more powerful than forward or backward model.
- Frames at start and end contain less information.

Future Work

- Try combining bidirectional with optical flow model.
- Try using gaussian sampling centred on video’s centre
- Is it more suitable for specific kinds of videos ? Like generating sports commentary ?

References

Sequence to Sequence Video to Text - Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko

Thank You :)