## Motivation

- Significant progress in image-based VQA with various datasets
- A few datasets use movie as data source for video VQA

### MovieQA

[Neelakantan et al., CVPR 2017]

### LSMDC 16

[LeCun et al., IJCAI 2017]

---

### New Tasks for Video VQA

1. **Counting Repetitions**
   - Q) How many times does the animal pump arms?
   - A) 2 times

2. **Reasoning State Transitions**
   - Q) What does the woman do after lowering the coat?
   - A) Pivot around
   - A) Stand

---

### A New Dataset for Video VQA: 165K QA Pairs

- **Repetition Count:** 30K
- **Repetition Action:** 23K
- **State Transition:** 59K

#### Methods for Generating QA Pairs

- **Template-based.**
  - Crowdsource via amazon mechanical turk
  - Pre-labeled Items
  - Match Answers
  - Strict quality control: Blacklist workers based on pre-labeled items
  - Synonyms are considered as correct answers
  - Generate four wrong answers based on a cosine similarity of the verbs

- **NLP-based QA generation**
  - Convert a declarative sentence to an interrogative sentence
  - Generate four wrong answers based on a cosine similarity of the verbs
  - Synonyms are considered as correct answers

### Results and Findings

- **Video-based model** works better than image-based models
- **Our model** with ST-attention shows the best result.

---

### References


[4] Li et al., Exploring Models and Data for Image Question Answering, in CVPR 2016

---

### Code and Dataset are available at

http://vision.snu.ac.kr/projects/tgif-qa