A Read-Write Memory Network for Movie Story Understanding

Seil Na† Sangho Lee† Jisung Kim‡ Gunhee Kim†

Seoul National University† SK Telecom‡

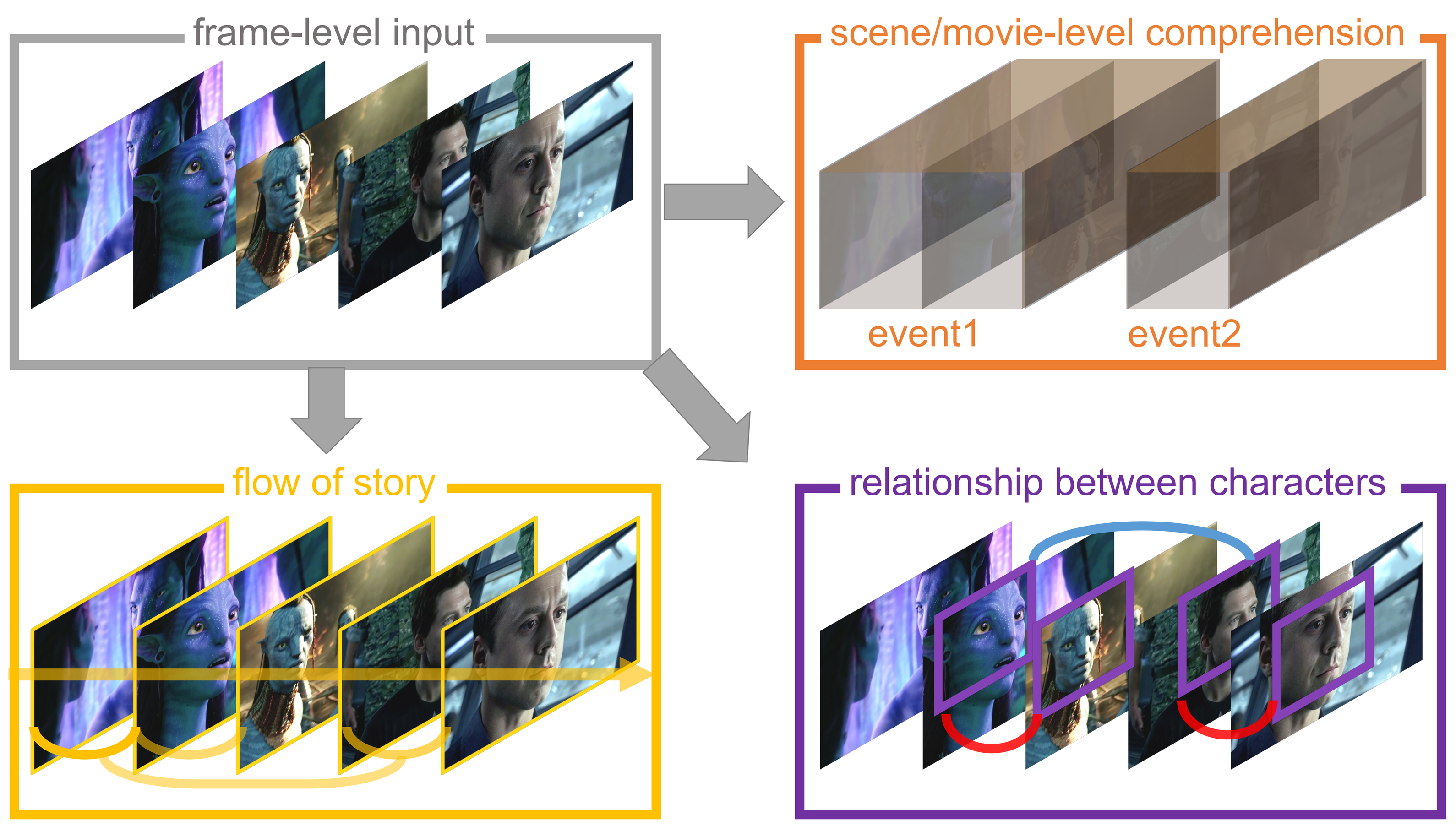
Code is available at https://github.com/seilna/RWMN

International Conference on Computer Vision 2017

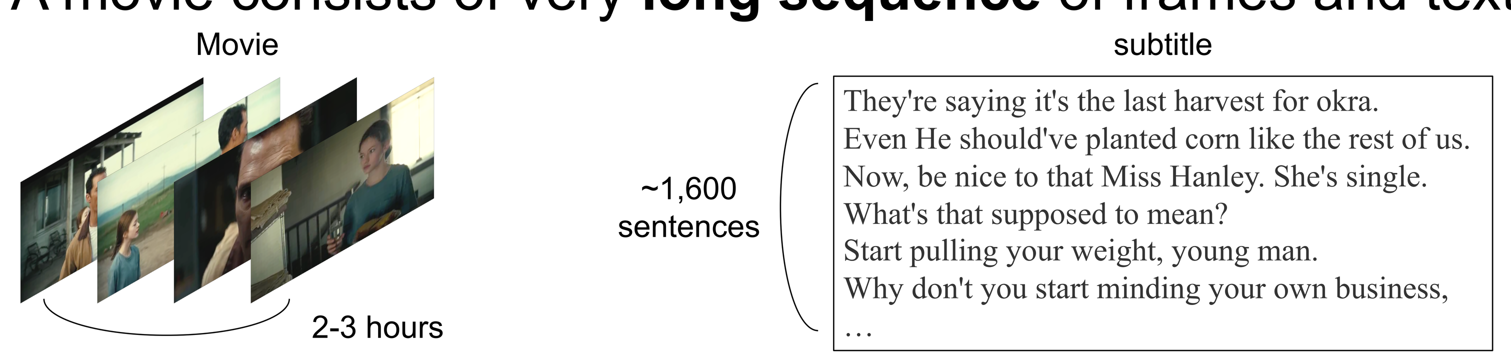
Motivation

It is hard to understand a long movie story

- It needs **high-level abstraction** given frame-level input only



- A movie consists of very **long sequence** of frames and text



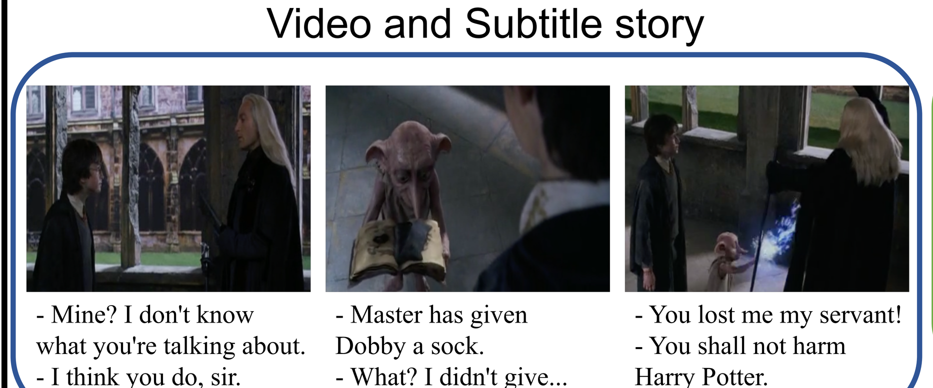
Objective

MovieQA [1]: Story Understanding Benchmark
Multi-choice Question & Answering with movie stories

Our model achieves **best** performance on **4** out of **6 tasks**

Video-based Q&A

Video and Subtitle story



Multi-choice Q&A

Q. What is Harry trick Lucius into doing?
A1. Releasing Dobby to Harry's care
A2. Releasing Dobby to Dumbledore's care
A3. Releasing Dobby to Hagrid's care
A4. Freeing Dobby
A5. Admitting he put Tom Riddle's diary in Ginny's cauldron

Text-based Q&A

Script story

- Harry's snowy owl, flies into the great hall skimming over their heads.
- It drops a long thin parcel.
- They unwrap the brown paper, to reveal a streamline broomstick with a highly polished handle.
- Stroking the white owl's feathers.
- Wearing leather gloves and Scarlett and gold cloaks, the Gryffindor Quidditch team assemble in the player's tunnel.

Multi-choice Q&A

Q. What sports they play in Hogwarts?
A1. They box
A2. They play golf
A3. They fight with brooms
A4. They play chess
A5. Quidditch

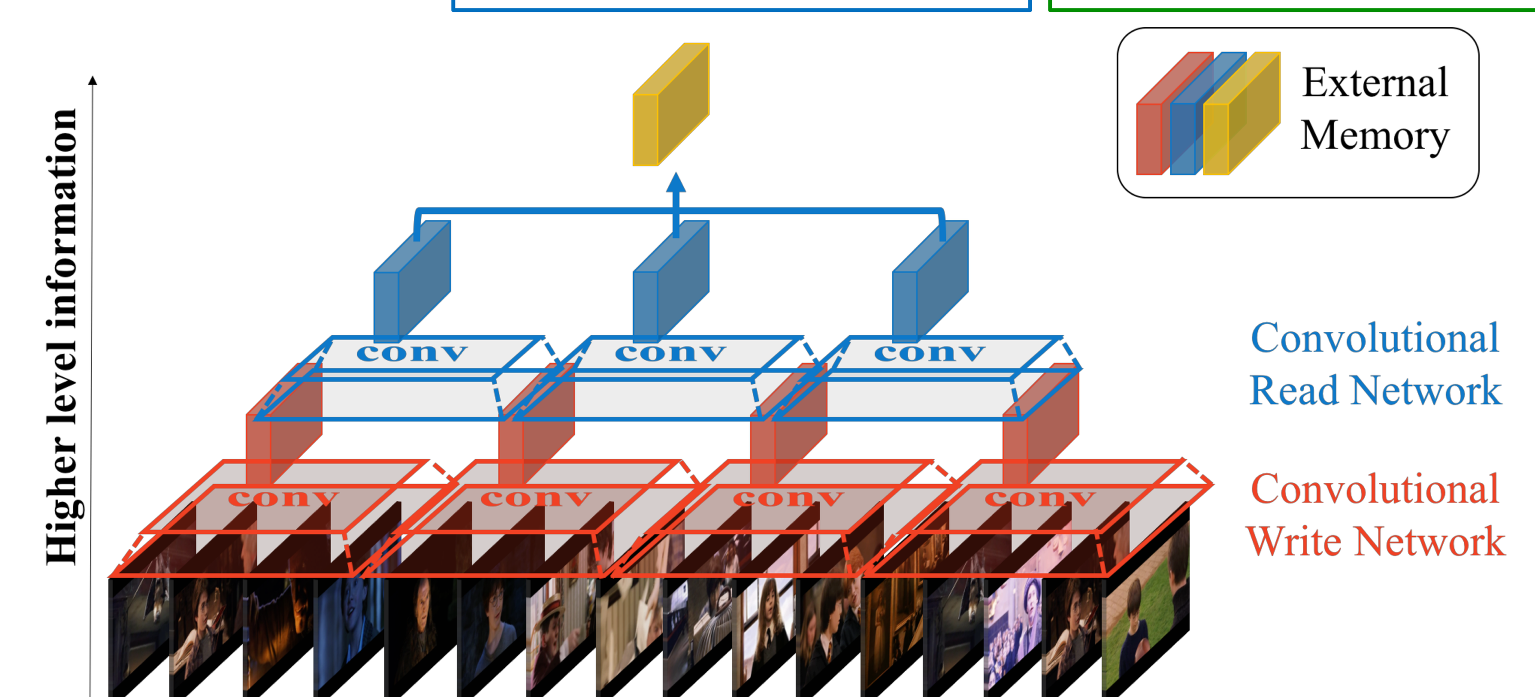
Our Solution – RWMN

Read Write Memory Network

External memory to fully utilize long input sequences

Convolutional write/ read to capture inter-actions btw. adjacent memory vectors

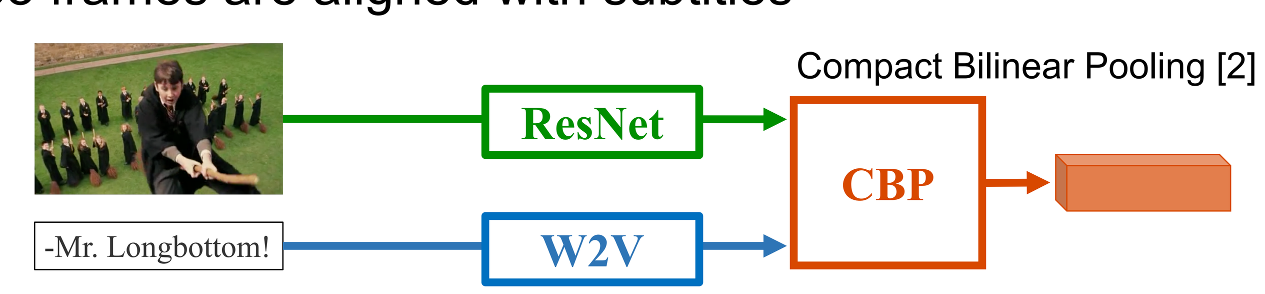
Query-dependent memory embedding to update memories conditioned on query



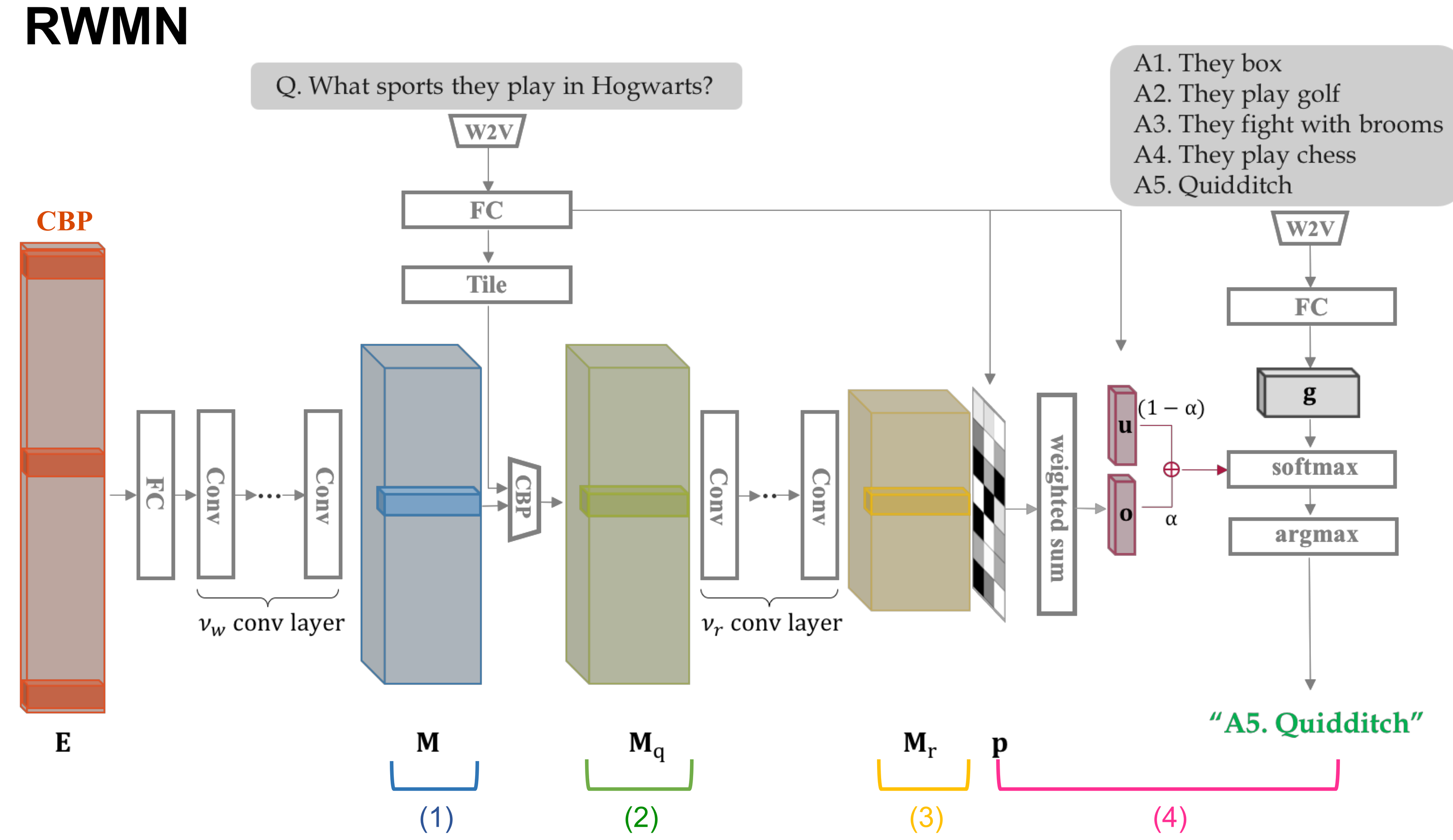
RWMN Architecture

Preprocessing & Feature extraction

- Video frames are aligned with subtitles



RWMN



(1): Write operation abstracts memory cells to higher-level via write convolutions

(2): Memory cells is updated conditioned on query via CBP

(3): Read operation abstracts updated memory cells appropriately for query

See the equations in the paper!

Quantitative Results

Results on MovieQA Benchmark

RWMN shows **best** performance on **4 tasks**
→ **Video+Subtitle / Subtitle / Script / Open-end task**

(RWMN-*): Ours and variants
(DEMN): [Kim et al. IJCAI 17]
(MEMN2N): [Tapaswi et al. CVPR16]
(SSCB-*): [Tapaswi et al. CVPR16]
(CNN Word Matching): [Wang et al. ICLR 17]

Video-based Q&A

Methods	Video
OVQAP	23.61
Simple MLP	24.09
LSTM + CNN	23.45
LSTM + Discriminative CNN	24.32
VCFSM	24.09
DEMN	29.97
RWMN	**36.25**

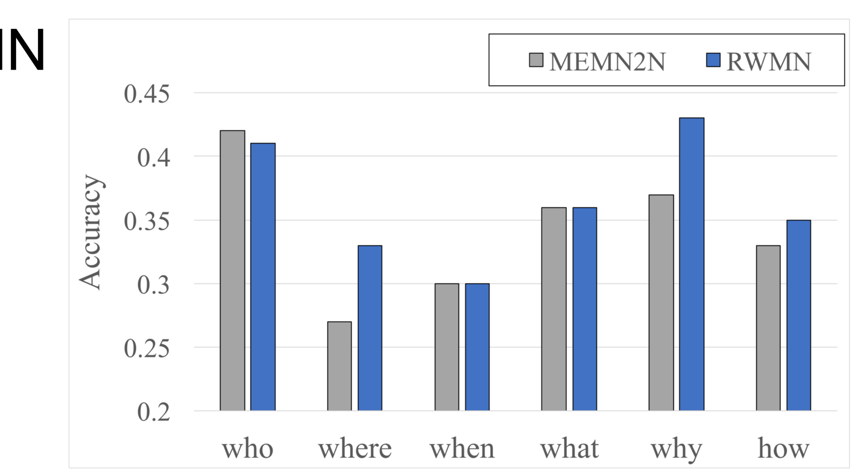
Text-based Q&A

Method	Subtitle	Script	DVS	Plot	Open-end
MEMN2N [24]	36.9	37.0	**35.0**	38.4	–
SSCB-W2V [24]	23.7	24.4	24.9	45.6	–
SSCB-TF-IDF [24]	26.5	23.9	23.3	47.4	–
Convnet Fusion	–	–	–	**77.6**	–
Longest Answer	–	–	–	–	25.6
RWMN	**38.5**	**39.4**	34.2	34.8	**36.6**

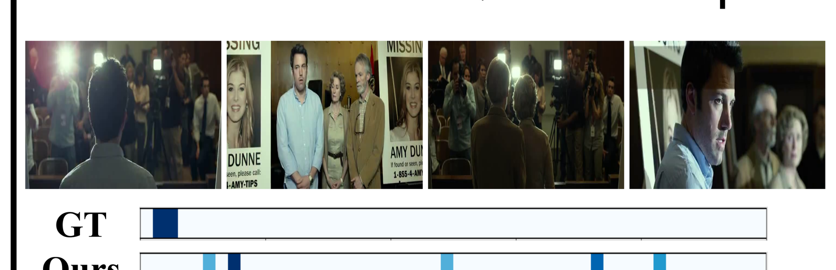
- All results as of the ICCV Submission Deadline, March 27, 2017 23:59 GMT

Qualitative Results

- Comparison between RWMN and MEMN2N according to **question types**



- Video-based Q&A examples with attention maps



Q. Why does Amy's disappearance receive heavy press coverage?
[0] Because her parents are popular
[1] Because Amy was the inspiration for the popular "Amazing Amy" children books
[2] Because Amy is a popular actress
[3] Because it happened on the day of her wedding anniversary
[4] Because her husband is popular

Q. What does Gandalf learn from Pippin's visions?
A1. Sauron will attack Minas Tirith
A2. Sauron will hide in Minas Tirith
A3. Sauron will attack Erebor
A4. Sauron will attack The Shire A5. Sauron will flee from Minas Tirith

Reference

[1] MovieQA: Understanding Stories in Movies through Question-Answering, M Tapaswi et al. CVPR 2016
[2] Compact Bilinear Pooling, Y Gao et al. CVPR 2016