MovieQA: Understanding Stories in Movies through Question-Answering

Answering and Evaluation

For any questions, email
tapaswi@kit.edu or
fidler@cs.toronto.edu

Registration and submissions are open!
Benchmark: http://movieqa.cs.toronto.edu

Generic QA framework

- all multiple-choice QA approaches

correct answer — story — answers

$$\text{[ ]} = \arg\max_a f(\text{[ ]}, \text{[ ]}, \text{[ ]})$$

question

- use a three-way function between story, question, and answers

e.g. a CNN-RNN approach on VQA
1. CNN(story = image); 2. RNN(question);
3. answer = softmax(vocabulary)

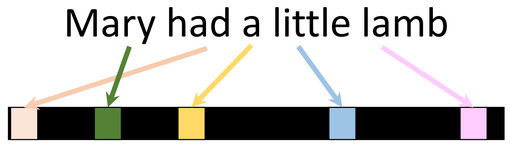
Data representations

represent data in vector space
- TF·IDF facilitates matching exact words
- Word2Vec meanings of words, allows synonyms
- SkipThought encodes semantics of sentence
- EncodeClip identifies objects/places, embeds in text space

EncodeClip

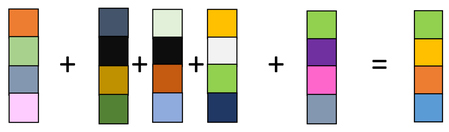
TF·IDF

Mary had a little lamb



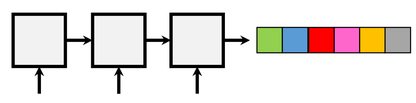
Word2Vec

Mary had a little lamb



SkipThought

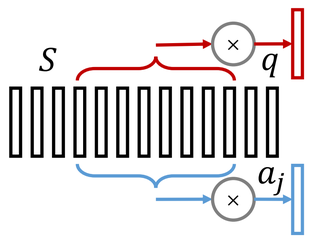
Mary had a little lamb



The Searching Student

core idea
- search within the story to find the best match for question and the answer
- windowed cosine similarity to compute how well a story fits a Q and A

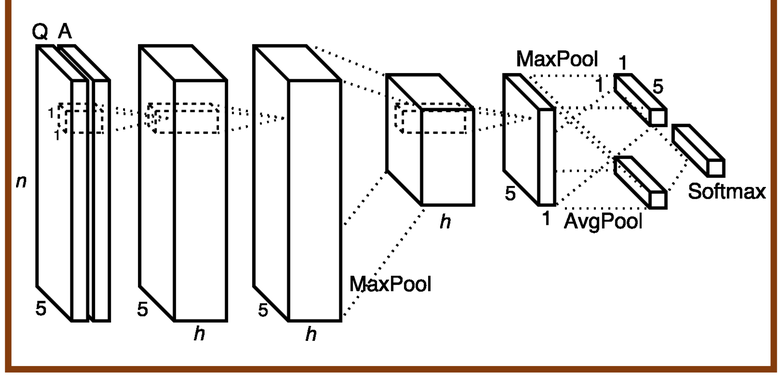
$$j^* = \arg\max_j f(S, q, a_j) \;\Big|\; \varphi(S, a_j) + \varphi(S, q)$$



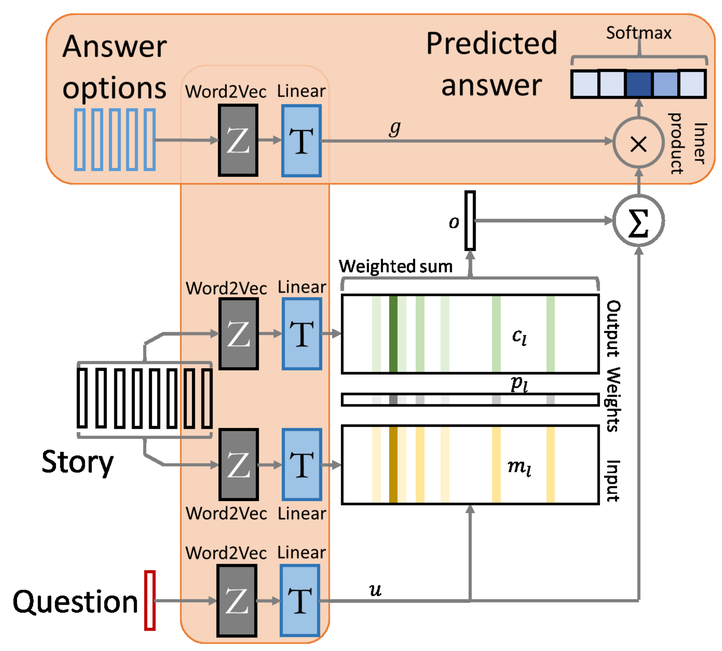
$$\varphi(s_k, a_j) = \langle s_k, a_j \rangle$$

SS with Convolutional Brain

core idea
- learn the three-way scoring function
- weighted combination of scores from $\langle \text{story}, \text{question} \rangle$ and $\langle \text{story}, \text{answer} \rangle$
- 1×1 convolutions



Modified Memory Network



The Hasty Student

- answer questions, don't look at story
- pick correct answer as the
  - longest answer
  - most similar/distinct answer
  - answer most similar to the question

answer length	longest	shortest	different
	25.3	14.6	20.4
within answer	TF·IDF	W2V	SkipT.
	21.7	**28.1**	25.4
question answer	TF·IDF	W2V	SkipT.
	13.0	19.3	**25.0**

The Hasty Turker

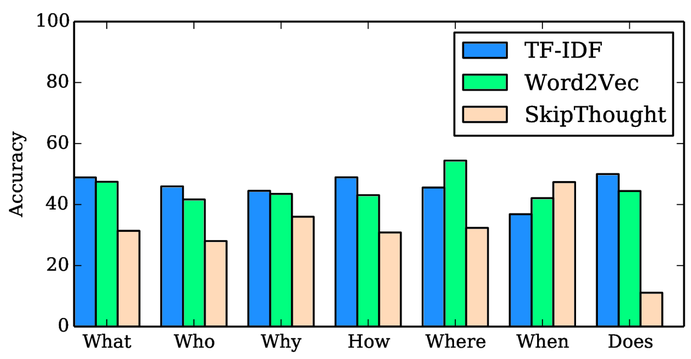
- 10 AMT workers answer questions without looking at the story

	200 QA subset	135 QA no names
overall accuracy	**27.6**	24.7
majority accuracy	**37.0**	30.4

Evaluation

Method	Plot	DVS	Subtitle	Script
Cosine TF·IDF	47.6	24.5	24.5	24.6
Cosine Word2Vec	46.4	26.6	24.5	23.4
Cosine SkipThought	31.0	19.9	21.3	21.2
SSCB TF·IDF	48.5	24.5	27.6	26.1
SSCB Word2Vec	45.1	24.8	24.8	25.0
SSCB SkipThought	28.3	24.5	20.8	21.0
SSCB Fusion	**56.7**	24.8	27.7	28.7
MemN2N 1 layer	40.6	**33.0**	**38.0**	42.3
MemN2N 3 layers	42.3	**33.0**	37.1	**43.0**

- plot-based answering easier, words repeated
- simple cosine similarity does not work with DVS, subtitles, scripts
- memory networks able to leverage this info.
- SSCB easily fuses all text representations



- Who: TF·IDF
- Where: Word2Vec
- When: SkipThought

	Clips	Video	Subtt.	V+S
SSCB	All	21.6	22.3	21.9
MemN2N	All	23.1	38.0	34.2
	QA	22.6	38.0	33.3

- video based answering needs more work
- individual vision modules may be required (e.g. identities, places)

CVPR 2016