Are we asking the right questions in MovieQA?

Bhavan Jasani, Rohit Girdhar and Deva Ramanan

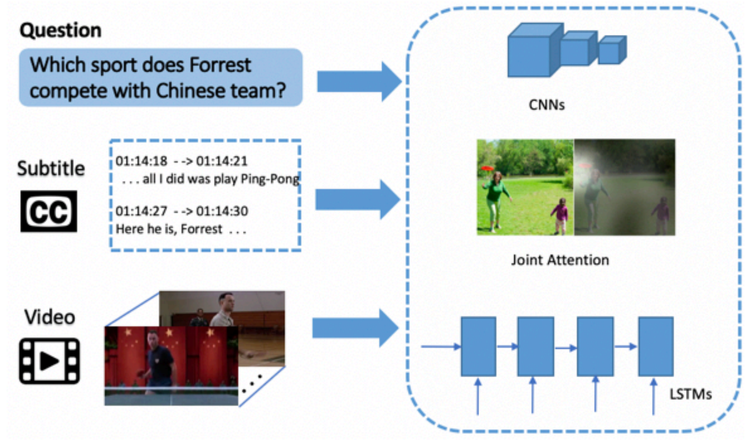
SCAN ME

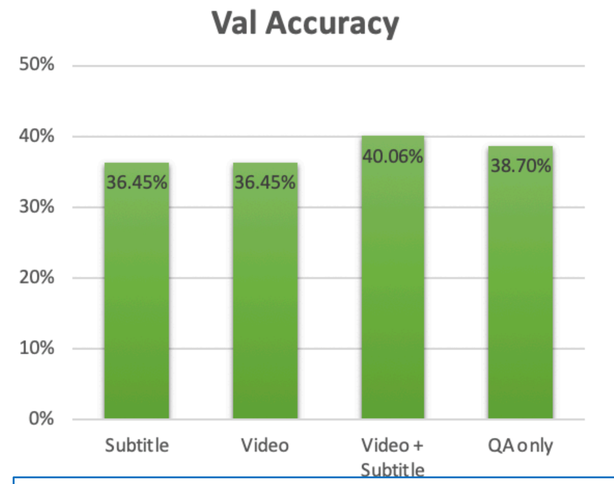
ICCV 2019 Seoul, Korea

Answering questions about movies without watching movies

Q. Who makes Indy return the crucifix after escaping from the grave robbers?

A1. Coronado
A2. No one, he keeps it
A3. The local sheriff
A4. The Boy Scout troop
A5. The grave robbers

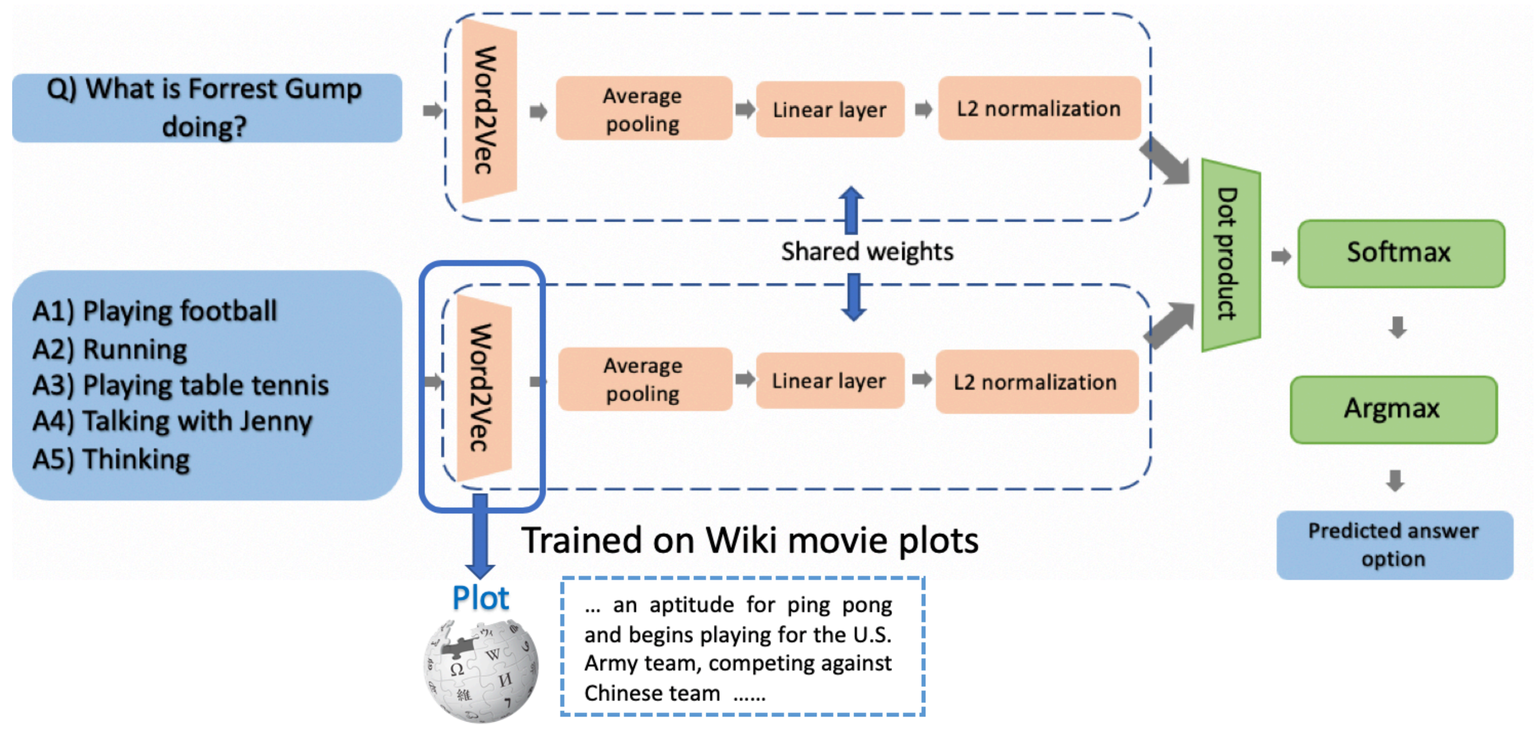




Results for best model on the leaderboard with available code [1]

MovieQA task: given a question, 5 answer choices, and a movie context (encoded with videos, scripts, and subtitles), select the correct answer.
Prior works: Use deep networks to incorporate information from videos and subtitles to do this task, but fail to utilize them.
Our approach: Much simpler model that achieves state of the art performance, without using any video or subtitle context. We attribute this to linguistic bias in the data.

WikiWords Embedding Model



WikiWords is simple – it selects the answer closest to the question in word2vec space. It does not use any movie context in the form of clips or subtitles. Importantly, it uses word2vec trained on Wikipedia movie plots (movie summaries).

Results - State of the art on 4 out of 5 MovieQA categories

Leader board submission	Videos + Subtitles
WikiWords model	**46.98**
New method to optimize all MEM network (anonymous)	45.31
Multimodal dual attention memory [2]	41.41

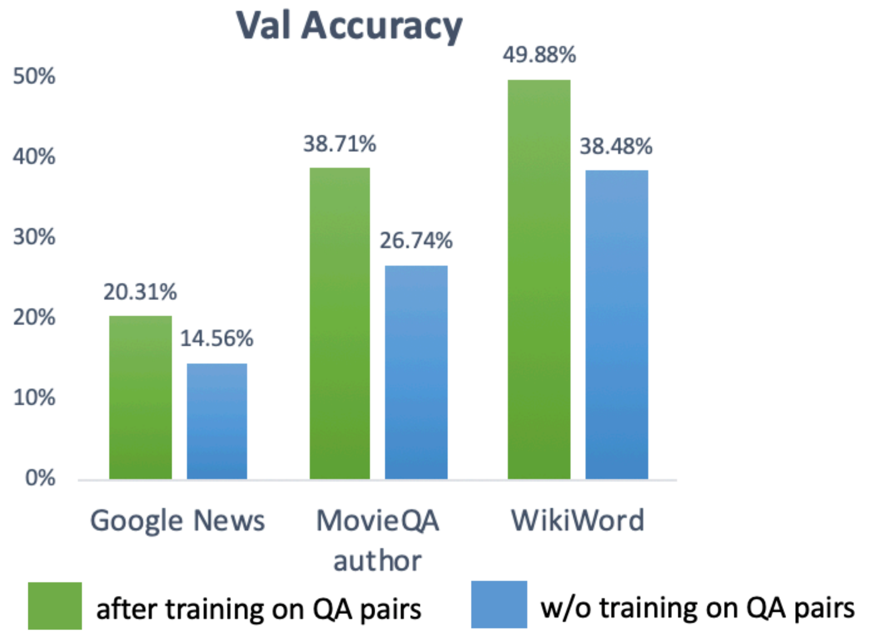
Leader board submission	Subtitles
WikiWords model	**44.01**
Speaker Naming in Movies [3]	39.36

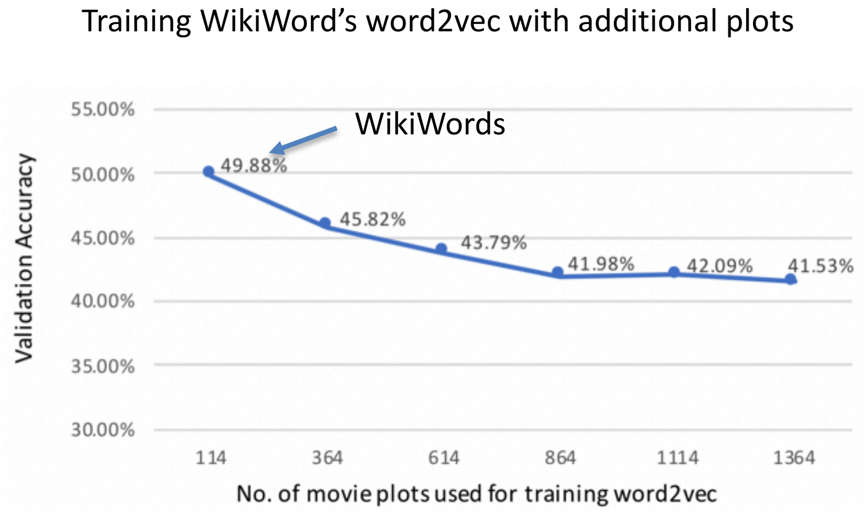
Leader board submission	DVS
WikiWords model	**49.65**
MovieQA benchmark [4]	35.09

Leader board submission	Scripts
WikiWords model	**45.49**
Read Write Mem. Net [5]	39.36

Compared to best published results, WikiWord improves accuracy by 5% for video + subtitle category, 5% for subtitle, 15% for DVS and 6% higher for scripts.

Ablation Study – Training data for word2vec

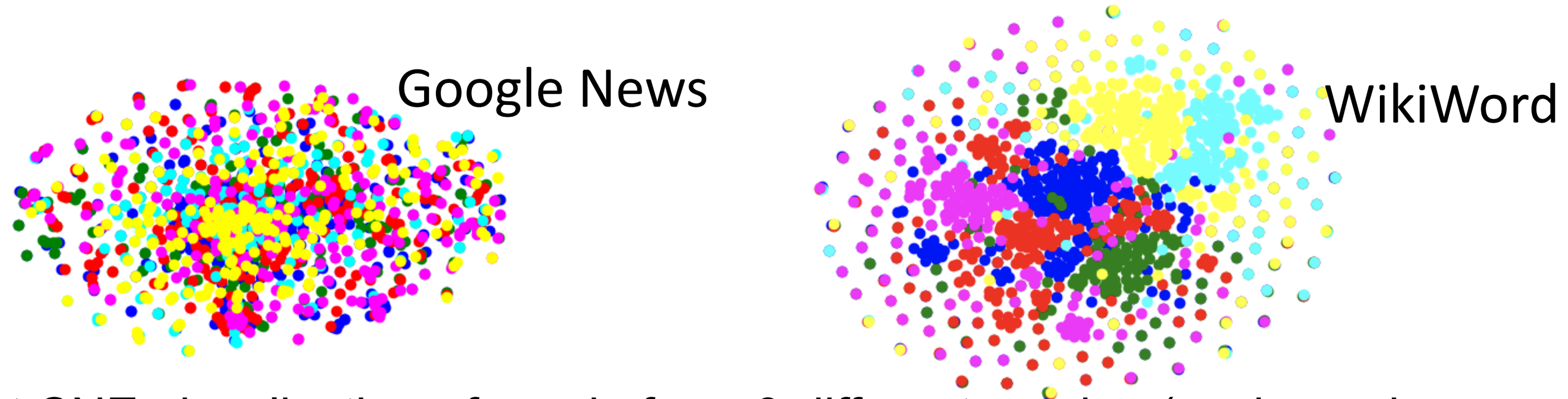




Less and concise is better!

Using a generic word2vec, like the one trained on Google news, results in chance level accuracy. Default MovieQA word2vec trained on a huge pool of movie plots gives intermediate accuracy. Our WikiWord embedding, trained on a subset of movie plots (which are part of the MovieQA dataset), performs best.

Ablation Study – t-SNE visualization of word2vec



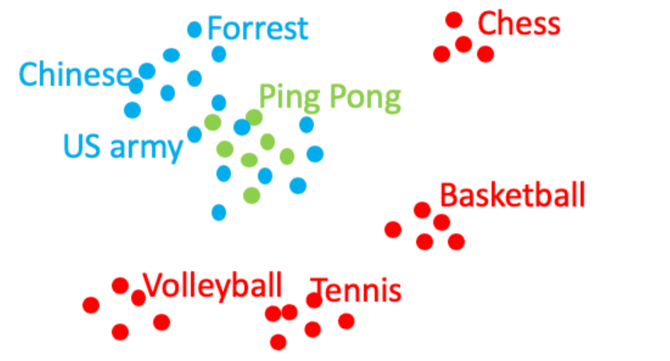
t-SNE visualization of words from 6 different movies (each movie corresponds to a different color). WikiWord embedding has separate clusters for each movie, hence it preserves movie semantics, while Google News loses movie semantics.

Source of bias

Q) Which sport does Forrest compete with Chinese team?

A1) Ping pong
A2) He competes in volleyball
A3) Tennis
A4) Basketball
A5) Chess

Wiki plot: Forrest discovers aptitude for ping pong and begins playing for the U.S. Army team, eventually competing against Chinese teams on a goodwill tour.



WikiWord's embedding space

Mechanical Turkers generated QA pairs by looking at Wikipedia plots, ignoring other sources of information like videos and subtitles. Pairs of questions and the correct answer tend to contain common keywords from these movie plots. This suggests that their WikiWord embeddings will tend to lie near each other. In the given example, the words "Forrest" and "Chinese" lie close to "Ping Pong" in WikiWord embeddings.

Ability to generate better benchmark

Type	WikiWord embedding	TVQA baseline model
Original dataset	49.88	32.50
Only biased	99.41	47.80
Only unbiased	25.68	22.50

Experiments on TVQA dataset

Model	Word embedding	Val accuracy
WikiWord embedding	Google News	32.76
	TVQA subtitles	32.66
TVQA baseline [6]	Random weights	39.61
	Wikipedia articles	40.18
	TVQA subtitles	39.65

WikiWord naturally allows us to find the subset of the dataset which is unbiased. We consider the QA's which WikiWord is unable to answer as unbiased. We show QA only models perform chance level on this split, indicating need for information from videos and subtitles. Also, our experiments on TVQA dataset shows the data used for training word embedding doesn't have any impact, indicating it is free from such bias.

References:

[1] B. Wang, Y. Xu, Y. Han, and R. Hong, Movie question answering: remembering the textual cues for layered visual contents. In *AAAI*, 2018.
[2] K.-M. Kim, S.-H. Choi, J.-H. Kim, and B.-T. Zhang, Multimodal dual attention memory for video story question answering. In *ECCV*, 2018
[3] M. Azab, M. Wang, M. Smith, N. Kojima, J. Deng, and R. Mihalcea, Speaker naming in movies. *arXiv preprint arXiv:1809.08761*, 2018
[4] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, Movieqa: Understanding stories in movies through question-answering. In *CVPR*, 2016.
[5] S. Na, S. Lee, J. Kim, and G. Kim, A read-write memory network for movie story understanding. In *CVPR*, 2017
[6] J. Lei, L. Yu, M. Bansal, and T. Berg, Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018