Are we asking the right questions in MovieQA?

Bhavan Jasani
Carnegie Mellon University
bjasani@cs.cmu.edu

Rohit Girdhar
Carnegie Mellon University
rgirdhar@cs.cmu.edu

Deva Ramanan
Carnegie Mellon University
deva@cs.cmu.edu

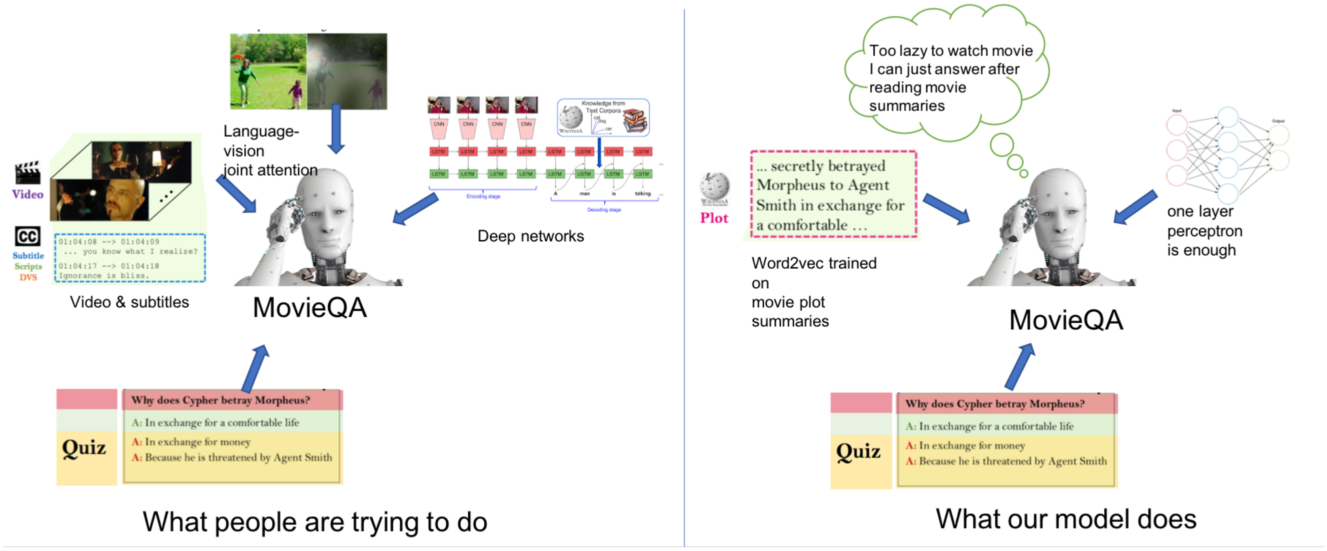

Video
Subtitle
Scripts
DVS
01:04:08 --> 01:04:09
... you know what I realize?
01:04:17 --> 01:04:18
Ignorance is bliss.
Video & subtitles
Language-vision joint attention
CNN
LSTM
Knowledge from Text Corpora
cat
dog
car
Wikipedia
A
man
is
talking
Encoding stage
Decoding stage
Deep networks
MovieQA
Quiz
Why does Cypher betray Morpheus?
A: In exchange for a comfortable life
A: In exchange for money
A: Because he is threatened by Agent Smith
What people are trying to do
Too lazy to watch movie I can just answer after reading movie summaries
Wikipedia
Plot
... secretly betrayed Morpheus to Agent Smith in exchange for a comfortable ...
Word2vec trained on movie plot summaries
Input
Output
one layer perceptron is enough
MovieQA
Quiz
Why does Cypher betray Morpheus?
A: In exchange for a comfortable life
A: In exchange for money
A: Because he is threatened by Agent Smith
What our model does

Figure 1: **Why watch a full movie when I can learn from wiki-plots?** Traditional video QA models use videos, subtitles in conjunction with question and option text to answer the question. We show a significantly simpler model, using a word embedding trained on plot synopses, using only question and options, is able to outperform all reported performance on the challenging MovieQA benchmark. Parts of the figure taken from [20, 22, 23].

Abstract

Joint vision and language tasks like visual question answering are fascinating because they explore high level understanding, but at the same time, can be more prone to language biases. In this paper we explore the biases in the MovieQA dataset and propose a strikingly simple model which can exploit them. We found that using right word embedding is of utmost importance. By using an appropriately-trained word embedding, about half the Question-Answers (QAs) can be answered by looking at the questions and answers alone, completely ignoring narrative context from video clips, subtitles, and movie scripts. Compared to the best published models on the leaderboard, our simple question+answer only model improves accuracy by 5% for video + subtitle category, 5% for subtitle, 15% for DVS and 6% higher for scripts. We further propose a solution to mitigate these language biases by creating a subset of hard questions that require additional contextual cues to answer.

1. Introduction

Language has long been an integral part of visual understanding. From objects [4, 13] to human actions [10], categorization of visual data has lead to rapid developments in computer vision techniques, especially with deep learning. However, language is a much more powerful tool, and researchers recently have started to be apply it to domains beyond simple classification. To that end, various tasks such as image captioning [25] and Visual Question-answering (VQA) [1] have been proposed. VQA has arguably emerged as one of the most popular vision tasks, primarily due to its simple setup and clear evaluation.

QA tasks are particular intriguing for videos, where they can explore cognitive storytelling concepts (such as intentions and goals) difficult to extract from static images. Unsurprisingly, there have been considerable efforts in bridging the gap between language and spatio-temporal understanding of videos. To that end, a recently released dataset, MovieQA [20], has extended the VQA philosophy to videos, by collecting short real-world movie clips, along with subtitles and wiki-plots, and defining multiple choice questions on them. Similar ideas have been pursued in other works as well [12].

While there has been a reasonably large amount of work in this direction, most methods [16, 24] do not make strong use of visual features and intead rely heavily on language-based cues such as subtitles or wiki-plots. This raises the question: are our video models unequipped to truly understand videos, or is the MQA task unfairly biased against actually needing visual information?

In this work, we explore this question in detail. We propose a very simple model, relying purely on the word embedding of the question and answer options. We report a surprising finding: this simple model, when trained appropriately, already outperforms all reported methods on MovieQA [20] test set. This includes models that use subtitles, scripts, and videos, while our model only uses the question and answer. We have submitted our results to the test evaluation server, and are ranked first in four out of five categories at the time of submission.

The role of plot-synopsis: It is worth noting the one category that we do *not* win is plot-synopsis, where the current state-of-the-art is quite high (85%). This is explained by the fact that the question and answers were *constructed* by inspection of movie plots from wikipedia. This category provides aligned training examples of $\{(\text{question,answer,plot})_i\}$ tuples for learning, which can be exploited by powerful language models that exploit such aligned data [3]. In contrast, we learn embeddings from unaligned training examples of $\{(\text{question,answer})_i\}$ pairs and movie plots $\{\text{plots}_i\}$, which are available in all benchmark category protocols. Our results demonstrate that *unsupervised* learning of word-embeddings from *un*aligned movie plots still captures a rich amount of narrative structure about the movies of interest. In some sense, our results may reveal a flaw in the underlying protocol of this benchmark.

Fixing the bias: Exposing this bias in the data enables us to build a better benchmark that is harder to solve trivially. To that end, we propose a simple technique using our model to convert the existing benchmark into a *harder* set. We show, for various well-known QA models including ours, that the performance is much lower, opening up avenues for further research on the truly hard part of joint video language understanding.

The paper is organized as follows. We start by discussing related work in Section 2,in section 3 we describe our simple QA-only model, what it learns and the source of the bias in the dataset. And we propose a way to mitigate the bias. In section 4 we describe the details of the dataset, the results of out model on leaderboard and show our experiments with for different word embeddings.

2. Related Work

Video and language: Joint learning of language and vision offers various possibilities. People have been working on lots of different tasks and lots of datasets have been released. For example, Movie descriptions [18], video understanding through fill in the blank [14], video retrieval [21], character co-referencing [17], image captioning [25]. Lots of work have focused on using movies [7, 17, 21], because movies provide with time synchronized audio, subtitles and videos.

Visual QA task: Humans learn through Question Answering (QA) and answering a question is an important way to show understanding towards a concept and it provides a an easy unambiguous evaluation metric for joint language and vision tasks. The Question Answering task is to to predit correct answer from a list of options for a given question based on a story which provides the context. Lots of visual question answering datasets have been recently realized that include image based question answering datasets like VQA [1], and more recently videos based QA like MovieQA [20] which is constructed from movies, TVQA [12] which is constructed from TV series and TGIF QA [8]. Also there are lots of reading comprehension [6] datasets which are the pure language based QA datasets. Movies provide wide variety of human activities including high-level semantics of human actions like intention, motivation, and emotion in a very concise story like way. Hence MovieQA is a very interesting dataset with movie clip and questions from 140 movies. Some of the recent video models on MovieQA include Multimodal Dual Attention Memory [11], Layered Memory Network [24], Read Write Memory Network [16].

Biases in Visual QA datasets: Text features are comparatively easier then visual features and because of which in joint language and vision tasks are more prone towards using language features and neglecting the visual clues, part of this happens because of language priors and biases which models can easily exploit. It can also happen that models can predict the correct answer by just looking at the questions and totally neglecting the story because of the inherent biases. It was shown that the origianl VQA 1.0 [1] dataset had lot of bias, [27] proposed a simple baseline model which had comparable performance to most of the complex approaches of the time. In VQA 2.0 [5], the authors tried to reduce the langauge bias in the dataset by augmenting the original dataset. Recently [9] compared different reading comprehension datasets to provide sensible baselines including comparison of models which predict the

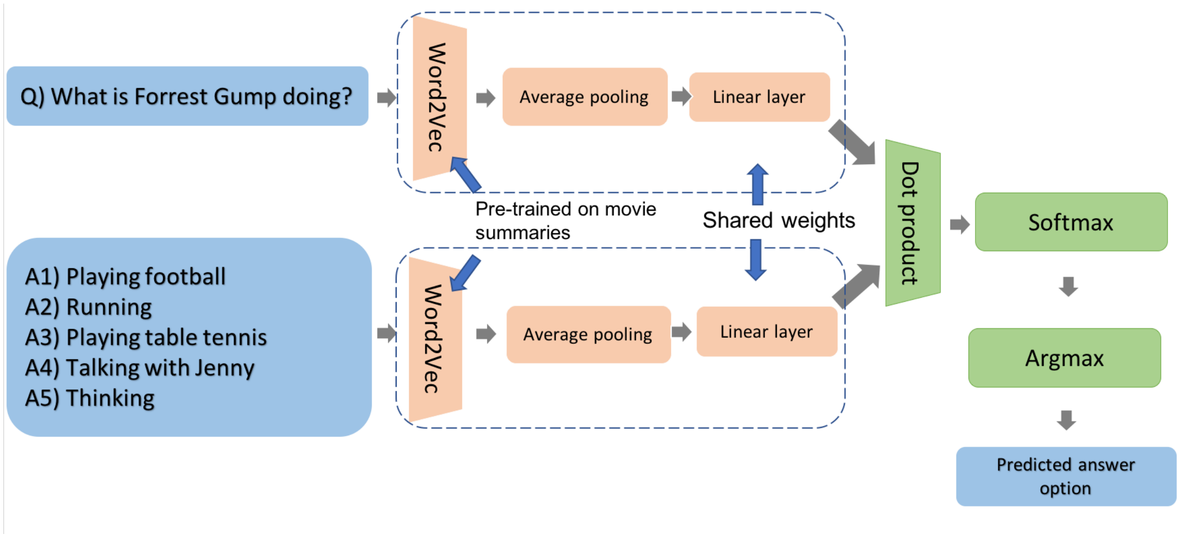


Figure 2: Our simple model which can find bias by predicting answer by just looking at the question. It takes the Question and 5 answer choices as input, computes there word embedding and picks the answer choice based on highest dot product similarity with the question.

answer by just looking at the question alone or by just looking at the story alone.

3. Approach

3.1. Model

The general Question-Answering framework is to model a three variable scoring function between the question, the answer choices and the story. The scoring function tells the score of every answer choice for the given question by looking at the associated story which provides the context, and the answer choice with the highest score is selected as the prediction of the function.

In our proposed model shown in figure 2 which we use to find bias in the dataset we further even remove the use of story (subtitles and videos in the MovieQA dataset). Our simple model just looks at the question to pick the correct answer and totally ignores the story. The pre-trained word embeddings we use provide a weak general context of all the movies in the dataset, which we found is sufficient and we can totally ignore the question specific context of stories from which ideally the question is supposed to be answered. In essence our model consists of embedding questions and the 5 multiple choice answers using an embedding layer and then we compute dot product similarity between question and each of the answer choices and pick the choice with highest value as the prediction.

The embedding layer takes as input the sentence (the raw question and all the answer choices) and for each word in the sentence computes a 300 dimensional vector representation using a pre-trained word2vec. [15] model. The 300-d vector representation of every word in sentence are then average pooled to get a 300-d vector representation of the sentence, which is then passed through a linear layer (initialized as an identity matrix) and then l2 normalized.

In accordance with the basic Question-Answering framework provided in MovieQA paper [20] the weights of the word2vec model are kept fixed. This is because there are about 26,000 words in the vocabulary of MovieQA dataset and which for a 300d vector representation makes 26000x300 = 7.8 million parameter, huge for a small dataset with about 6000 QAs (for video category) and hence can easily overfit. So instead in accordance to MovieQA [20] we use a linear projection layer (300x300) initialized as an identity matrix to reduce the number of trainable parameters. The word embedding layer is shared for question and all the answer choices, and also for subtitles used in our other experiments.

The word2vec model is trained unsupervised on plot synopses which are movie summaries written by movie fans, they range from one to twenty paragraphs. For the MovieQA dataset these are actually used by AMT workers while generating question and answers. There is one category of text based QA task in MovieQA in which one is allowed to use these plot synopses as the story for answering questions.

An important point to emphasis is that the plot synopses are used in an unsupervised manner just for training the word2vec. There purpose is to provide a movie specific general knowledge about the characters and entities particular movies. It turns out they provide sufficient information to solve half the dataset without use of any other context through videos or texts which is supposed to be used during training. MovieQA authors also provide a similar word2vec trained on movies plots.

This is probably as simple a model as it can get. The most important thing in our simple model is the specific data on which our word2vec is pre-trained as that's the only thing which drives our model to pick the answer based on just looking at the question. Details and ablation studies of which are mentioned in section 4.

3.2. What our simple QA only model learns?

In a way our simple QA model, the pre-trained word2vec model is trying to memorize the occurrence of nearby words in the movie plot synopsis and since the question-answers are made by AMT workers by only looking at the movie plot synopsis, it is able to correctly answer the QAs in half the dataset.

Figure 3 shows the predictions of our simple model with 'train+val' word2vec which are correct and figure 4 shows the predictions which are incorrect That also the highlights the prominent words in the question, the correct answer and the the line in the movie plot form which the QA was made by the AMT workers.

We found that the model first tries to select the answer choice which has most movie specific words as that in the question, this happens because in this case the word embedding of question and the selected answer would be very close. The another thing which model tries to do if the previous thing doesn't hold is to select the answer whose movie specific word(s) occur adjacent to the movie specific word(s) of the question in the movie plots (since in word2vec space nearby text words have very high dot product similarity). Again this ensures that the word embeddings of question and the selected answer choice would have very high similarity. And surprisingly just doing this our simple model close to 50 percent accuracy on the video based QA task with only looking at question and picking the answer.

3.3. Source of bias

Apart from finding the bias in the video based QA task of MovieQA, we also tried our same approach on full dataset (which contains text based QA's as well). Again our simple word2vec model which just looks at question was able to get similiar results, of 44% indicating that the bias is present in the full dataset as well.

From the analysis of our model in the previous section it's evident the main reason for the bias in the dataset is that movie specific entity names in the question and the corresponding correct answer choice made by AMT worker resemble very closely to the words in movie plot synopsis. There are two main issues here:

1) For many of the QA's the movie specific words in the correct answer choice and the questions are very similar and are directly copied (without rephrasing) from the wiki-plots

2) For many QA's the incorrect answer choices are very different from the correct answer choice and at the same time the correct answer choice has movie specific words adjacent to the movie specific words in the question with respect to the movie plots but that's not the case for the four other incorrect answer choices.

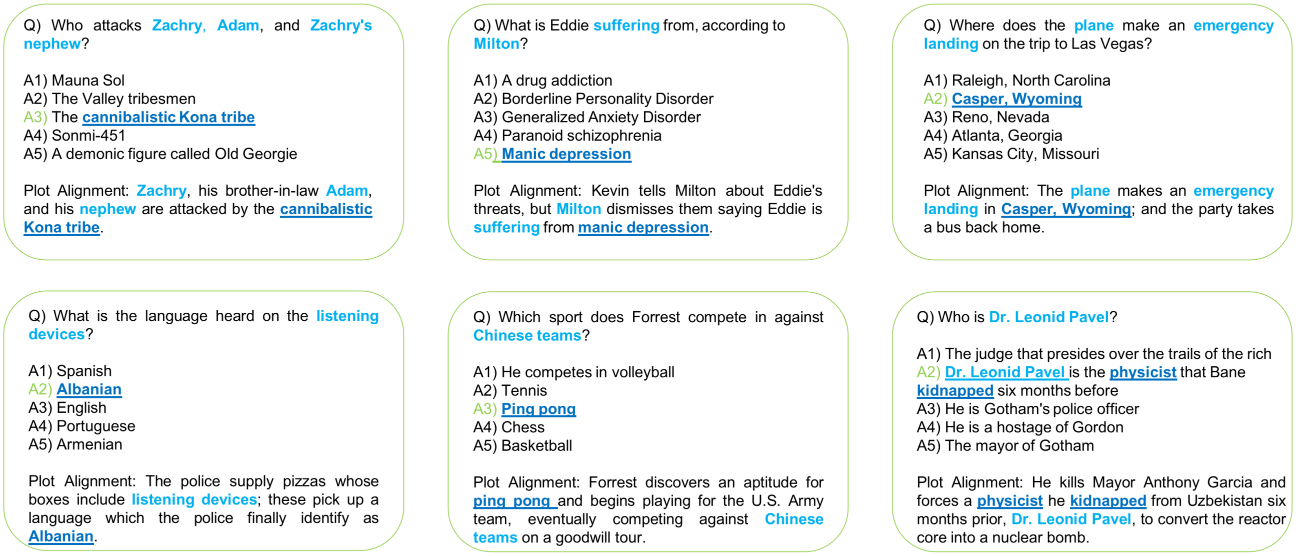
Due to this reasons the correct answer choice is very similar to the question and very distinct from the incorrect answer choices in their word embedding space, and because of which our model needs no information from the contexts (subtitles and video clips) to figure out the correct answer for atleast about half the dataset. Just a well trained word embedding model suffices.

As mentioned earlier the important thing is the QA were created by AMT workers by just looking at the wiki/movie-plots. If one had access to movie plots during training and test time (unlike in our case) then it would very trivial to find the correct answer. One would just need to search for line in the wiki/movie-plots (on average 40 lines) most similar to the question and the correct answer choice would be similiar to that text line. In fact MovieQA has a separate task where in one is allowed to look at movie plots in supervised manner (use them as the story) and answer the question, not surprisingly the highest accuracy on the leaderboard for it is 85.12%. That's the only category where our simple QA-only model achieves an accuracy of 44% on the leaderboard and doesn't beat the state of the art model.

	Google	MovieQA	Our best W2V
QA only	24.71	38.70	50.00
Subtitle	25.16	36.45	47.62
Video	27.87	36.45	50.67
Videos + subtitle	25.39	40.06	48.87

Table 1: Validation experiments with different word2vec on the model in [24]. It can be seen that using subtitle or videos doesn't help much as they have similar accuracy as that when the model just uses questions to predict the answer. This shows the question-answer bias. Also the table shows that in general the accuracy of the model (for all different modalities) increases as we use better and better word embedding from generic one like Google to movie specific one like ours

Videos were later automatically aligned with movie plot lines and hence indirectly associated with the QA. And so usefulness of videos for the QA task can be quite So it's not very surprising that the videos are not much useful. A good way to generate QA's would be to actually show videos to the AMT workers and avoid showing any other text data. The authors of MovieQA [20] mention the reason for not doing this is because it would be very expensive and time consuming for them to show full movies to AMT workers,



Q) Who attacks **Zachry**, **Adam**, and **Zachry's nephew**?

A1) Mauna Sol
A2) The Valley tribesmen
A3) The **cannibalistic Kona tribe**
A4) Sonmi-451
A5) A demonic figure called Old Georgie

Plot Alignment: **Zachry**, his brother-in-law **Adam**, and his **nephew** are attacked by the **cannibalistic Kona tribe**.

Q) What is Eddie **suffering** from, according to **Milton**?

A1) A drug addiction
A2) Borderline Personality Disorder
A3) Generalized Anxiety Disorder
A4) Paranoid schizophrenia
A5) **Manic depression**

Plot Alignment: Kevin tells Milton about Eddie's threats, but **Milton** dismisses them saying Eddie is **suffering** from **manic depression**.

Q) Where does the **plane** make an **emergency landing** on the trip to Las Vegas?

A1) Raleigh, North Carolina
A2) **Casper, Wyoming**
A3) Reno, Nevada
A4) Atlanta, Georgia
A5) Kansas City, Missouri

Plot Alignment: The **plane** makes an **emergency landing** in **Casper, Wyoming**; and the party takes a bus back home.

Q) What is the language heard on the **listening devices**?

A1) Spanish
A2) **Albanian**
A3) English
A4) Portuguese
A5) Armenian

Plot Alignment: The police supply pizzas whose boxes include **listening devices**; these pick up a language which the police finally identify as **Albanian**.

Q) Which sport does Forrest compete in against **Chinese teams**?

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

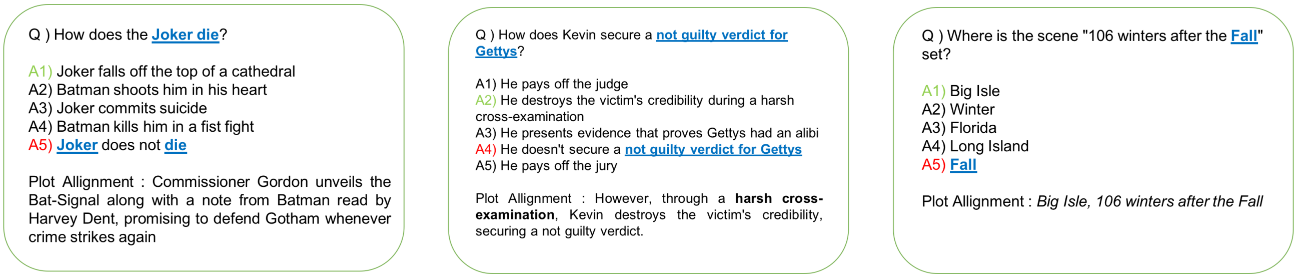
Plot Alignment: Forrest discovers an aptitude for **ping pong** and begins playing for the U.S. Army team, eventually competing against **Chinese teams** on a goodwill tour.

Q) Who is **Dr. Leonid Pavel**?

A1) The judge that presides over the trails of the rich
A2) **Dr. Leonid Pavel** is the **physicist** that Bane **kidnapped** six months before
A3) He is Gotham's police officer
A4) He is a hostage of Gordon
A5) The mayor of Gotham

Plot Alignment: He kills Mayor Anthony Garcia and forces a **physicist** he **kidnapped** from Uzbekistan six months prior, **Dr. Leonid Pavel**, to convert the reactor core into a nuclear bomb.

Figure 3: QAs which are correctly predicted by our QA only model and hece are the biased QA's. Correct answer is highlighted in green. Light blue coloured words are the movie specific words common between the question and the line in wiki-plot from which the question was made by Turkers. Dark blue underlined words are the movie specific words common between the correct answer and the line in wiki-plot. For example in 5th Question, the model predicted A3 because Ping pong (movie specific word in the correct answer) is the only word that appears close to Chinese teams (movie specific word in the question) in the movie plot, all other answer options don't occur in the movie plot and hence are very different, hence making it a biased QA.



Q) How does the **Joker die**?

A1) Joker falls off the top of a cathedral
A2) Batman shoots him in his heart
A3) Joker commits suicide
A4) Batman kills him in a fist fight
A5) **Joker** does not **die**

Plot Allignment : Commissioner Gordon unveils the Bat-Signal along with a note from Batman read by Harvey Dent, promising to defend Gotham whenever crime strikes again

Q) How does Kevin secure a **not guilty verdict for Gettys**?

A1) He pays off the judge
A2) He destroys the victim's credibility during a harsh cross-examination
A3) He presents evidence that proves Gettys had an alibi
A4) He doesn't secure a **not guilty verdict for Gettys**
A5) He pays off the jury

Plot Allignment : However, through a **harsh cross-examination**, Kevin destroys the victim's credibility, securing a not guilty verdict.

Q) Where is the scene "106 winters after the **Fall**" set?

A1) Big Isle
A2) Winter
A3) Florida
A4) Long Island
A5) **Fall**

Plot Allignment : *Big Isle, 106 winters after the Fall*

Figure 4: QA which are wrongly predicted by our QA only model. These are the QA's which are more likely to be less biased. Prediction of the model is in red and the correct answer is in green. For example in 3rd QA, the model predicted A5 because it's the only word as compared to all other options, common with a word in question and hence would have very dot product similarity in the word embedding space with the question.

which is a legitimate reason. Table 1 shows the performance of different modalities (QA only, subtitles, videos and vidoes+subtitles) for the top model on leaderboard with publicly released code. We make modifications to their code to do our experiments with different modalities and word embeddings and so the numbers don't exactly match with those in their paper.

3.4. Reducing the bias

It's not uncommon to have some inherent bias in machine learning datasets, specifically joint language and vision tasks are more prone to have language biases as has been earlier [5] [27].

In VQA 2.0 [5] authors create a 'balanced' dataset by augmenting the original 'unbalanced' dataset which has high amount of inherent language biases. For a given a question-answer-image triplet, they collect a similar looking image which for the same question has a different answer. They hypothesis that in that case the only way to get the answer correct to both the question-answer-image triplet would be for models to be forced at looking at the images. They also show that existing state of the art models perform worse on the new balanced dataset. Further they have shown that existing models perform better when trained and tested on the balanced dataset in comparison to when trained on unbalanced dataset and tested on the balanced dataset. Indicating that for the unbalanced dataset the models were exploiting the bias to a high extent but they actually start to use visual clues when they are trained on the balanced dataset.

With the language bias which we discovered in the MovieQA dataset makes it difficult to properly utilize videos. Dataset collection is a time consuming process and so to reduce bias in the existing dataset we propose a very quick fix to mitigate the problem. Our fix is to remove the biased QAs which can be answered from just the questions with our very simple QA only model. In order to ensure that we don't overfit while finding these biased questions from training set simply use the predictions of our untrained QA-only model. That is we simply use our best pre-trained word2vec (which is trained on train+val movies in the dataset) in our simple QA-only model and remove questions it can correctly answer without training. These questions which our model gets correct are the really easy one and the and the most biased ones and most of which any other other QA only model can as well easily answer correctly.

As per the table 3 we achieve around 40% accuracy on train and val sets without training and so we propose to create a unbiased dataset by removing these 40% QAs from the original QA dataset for the video category. We show that the remaining QAs are much more difficult for even other QA-only models. In table 2 we show the performance of our QA-only model and that of the baseline QA-only model proposed in the recently released TVQA [12] dataset.

The TVQA QA only baseline model is based on the context matching module followed by an LSTM layer, unlike our QA only mode it avoids average pooling of the embeddings of words to get sentence level embeddings. The context matching module is based on the context query attention layer [19, 26], which takes as input the word embeddings of every word in the questions as the context vector and word embeddings of every words in answers as the query vectors and produces context-aware query vectors (answer-aware question embedding) which are like attention weights. The question word level embeddings, answer word level embedding and the answer-aware question embedding element wise multiplied with question word level embeddings are then concatenated and passed through an LSTM layer and the outputs are temporally max pooled. The pooled features are then passed through a linear layer to produce the final softmax scores for the 5 answer choices.

Type	Our model	TVQA baseline [12]
Original dataset	49.88%	32.50%
Only biased	99.41%	47.80%
Only unbiased	25.68%	22.50%
New augmented	31.82*%	28.44*%

Table 2: Comparison of performance on different splits of MovieQA dataset

Further we also tried an another quick fix of instead of removing these biased QAs we augmented them. That is for a given question to find the top 4 nearest neighbours in the word embedding space of the correct answer (from the list of all the answer choices across the different questions of the same movie) and replace the 4 incorrect options with these top 4 nearest neighbours. Thus we create a new set of QAs which consists of unbiased QAs plus biased QAs which are augmented by the nearest neighbours to reduce the bias. We call this as the augmented dataset. The new answer choices with nearest neighbour approach looks less meaningful options to humans and are still easier then the unbiased QAs but are difficult then original biased QAs. Table 2 compares performance when QA only models are trained and tested with the original dataset, only the biased subset, only the unbiased subset and new augmented dataset.

We show a very simple model which outperforms all the existing video based complex models which supposedly are utilizing videos, on 4 out of 5 categories by predicting answers by just looking at questions. Because of the language biases, the the usefulness of videos by the models can be very questionable.

We suggest a good evaluation criteria when proposing a new video based QA models would be to check the performance of there model after turning off the video input. The accuracy of this simplified QA only model (which learns to predict answer by just looking at the question) would serve as the appropriate baseline for that model. The delta in the performance of the simplified version and the full version of the same model would be a better evaluation criteria for the model. In the TVQA dataset [12] the authors show such an example by showing the performance of there QA only baseline model after turning off the video and subtitle streams from there proposed video+subtitle based QA model. Additionally training and evaluating models on the unbiased QA subset can as well be a better evaluation measure.

Word2vec type	Movie plots for training w2v	Train accuracy (w/o training)	Train accuracy (stopping epoch)	Val accuracy (w/o training)	Val accuracy (stopping epoch)
MovieQA w2v [20]	General + train + val	27.70%	41.67%	26.74%	38.71%
Google w2v [15]	Google News	17.84%	30.40%	14.56%	20.31%
Ours	Val	20.30%	24.43%	40.51%	41.98%
Ours	Train	40.19%	57.46%	18.39%	19.30%
Ours	Train + val	39.90%	51.64%	38.48%	49.88%
Ours	General	21.34%	21.44%	17.17%	18.17%
Ours	General + val	21.31%	27.26%	34.76%	36.11%
Ours	General + train	36.77%	55.33%	16.59%	19.63%
Ours	General + train + val	36.01%	54.40%	32.73%	41.53%

Table 3: Experiments with QA only model (for movies+subtitle task) with different data used for training word2vec. This table shows the importance of different word embeddings. Generic word embedding like Google's gives really bad accuracy. And using a good word embedding can give really high accuracy even without training the QA only model. Highest accuracy is achieved when we use Our(train + val) word2vec.

4. Experiments

4.1. Dataset

The purpose of MovieQA dataset is for building models which can understand stories both in videos and text through the task of question-answering. The dataset is built from texts and videos clips taken from commercial movies. The task here is to look at the story and given a question and 5 possible choices pick the correct one. The story can be in the form of text or video clips, and based on this the dataset is divided into text based QA tasks and video based QA task. The full dataset consists of 14944 QAs taken from 408 movies. For the video based task in which movie clips and the corresponding movie subtitles are the story there 6462 QAs taken from 140 movies and are the main focus of our work. For the text based task there are multiple kind of text sources which are the stories and include subtitles, movie scripts, movie plots and DVS. Each of this have associated QAs. Subtitles and movie plots are available for all the movies and hence they can be used to answer all the QAs in the dataset. The difficulty of question ranges from simple ones based ”what” to ones which require high level of understanding and reasoning like ”why”. And as per the authors some can be answered by just text, some by just videos and some require combination of both.

4.2. Evaluation

Leaderboard submission	Movie: Video+Subtitles
Ours QA only model	46.98%
Multi-modal End-to-end Memory Network (no details)	43.08%
Multimodal dual attention memory [11]	41.41%

Table 4: Leaderboard: Video+Subtitles

Leaderboard submission	Subtitles only
Our QA only model	44.01%
Speaker Naming in Movies [2]	39.36%

Table 5: Leaderboard: Subtitles only

Leaderboard submission	DVS only
Our QA only model	49.65%
MovieQA benchmark [20]	35.09%

Table 6: Leaderboard: DVS only

Leaderboard submission	Scripts only
Our QA only model	45.49%
Read Write Memory Network [16]	39.36%

Table 7: Leaderboard: Scripts only

The dataset is divided into train, val and test splits. The ground truth answers to QAs in test set are not released and one has to submit the predictions on the server for evaluation. Since the test server submissions are limited to once every 3 days, we follow the standard practice and do all our ablation experiments on the validation set. The train set is further divides into train (90%) and dev set (10%), the later is used for hyper-parameter tuning. All the splits are movie specific.

4.3. Importance of word embeddings

Since questions and answers both are text, choosing the right word embedding matters the most as per our analysis on this dataset, and that's what can really exploit the bias. We experimented with word2vec trained on different data - 1) Google w2v (trained on 100 billion words from Google News dataset, has vocab of 3 million words) 2)MovieQA W2V (provided by the authors ,which is trained on about 1400 movie plot synopses, this includes plots for all 408 movies in the MovieQA dataset) 3) Our different versions of w2v trained on different amount of movie plots.

Since the questions and answers are based on movies, they contain lots of movie specific vocabulary. It's really important for models to know movie characters and movie specific vocabulary to even understand the questions. For example, given a question related to Quidditch, the model should have some sense about what it is. That it's from the movie Harry Potter. That's where unsupervised word embedding trained on appropriate data comes comes into picture.

If one uses a general word embedding trained on huge datasets like the Google's W2V it might not have the movie specific words in it's vocabulary, typically the out of vocabulary words are initialized as random vectors, due to which they totally loses the connection to the movie. And also even if the movie related words are there in the vocabulary, all the movie related words of the same movie would most likely be grouped tightly in the word embedding space, which means the models won't be able to differentiate the fine grained importance of the words. The missing words in vocabulary of general word embeddings are mostly character names and other movie specific terms, hence if these words occur in question-answers the model would have very poor understanding of the question themselves, forget the story. So it's important to train word embedding with movie related content especially for small datasets like MovieQA wherein the models cannot learn directly from the raw data. Good word embeddings hence can provide the models with general knowledge about the movie related entities essential to learn from a small dataset.

Table 3 shows the performance of different pre-trained word2vecs when we are just using our simple question only mode to predict the answers. It shows the accuracy of validation and train sets without training of our QA only model (just using the pre-trained word2vec) and after training our simple QA model. As can be seen in the second row of the table 3 Google w2v performs really bad and gives close to chance accuracy (20.31%), primarily because the movie specific words are missing in it's in vocabulary. MovieQA's word2vec itself gives about 38.71% accuracy (first row)on the val set after training. Since the MovieQA word2vec is trained on about 1400 movie plots, majority of them are from outside the movies in the dataset, and hence they are also remembering the movie specific knowledge of movies which are not in the dataset, thereby reducing there efficiency.

We find that using less and just the relevant data gives the best performance of the word embeddings and hence we show experiments with different subsets of these movie plots as an abalation study. In table 3 the terminology is 'val' is word2vec trained on movie plots of the movies in val set, 'train' on movies in train set, and 'general' are the 1400 movie plots minus those in train and val movies. It's interesting to see that accuracy of our simple QA only model without even training on the train and val sets. When we use train word2vec, we get 41.19 % on train set and close to chance accuracy in val set and opposite trend with val word2vec. We found the best model to be the 'train+val' word2vec which gives the highest accuracy. Adding 'general' movie plots, degrades the performance and we get results and trend similar to that of MovieQA (as these both word2vec are trained on similar data, but with different hyper-parameters). So to summarize we found a word2vec trained on just movies of which the questions are there in the train and val set (if we were evaluating our model on test set then we would additionally include movie plots of test set as well) gives us the highest accuracy and is able to fully exploit the bias in the dataset. This is because adding more data means we are reducing the fine grained movie specific clusters in word2vec space. So training word2vec on less data is better as it gives more fine grained word embeddings for the movies.

5. Conclusion

We show that the MovieQA dataset has language bias and present a simple QA only model that exploits it. We train it in unsupervised manner on movie plots and achieve state of the art performance on four of the five categories on the leaderboard at the time of submission. These language biases make it harder to analyze the effect of visual input in existing state of the art models. To mitigate this we propose a simple fix of removing the QAs which our simple QA-only model gets correct. We believe that this unbiased QAs could provide for a better evaluation metric for video based models.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015.

[2] M. Azab, M. Wang, M. Smith, N. Kojima, J. Deng, and R. Mihalcea. Speaker Naming in Movies. *ArXiv e-prints*, Sept. 2018.

[3] M. Blohm, G. Jagfeld, E. Sood, X. Yu, and N. T. Vu. Comparing attention-based convolutional and recurrent neural networks: Success and limitations in machine reading comprehension. *arXiv preprint arXiv:1808.08744*, 2018.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[5] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *CoRR*, abs/1612.00837, 2016.

[6] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340, 2015.

[7] Q. Huang, Y. Xiong, Y. Xiong, Y. Zhang, and D. Lin. From Trailers to Storylines: An Efficient Way to Learn from Movies. *ArXiv e-prints*, June 2018.

[8] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim. TGIF-QA: toward spatio-temporal reasoning in visual question answering. *CoRR*, abs/1704.04497, 2017.

[9] D. Kaushik and Z. C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. *CoRR*, abs/1808.04926, 2018.

[10] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

[11] S.-H. . K. J.-H. . Z. B.-T. Kim, Kyung-Min Choi. Multimodal dual attention memory for video story question answering. In *15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV. 698-713. 10.1007/978-3-030-01267-0*$_41, September 2018.

[12] J. Lei, L. Yu, M. Bansal, and T. L. Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

[13] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[14] T. Maharaj, N. Ballas, A. Rohrbach, A. C. Courville, and C. J. Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[15] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[16] S. Na, S. Lee, J. Kim, and G. Kim. A read-write memory network for movie story understanding. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[17] A. Rohrbach, M. Rohrbach, S. Tang, S. J. Oh, and B. Schiele. Generating descriptions with grounded and co-referenced people. *CoRR*, abs/1704.01518, 2017.

[18] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. Movie description. *International Journal of Computer Vision*, 2017.

[19] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *CoRR*, abs/1611.01603, 2016.

[20] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[21] A. Torabi, N. Tandon, and L. Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint*, 2016.

[22] S. Venugopalan, L. A. Hendricks, R. J. Mooney, and K. Saenko. Improving lstm-based video description with linguistic knowledge mined from text. *CoRR*, abs/1604.01729, 2016.

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014.

[24] B. Wang, Y. Xu, Y. Han, and R. Hong. Movie question answering: Remembering the textual cues for layered visual contents. In *AAAI*, 2018.

[25] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[26] A. W. Yu, D. Dohan, M. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *CoRR*, abs/1804.09541, 2018.

[27] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. *CoRR*, abs/1512.02167, 2015.