Adversarial Multimodal Network for Movie Question Answering

Zhaoquan Yuan[*1], **Siyuan Sun**[*2], **Lixin Duan**[†2], **Xiao Wu**[1], **Changsheng Xu**[3]

[1] School of Information Science and Technology, Southwest Jiaotong University
[2] Big Data Research Center, University of Electronic Science and Technology of China
[3] National Lab of Pattern Recognition, Institute of Automation, CAS

zqyuan@swjtu.edu.cn, jacobsun1996@gmail.com,
lxduan@uestc.edu.cn, wuxiaohk@swjtu.edu.cn, csxu@nlpr.ia.ac.cn

Abstract

Visual question answering by using information from multiple modalities has attracted more and more attention in recent years. However, it is a very challenging task, as the visual content and natural language have quite different statistical properties. In this work, we present a method called Adversarial Multimodal Network (AMN) to better understand video stories for question answering. In AMN, as inspired by generative adversarial networks, we propose to learn multimodal feature representations by finding a more coherent subspace for video clips and the corresponding texts (e.g., subtitles and questions). Moreover, we introduce a self-attention mechanism to enforce the so-called consistency constraints in order to preserve the self-correlation of visual cues of the original video clips in the learned multimodal representations. Extensive experiments on the MovieQA dataset show the effectiveness of our proposed AMN over other published state-of-the-art methods.

1 Introduction

With the recent development of deep learning techniques, tasks related to computer vision and natural language processing (such as object detection, text classification, machine translation, etc.) have been rapidly advanced. As a research topic where both computer vision and natural language processing play important roles, Visual Question Answering (VQA) has attracted a lot of attention in the research community. Existing VQA methods have shown that understanding video stories through only visual cues is a very hard problem [Agrawal *et al.*, 2017; Wang *et al.*, 2018b; Mun *et al.*, 2017; Zhao *et al.*, 2018]. Since nowadays videos often come with rich textual information (e.g., descriptions, subtitles, etc.), it becomes increasingly interesting to learn from those multiple modalities for VQA [Tapaswi *et al.*, 2016; Kembhavi *et al.*, 2017]. However, it is still a challenging task, since the visual content has quite different statistical proper-

---

The superscripts [*] and [†] indicate equal contribution and corresponding author, respectively.

Figure 1: Illustration of the multimodal movie question answering task. Given the question (red box), the goal is to infer the correct answer from multiple candidate choices (blue box), based on the information of visual clips and subtitles (green box).

ties from the natural language. How to bridge such a gap is the goal of this multimodal VQA challenge.

This work focuses on multimodal movie question answering (MovieQA) [Tapaswi *et al.*, 2016], which requires the agent to understand the story of a movie based on visual clips and subtitles to infer the correct answer from multiple candidate choices, as shown in Fig. 1. Compared with typical multimodal VQA, the problem of cross-modal gap gets more serious in the context of multimodal MovieQA. Firstly, different movies may have different background, sheme, and shooting style, which makes it difficult for learning robust multimodal representations [Wang *et al.*, 2018a]. Secondly, it is common that visual clips and textual subtitles are not aligned in the time coordinate. For many video clips, the semantics from different modalities may have discrepancies at some time point, and story cues are usually hidden in different points of time in the longstanding movie.

There are three kinds of methods combining visual clips and subtitles for multimodal MovieQA. The first one is based on embedding mapping, which learns a visual embedding matrix and takes the sum of visual embedded representations and textual features as the combined multimodal representations [Kim *et al.*, 2017]. The second kind of methods are based on attention mechanism, which attend to the textual memory of the words and sentences for each visual regional features [Tapaswi *et al.*, 2016; Wang *et al.*, 2018a]. The third kind methods are based on compact bilinear pooling [Fukui *et*

al., 2016], which compute a compact analogous outer product between visual features and textual features to get a joint representation [Na *et al.*, 2017]. Among the methods mentioned above, the attention-based models achieve better performance, which project visual representations into textual space [Wang *et al.*, 2018a]. We think this benefits from the textual forms of the combined representations. As questions and answers are textual and the video-subtitle integrated representations must interact with them in the end, projecting visual clips into the language embedding space could make them coherent with the textual questions and answers. However, they mainly concern on which parts the model should attend over visual parts, while ignore the characteristics of learned multimodal representation distribution and the issue of information retaining about story cues. The challenges are still far from being solved.

In this paper, we solve the problem of the multimodal movie question answering from a different perspective. We consider the multimodal representation learning from visual clips and subtitles as a process of data generation, during which robust multimodal features are obtained and the core story cues are preserved. The Generative adversarial network (GAN) [Goodfellow *et al.*, 2014] is a powerful framework for data generation and achieves great successes in related fields, which provides a way to generate the multimodal representations for movies.

Based on above ideas, we propose a novel Adversarial Multimodal Network (AMN) model for the multimodal MovieQA task. We place the process of multimodal representation learning in the framework of GAN. Specifically, for the generator, visual clips and subtitles are projected into multimodal representation features, which is composed of two layered form of attention mechanisms in the frame level and clip level separately. For the discriminator, it is encouraged to distinguish the learned multimodal representations from the textual features (e.g., subtitles and questions). The generator and discriminator constitute a couple of adversarial modules, which help us find a more coherent subspace for video clips and the corresponding texts. Moreover, in order to preserve the correlation from the story cues, we introduce the self-attention mechanism to enforce a consistency constraints on the learned multimodal representation.

The contributions of our work can be summarized as follows:

- We propose a novel Adversarial Multimodal Network (AMN) model for MovieQA. To the best of our knowledge, AMN is the first work to introduce the generative adversarial framework for multimodal question answering.

- Different from existing methods, we take the multimodal representation learning as a data generation process, where an adversarial multimodal representation learning module and a self-attention mechanism to ensure to satisfy consistency constraints are introduced. This obtains better multimodal representations in the scenario of multimodal movie question answering.

- We conduct comprehensive experiments on the public MovieQA dataset [Tapaswi *et al.*, 2016] to validate the effectiveness of our work, and it achieves the state-of-art performance among the published models.

2 Related Work

2.1 Visual Question Answering

Visual Question Answering (VQA) has received increasing attention from both the computer vision and the natural language processing communities, and several datasets are developed [Agrawal *et al.*, 2017; Malinowski and Fritz, 2014]. Some comprehensive surveys about VQA are introduced by [Wu *et al.*, 2017; Pandhre and Sodhani, 2017]. Typically, [Malinowski *et al.*, 2015] uses neural networks to learn joint embeddings of images and sentences into a common feature space, where further reasoning over both modalities together is performed. [Zhu *et al.*, 2016] utilizes attention mechanism to model interactions between the question and specific regions of these feature maps. [Andreas *et al.*, 2016b; 2016a] build compositional model architectures consisting of distinct modules designed for specific desired capabilities, where structures of questions and images. [Wang *et al.*, 2015] introduces external knowledge-bases to reinforce the content of structured knowledge and the ability of reasoning.

2.2 Multimodal Question Answering

Multimodal question answering integrates different modal data source to infer correct answers for questions. The typical scenarios include movie or TV question answering with video clips and subtitles for story comprehension [Tapaswi *et al.*, 2016], textbook question answering with diagrams and languages for lessons understanding [Kembhavi *et al.*, 2017], and visual dialogue which requires an agent to hold a meaningful dialog with humans in natural, conversational language about visual content. Specifically, given an image, a dialog history, and a follow-up question about the image, the task is to answer the question [Das *et al.*, 2017].

Movie question answering this paper focus on has attracted research interests and some efforts have been conducted for this task. [Tapaswi *et al.*, 2016] introduces the movie question answering (MovieQA) dataset and proposes a baseline method based on the End-to-end Memory Network. [Kim *et al.*, 2017] develops a deep embedded memory networks for video-story learning, where an LSTM-based attention model uses the long-term memory to recall the best question-story-answer triplet by focusing on specific words containing key information. [Na *et al.*, 2017] provides a read-write memory network where the read network and write network consist of multiple convolutional layers, which enable memory read and write operations to have high capacity and flexibility. [Wang *et al.*, 2018a] focuses on the video representation, and puts forward a layered memory network to represent frame-level and clip-level movie content by a static word memory module and another dynamic subtitle memory module respectively.

3 Adversarial Multimodal Network

In this section, we introduce the proposed novel adversarial multimodal network (AMN) model combining the visual contents and subtitles for the MovieQA task.
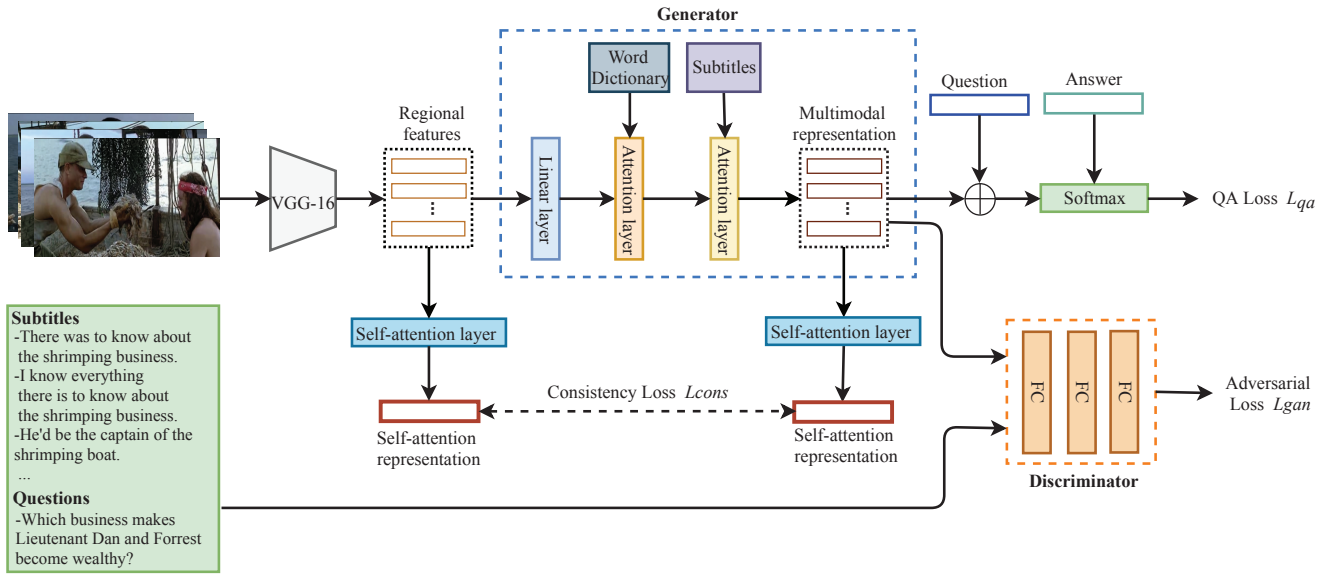
Figure 2: The framework of our proposed AMN method for multimodal movie question answering.

Given video clips, the corresponding subtitles, and a natural language question about the movie, the task is to predict an accurate natural language answer from multiple choices. Our basic idea is to learn multimodal representations for movie clips by projecting features of visual clips into textual spaces, and place the task in the framework of GAN. In the framework, the generator maps visual clip features into multimodal representations which is expected indistinguishable from subtitles and questions, while the discriminator is encouraged to distinguish them. Correct answers are inferred based on the learned multimodal representation and corresponding questions and candidate answers. Meanwhile, to preserve the information of story cues during the projecting, a self-attention mechanism to enforce a consistency constraints on the learned multimodal representation.

The framework of our proposed method is illustrated in Fig. 2. Our formulation contains three kinds of terms: adversarial multimodal representation learning, self-attention based consistency constraints, and answer inference. For convenience, the main symbols used in this paper are summarized in Table 1.

3.1 Adversarial Multimodal Representation Learning

We consider the multimodal representation learning process from visual clips and subtitles to the multimodal features as a mapping function $G : (\mathbf{V}, \mathbf{S}) \rightarrow \mathbf{H}$, which takes the visual feature maps $\mathbf{V}$ extracted by pre-trained CNN network and textual subtitle $\mathbf{S}$ computed based on word dictionary $\mathbf{W}_e$ as input, and output the adversarial multimodal representations $\mathbf{H}$.

We are looking forward to the $\mathbf{H}$ coherent with textual questions and answers. Inspired with the idea of generative adversarial network (GAN), we use its framework for this problem. We take the mapping function G as the generator in GAN, and consider another module D as the discriminator.

For the generator, we encourage its output $\mathbf{H}$ indistinguishable from the coherent subspace of the corresponding texts. For the discriminator, it is encouraged to distinguish them. The objective is expressed as:

$$\mathcal{L}_{gan}(G, D) = \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})}[\log D(\mathbf{z})] + \\ \mathbb{E}_{\mathbf{v} \sim p_v(\mathbf{v})}[\log(1 - D(G(\mathbf{v})))]. \quad (1)$$

where $\mathbf{z}$ denotes real textual data. In more detail, we use subtitles and questions as real textual data, and a multiple neural network is utilized as the discriminator.

As for the generator, we adopt a two layered attention learning process, similarly as in [Wang *et al.*, 2018a]. In the first attention layer, we attend the $H \times W$ regional features to a word dictionary. For each regional features $\mathbf{v}_{tj}$ in frame $\mathbf{I}_t$, we first map it into the word space through a mapping matrix $\mathbf{W}_l$ linearly:

$$\mathbf{o}_{tj} = \mathbf{W}_l \mathbf{v}_{tj}. \quad (2)$$

Then we attend to a word embedded memory $\mathbf{W}_e$ for $\mathbf{o}_{tj} \in \mathbf{R}^d$. The result is:

$$\mathbf{e}_{tj} = \sum_{k=1}^{|\mathcal{V}|} \mathbf{o}_{tj}^T \mathbf{W}_{e,k} \mathbf{W}_{e,k}. \quad (3)$$

where the $\mathbf{W}_{e,k}$ is the k-th row vector of the word memory, and the frame features are computed by sum the regional features $\mathbf{u}_t = \sum_{j=1}^{H \times W} \mathbf{e}_{tj}$.

In the second attention layer, we attend to embedded subtitle features $\{\mathbf{s}_1, ..., \mathbf{s}_K\}$ for $\mathbf{u}_t$:

$$\mathbf{h}_t = \sum_{k=1}^{K} \mathbf{u}_t^T \mathbf{s}_k \mathbf{s}_k. \quad (4)$$

Since clip features are computed according to both visual contents and textual subtitles, the generated $\mathbf{H} =$

Table 1: Summary of the symbols

Symbol	Description
C, H, W	The channel, height, and width of feature maps respectively.
d	The word embedding dimension.
$\mathbf{W}_e$	Word dictionary with $\mathbf{W}_e \in \mathbf{R}^{|\mathcal{V}| \times d}$, where $|W|$ denotes the size of word dictionary.
$\mathbf{V}$	Video clips with T feature frames $\{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_t, ..., \mathbf{I}_T\}$, where $I_t \in \mathbf{R}^{C \times H \times W}$.
$\mathbf{W}_l$	Projecting matrix with $\mathbf{W}_l \in \mathbf{R}^{d \times C}$.
$\mathbf{v}_{tj}$	the j-th projected regional feature of t-th movie frame with $\mathbf{v}_{tj} \in \mathbf{R}^C$, where $t \in \{1, ..., T\}$ and $j \in \{1, ..., H \times W\}$.
$\mathbf{S}$	Subtitle features of the movie clip, where $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, ...\mathbf{s}_K\} \in \mathbf{R}^{K \times d}$, and K is the number of sentences of corresponding subtitles.
$\mathbf{q}, \mathbf{a}_n, \widehat{\mathbf{a}}$	$\mathbf{q}$ denotes features of the question, $\mathbf{a}_n$ is the features of the n-th candidate answer, and $n \in \{1, .., 5\}$. $\widehat{\mathbf{a}}$ denotes the predicted answer of our model.
$\mathbf{H}$	The Learned multimodal features $\mathbf{H} = \{\mathbf{h}_1, ..., \mathbf{h}_T\} \in \mathbf{R}^{T \times d}$, where $t \in \{1, ..., T\}$.
α	Intra-clip relationship matrix and $\alpha \in \mathbf{R}^{T \times T}$.
$\mathbf{g}, \mathbf{c}$	Self-attention representations of the visual clip and multimodal representational features respectively.
$\mathcal{L}_{qa}$	Loss function between the predicted answer and the ground-truth answer.
$\mathcal{L}_{gan}$	Loss function for the generative adversarial network.
$\mathcal{L}_{cons}$	Loss function between the representation after generator module and the visual representation before generator module.
G, D	Generator and discriminator of the GAN.

$\{\mathbf{h}_1, ..., \mathbf{h}_T\}$ from G can be considered as the learned multimodal representations for movies.

3.2 Self-attention based Consistency Constraints

Since the projection process above may loss information about the movie, especially for the story cues, we propose a consistency constraint for the projection and try to reconstruct the representations of video clips from the learned multimodal representations. However, it is difficult, for the dimension of the elements of original visual representation $\mathbf{V}$ is different from that of $\mathbf{H}$.

To hand this problem, we propose to utilize a self-attention mechanism [Vaswani *et al.*, 2017] to map them into a fixed dimensional space. Apart from the issue of dimensionality, we think that the self-attention also reflects the intrinsic relationship and captures the long time dependency inside the movie, which is important for long story comprehension. If the results of the self-attention of them are closed to each other, the whole story cue is retained during the two layered projections. For the $\mathbf{V}$, we first reduce its dimension by meaning its regional features $\mathbf{V}$ to obtain $\mathbf{V}' \in \mathbf{R}^{T \times c}$. Then compute the interactions among the T features, which capture relationship

between T features in movie:

$$\alpha = \mathbf{V}' \mathbf{V}'^\top. \tag{5}$$

Based on the intra-clip relationship matrix α, the final self-attention representation computed by $\mathbf{g} = \sum_{t=1}^T \alpha_t$, where α_t denotes the t-th row of the matrix α. Likewise, this self-attention mechanism is also carried out for $\mathbf{H}$ except that there is no mean operation, for $\mathbf{H}$ is second order tensor itself, and we obtain the self-attention representation $\mathbf{c}$.

To make the self-attention representations $\mathbf{g}$ and $\mathbf{c}$ as close as possible, we take the *cos* distance as consistency loss:

$$\mathcal{L}_{cons}(G)) = \frac{\mathbf{g} \cdot \mathbf{c}}{\|\mathbf{g}\|\|\mathbf{c}\|}. \tag{6}$$

In preliminary experiments, we also try replacing the *cos* distance in this loss with an L_2 loss between $\mathbf{g}$ and $\mathbf{c}$, but do not observe improved performance.

3.3 Answer Inference

Based on the learned multimodal representations $\mathbf{H}$, the features of clips which are used for movie question answering are computed by: $\mathbf{x} = \sum_{j=1}^T \mathbf{h}_t$. Then we use the method from [Tapaswi *et al.*, 2016] to answer the movie questions with open-ended answers as follows:

$$\widehat{\mathbf{a}} = \mathrm{softmax}\left((\mathbf{x} + \mathbf{q})^\top \mathbf{a}\right). \tag{7}$$

We minimize the cross-entropy loss for answer inference:

$$\mathcal{L}_{qa}(G) = -\sum_{k=1}^n I\left(k, t_{correct}\right) \log \widehat{\mathbf{a}}_k. \tag{8}$$

where $t_{correct}$ is the label of the ground-truth correct answer, I denotes the indicative function, and there are 5 candidate answer choices and only one is correct (i.e., $n = 5$).

So far, the movie representation $\mathbf{x}$ is only based on visual contents and subtitles and has nothing to do with questions and candidate answers, which makes it difficult to discover the required relevant information to questions and answers. Also, there are some irrelative information with clips in subtitles, which should be removed. To solve this problem, we update subtitles with $\mathbf{q}$, $\mathbf{a}$, and $\mathbf{x}$:

$$\begin{cases} \beta_n^{(i)} &= \mathrm{ReLU}\left((\mathbf{x}^{(i)} + \mathbf{a} + \lambda\mathbf{q})^T \mathbf{s}_n^{(i)}\right), \\ \mathbf{s}_n^{(i+1)} &= \beta_n^{(i)} \mathbf{s}_n^{(i)}. \end{cases} \tag{9}$$

where $\mathbf{s}_n^{(i)}$ donotes the n-th sentences of the the i-th subtitle representations, and λ is a tradeoff hyper-parameter between the question and the rest of the query with the value of 0.45 in our model. Note that previous work [Wang *et al.*, 2018a] only utilizes $\mathbf{x}$ to update subtitles to get rid of the irrelative information with $\mathbf{x}$, which is different from our method.

3.4 Overall Formulation

Based on the above, our goal is to solve the following optimization problem:

$$\min_{G} \max_{D} \ \mathcal{L}(G, D). \tag{10}$$

where $\mathcal{L}(G, D) = \mathcal{L}_{gan}(G, D) + \mathcal{L}_{cons}(G) + \mathcal{L}_{qa}(G)$. By iteratively updating G and D in Eq. (10), we can obtain the final solution:

$$G^*, D^* = \arg \min_{G} \max_{D} \mathcal{L}(G, D). \tag{11}$$

In the learning process, stochastic gradient descent is used for optimization.

4 Experiments

We conduct extensive experiments on the MovieQA dataset [Tapaswi *et al.*, 2016] by comparing our proposed AMN method with existing state-of-the-art, such as SSCB [Tapaswi *et al.*, 2016], MemN2N [Tapaswi *et al.*, 2016], Read-Writer [Na *et al.*, 2017], DeepStory [Kim *et al.*, 2017], FVTA [Liang *et al.*, 2018], LMN [Wang *et al.*, 2018a] and so on.

4.1 Dataset Description

The MovieQA dataset [Tapaswi *et al.*, 2016] is composed of $14,944$ multiple-choice questions about 408 movies with high semantic diversity. It contains diverse sources of information, including video clips, plots, subtitles, scripts and DVS transcriptions. And for each question, There are *five* candidate answers where only one of them is correct. Similar to [Wang *et al.*, 2018a], we focus on the "Video+Subtitles" task in this work, which leaves us with $6,462$ question-answer pairs from 140 movies in total.

4.2 Experimental Setup

We strictly follow the same setup as in [Tapaswi *et al.*, 2016; Wang *et al.*, 2018a]. To be more specific, the $6,462$ question-answer pairs are officially split into three sets with $4,318$ for training, 886 for validation and $1,258$ for test. Since the official test set ($1,258$ pairs) can only be tested once per 72 hours in an online evaluation fashion, again following [Tapaswi *et al.*, 2016; Wang *et al.*, 2018a], we also adopt another set of experiments for performance evaluation by first dividing the official training set ($4,318$ pairs) into 90% for training and 10% for development, and then testing against the official validation set (886 pairs).

Moreover, for each video clip, we first extract 32 frames, as done in [Wang *et al.*, 2018a]. And the regional feature of each frame is obtained by extracting from the "pool5" layer of the ImageNet-based VGG-16 model, which is a tensor of size $512 \times 7 \times 7$. As a result, each video clip is represented as a tensor of size $32 \times 512 \times 7 \times 7$. For the static word dictionary, we have a vocabulary size of $26,630$, and each word in the dictionary is represented as a 300-dim feature vector, based on the word2vec model supported by [Tapaswi *et al.*, 2016; Mikolov *et al.*, 2013].

For the discriminator network in AMN, we employ a three-layer neural network to differentiate bewteen the real text

Table 2: Results of different methods on the official validation and test sets of the MovieQA dataset.

Methods	Accuracy on Val (%)	Accuracy on Test (%)
SSCB [Tapaswi *et al.*, 2016]	34.20	-
Read-Writer [Na *et al.*, 2017]	38.67	36.25
DeepStory [Kim *et al.*, 2017]	44.70	29.97
FVTA [Liang *et al.*, 2018]	41.00	37.30
LMN [Wang *et al.*, 2018a]	42.50	39.03
AMN	**45.37**	**45.31**

Table 3: Ablation study of our AMN method on the official validation and test sets of the MovieQA dataset.

Methods	Accuracy on Val (%)	Accuracy on Test (%)
AMN_{deg}	42.66	42.45
AMN_{cons}	43.68	43.56
AMN_{adv}	44.13	44.12
AMN	**45.37**	**45.31**

data (i.e., subtitles and questions) and the learned multimodal representations (i.e., $\mathbf{H}$). In the discriminator network, each layer contains 50 hidden nodes and ReLU is used for activation function. For the training of our overall AMN model, we set the batch size as 8 and the learning rate as 0.01. We train our model up to 100 epochs.

For performance evaluation of different methods, we use accuracy as the metric throughout the experiments. Since this work adopts the same experimental protocol for all the methods, we directly take the reported results of existing baselines in their original papers.

4.3 Results

We compare our proposed AMN method with several state-of-the-art on the MovieQA dataset for the multimodal video question answering task. It is worth pointing out that Deep-Story [Kim *et al.*, 2017] sums up the visual embeddings and textual features as the final multimodal representation. And Reader-Writer [Na *et al.*, 2017] is built based on multi-layer CNNs in order to capture and store the sequential information of movies into the memory. FVTA [Liang *et al.*, 2018] is a focal visual-text attention network which is able to capture correlation between visual and textual sequences. Moreover, SSCB [Tapaswi *et al.*, 2016] and LMN [Wang *et al.*, 2018a] adopt the attention mechanism but do not consider the adversarial mechanism or the self-attention based consistency based on visual cues as done in our work.

We evaluate different methods and report their accuracies on the official validation and test sets of the MovieQA dataset in Table 2. We can observe from the results that our proposed AMN achieves superior performance than the state-of-the-art like FVTA [Liang *et al.*, 2018] and LMN [Wang *et al.*, 2018a], which demonstrates the effectiveness of the newly developed adversarial multimodal representation learning based on GAN and consistency constraints in our AMN. It is worth mentioning that the accuracies of existing baselines drop considerably on the official test set of the MovieQA
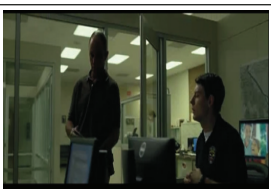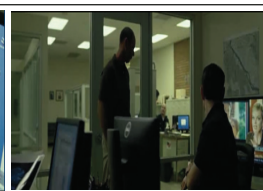
Question:
What happens when the boat gets hit again?
Answers:
1. It gets caught in a nearby tree
2. Everybody falls into the water
3. It gets damaged
4. The engine stalls
5. It is forced over a cliff

Top-3 weight and corresponding subtitles (Our method):
0.000927 : The rope at the front of the boat, can you reach it
0.000783 : I'm not going into the water again
0.000634 : Lee, get out of the water now!

Top-3 weight and corresponding subtitles (LMN [Wang et al., 2018a]):
0.001071 : Get up the tree.' 'Get up the tree!
0.001071 : We're also gonna need the boat.
0.001069 : The rope at the front of the boat, can you reach it

Question:
Why do the police conclude that Amy was murdered?
Answers:
1. Becausethe killer left a confession in the kitchen of their house
2. Because her murder was recorded on a surveillance camera
3. Becausethey found a body that resembled her floating
4. BecauseNick confesses to killing her
5. Because they find remnants of cleaned blood stains in the house

Top-3 weight and corresponding subtitles (Our method):
0.000933 : Amy lost a lot of blood in there
0.000788 : A pool of blood and no body suggests homicide.
0.000724 : Let's find Amy..

Top-3 weight and corresponding subtitles (LMN [Wang et al., 2018a]):
0.000700 : Amy's medical records come in?
0.000645 : Because in April, you bumped up Amy's life insurance...
0.000621 : Because it's the most abnormal situation in the world.

Figure 3: Qualitative results of two test samples based on LMN and our AMN. For both methods, we show movie frames and their corresponding subtitles which have three largest attention weights. The answers/weights in green are obtained by our AMN, while those in red are by LMN. Best viewed in color.

dataset (compared to those on the official validation set), while our AMN has consistent performance on both settings. Such an observation clearly shows the good generalization ability of AMN.

Ablation study. We further provide the ablation study of AMN in Table 3 to analyze the importance of the two proposed components: adversarial multimodal representation learning and self-attention based consistency constraints. We use AMN_{adv}, AMN_{cons} and AMN_{deg} to represent AMN using only adversarial multimodal representation learning, AMN using only self-attention based consistency constraints and AMN using none of them (this degenerated version is almost the same to LMN except that we update subtitles by additionally using answers as in Eq. (9). As the ablation study shown in Table 3, we can see that both AMN_{adv} and AMN_{cons} performs better than AMN_{deg} by a noticeable margin, which shows the usefulness of the two corresponding components of AMN. And the two components are complementary to further boost the overall performance of AMN.

4.4 Qualitative Analysis

Illustration of test samples for multimodal movie question answering. In Fig. 3, we compare our AMN with the most competitive baseline LMN by presenting the qualitative results of a couple of test samples, where we show the answers that are correctly predicted by our AMN but not by LMN [Wang *et al.*, 2018a]. It is clear that our AMN is able to associate specific subtitles with more relevant movie frames and also to find better correspondences between subtitles and question-answer pairs.

Visualization of feature distributions. We further investigate the good performance of AMN in more depth by comparing the distributions of the learned adversarial multimodal
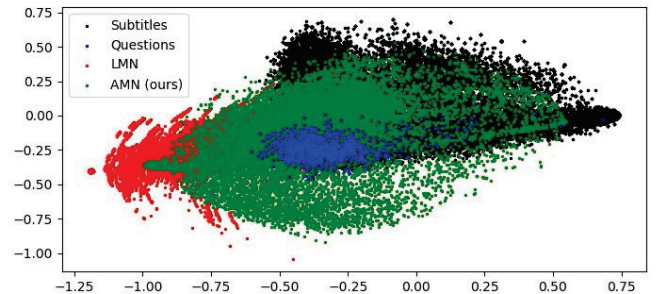

Figure 4: Visualization of feature distributions of subtitles, questions, LMN and our AMN. Best viewed in color.

representations with the learned features from LMN as well as the word2vec representations of the original textual data (i.e., subtitles and questions). Since each feature representation is 300-dim, we use PCA to reduce their dimensionality and project them into a 2D space. As shown in Fig. 4, we can see that the distribution of questions is part of subtitles, and the distribution of learned features from LMN barely overlaps with the text distribution, while there is a relatively large overlap between the feature distributions of our AMN and original texts. This intuitive observation shows that our AMN is promising to learn a better multimodal representation and thus achieves the best performance over other baselines.

5 Conclusion

In this work, we focus on the multimodal movie question answering task. To better understand a movie story based on different modalities (e.g., movie clips and subtitles), we propose a novel method coined as Adversarial Multimodal Network

(AMN) by introducing a newly developed adversarial multi-modal representation learning mechanism based on GAN as well as a self-attention based consistency constraints based on visual cues from movie clips. As demonstrated by the comprehensive experiments on the benchmark MovieQA dataset, our AMN method outperforms recently published state-of-the-art and also shows its good generalization ability.

References

[Agrawal *et al.*, 2017] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. VQA: visual question answering - www.visualqa.org. *International Journal of Computer Vision*, 123(1):4–31, 2017.

[Andreas *et al.*, 2016a] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, 2016.

[Andreas *et al.*, 2016b] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[Das *et al.*, 2017] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[Fukui *et al.*, 2016] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[Kembhavi *et al.*, 2017] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4999–5007, 2017.

[Kim *et al.*, 2017] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*, 2017.

[Liang *et al.*, 2018] Junwei Liang, Lu Jiang, Liangliang Cao, Li-Jia Li, and Alexander G Hauptmann. Focal visual-text attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6135–6143, 2018.

[Malinowski and Fritz, 2014] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014.

[Malinowski *et al.*, 2015] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Mun *et al.*, 2017] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875, 2017.

[Na *et al.*, 2017] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 677–685. IEEE, 2017.

[Pandhre and Sodhani, 2017] Supriya Pandhre and Shagun Sodhani. Survey of recent advances in visual question answering. *arXiv preprint arXiv:1709.08203*, 2017.

[Tapaswi *et al.*, 2016] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[Wang *et al.*, 2015] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015.

[Wang *et al.*, 2018a] Bo Wang, Youjiang Xu, Yahong Han, and Richang Hong. Movie question answering: Remembering the textual cues for layered visual contents. *arXiv preprint arXiv:1804.09412*, 2018.

[Wang *et al.*, 2018b] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2018.

[Wu *et al.*, 2017] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

[Zhao *et al.*, 2018] Zhou Zhao, Xinghua Jiang, Deng Cai, Jun Xiao, Xiaofei He, and Shiliang Pu. Multi-turn video

question answering via multi-stream hierarchical attention context network. In *IJCAI*, pages 3690–3696, 2018.

[Zhu *et al.*, 2016] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004, 2016.