Conditional cross correlation network for video question answering

Kaouther Ouenniche, Ruxandra Tapu, Titus Zaharia

Conditional Cross Correlation Network for Video Question Answering

Kaouther Ouenniche, Ruxandra Tapu, Titus Zaharia

Institut Polytechnique de Paris, Télécom SudParis, Laboratoire SAMOVAR
9 rue Charles Fourier, 91011Evry-Courcouronnes Cedex
Email : {kaouther.ouenniche, ruxandra.tapu, titus.zaharia}@telecom-sudparis.eu

***Abstract* — Video question answering (VideoQA) is the process that aims at responding to questions expressed in natural language, according to the semantic content of a given video. VideoQA is a highly challenging task and demands a comprehensive understanding of the video document, including the recognition of the various objects, actions and activities involved together with the spatial, temporal and causal relations between them. To tackle the challenge of VideoQA, most methods propose efficient techniques to fuse the representations between visual and textual modalities. In this paper, we introduce a novel framework based on a conditional cross-correlation network that learns multimodal contextualization with reduced computational and memory requirements. At the core of our approach, we consider a cross-correlation module designed to learn reciprocally constrained visual/textual features combined with a lightweight transformer that fuses the intermodal contextualization between visual and textual modalities. We test the vulnerability of the composing elements of our pipeline using black box attacks. To this purpose, we automatically generate semantic-preserving rephrased questions. The ablation study conducted confirms the importance of each module in the framework. The experimental evaluation, carried out on the MSVD-QA benchmark, validates the proposed methodology with average accuracy scores of 43.58%. When compared with state-of-the-art methods the proposed method yields gains in accuracy of more than 4%and achieves a 43.58% accuracy rate on the MSVD-QA data set.**

Keywords—video question answering, multimodal learning, cross-correlation.

I. INTRODUCTION

The availability of large, annotated datasets have accelerated the progress of both computer vision and natural language processing methodologies, with a significant impact over a wide field of application domains, including image classification, speech recognition, reading comprehension and action recognition. Such advances have encouraged researchers to develop systems able to provide a holistic understanding of the scene, close to a human-level knowledge. Recently, Video Question Answering (VideoQA) has emerged as a testing ground to push boundaries in both domains. Given a video and an arbitrary question, the goal of a VideoQA model is to extract question-relevant semantic information and to infer an answer close to the ground truth. The task is highly challenging as it requires joint reasoning with both visual and textual elements, while taking into account their corresponding spatial, temporal and causal relations.

The core of multimodal learning lies in effectively combining the representations between multiple modalities in order to leverage the complementary information involved. The fusion is not straightforward as the modalities exhibit different types and levels of detail. In the state of the art, there can be identified three families of techniques for multimodal data fusion. The *early fusion* approaches concatenate multiple modalities from raw or pre-processed data. The output is then directly passed to a classifier to predict the answer. In the case of *late fusion* technique, each modality is independently processed, followed by concatenation in the prediction phase. The *intermediate fusion* methods combine the features of each modality in multiple stages of the model to produce new representations that are more expressive than the original, individual ones. This last, hybrid technique is often more adequate for video question answering purposes and generally achieves superior performances.

Recently, in the natural language processing community (NLP), there has been a paradigm shift from monolithic models with strong inductive biases such as CNNs and RNNs to general architectures based on attention [1]. More specifically, transformers have become the *de facto* standard for NLP task. Inspired by the success of attention mechanisms, there has been a growing interest to explore their potential application in multimodal learning. Recent works show that pre-training a transformer on large datasets followed by fine-tuning outperform previous state of the art methods on various multimodal tasks such as video captioning, visual common sense reasoning and video question answering.

The objective of this paper is to learn grounded joint visual and textual features to predict correctly the answer. Inspired by the success of transformers in AI applications, we use attention bottlenecks to fuse the modalities at multiple layers of the model. Developing and testing the performance of this model is challenging, due to multiple reasons. First, the complex interactions between multimodal data require very deep models. Hongsuck *et al.* [2] present a multimodal network composed of four transformers which learn intra-modal contextualizations through the mechanism of self-attention and inter-modal contextualization by computing question-guided attention over visual features. Similarly, in [3], authors present a heterogeneous memory network which learns the semantic visual and textual features independently through two self-attention-based transformers, then join them in a co-attention transformer. The proposed MERLOT [4] framework extracts the frame features using a CNN-based image encoder and the textual features with the help of a BERT-like transformer. Then, both representations are jointly encoded with a 12-layer transformer.

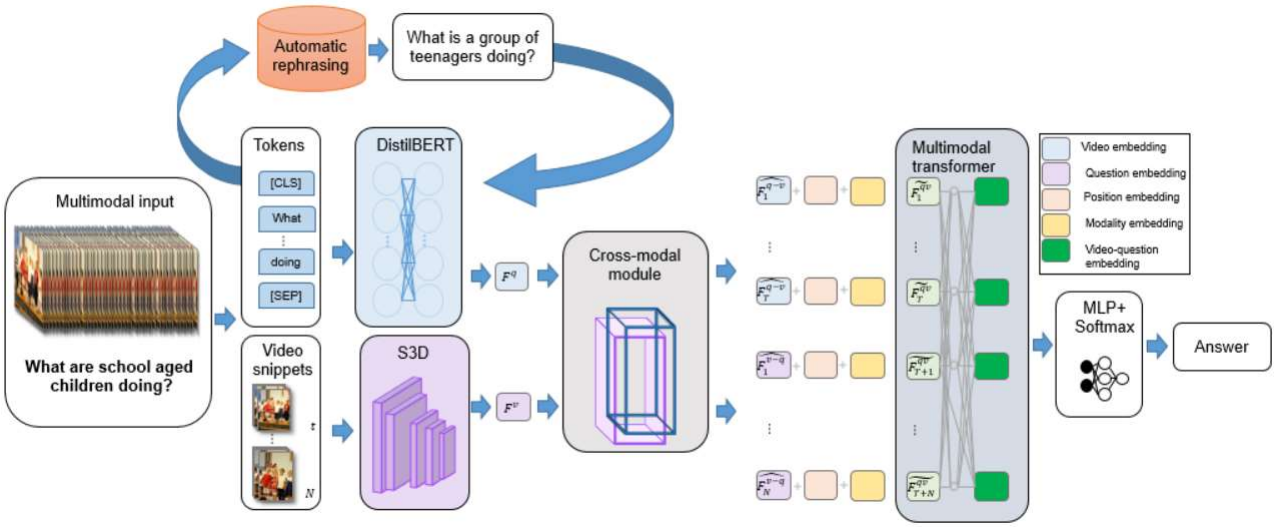

Fig. 1. The proposed framework architecture.

The main drawback of such attention mechanisms is related to the quadratic time and space complexity, which penalizes the wide adoption of a transformer-only fusion network.

To overcome such limitations, we design a novel framework based on a lightweight transformer that runs in conjunction with a cross-modality module. The latter uses cross correlation to reciprocally learn question-conditioned visual features and video-conditioned textual features. We finally feed the obtained constrained representations to a 2-layer transformer that provides the final multi-modal representation (Fig.1).

Generally, transformers require pre-training on large datasets to achieve competitive performance. Most techniques pre-train the model on task-agnostic datasets (usually videos with automatic speech recognition transcripts) then fine-tune on the considered VideoQA task. For example, Hongsuck *et al.* [2] pre-train the model on next utter prediction using the script extracted from HowTo100M [5]. MERLOT [4] uses general objectives (masked language modeling, frame-transcript matching and temporal reordering to pre-train on YT-temporal 180M) that are not designed for video question answering. In contrast with such approaches, we use a task-specific training process to learn more grounded features for the VideoQA task. We use the recently introduced HowToVQA69M data set, with over 69M video-question-answer triplets. To reduce the memory and computational requirements, we randomly sample 164148 training examples and achieve competitive results.

In addition, we test our model on the MSVD-QA [5] publicly available benchmark. Following previous state of the art methods, we have retained the accuracy metric to test the generalization of the network on the test set. However, several studies [7-9] have shown that despite recent advances, current models infer the answer without reasoning, relying instead on superficial correlations (i.e., biases) inherited from the training dataset. One reason of this behavior is the IID (independent and identically distributed) train-test split method. This suggests that models relying on priors during training demonstrate acceptable performance on the set. It is still unclear how VideoQA models perform in real-world situations. For this reason, we generate a novel test dataset to validate the robustness of our framework. More specifically, we apply adversarial attacks in a black-box scenario by distracting the model with rephrased questions (we consider here utilization of synonyms, changes in the word order or question length, various levels of redundancy) that preserve the over-all semantic similarity, automatically generated from a different distribution of the training set. Rephrasing attacks expose the brittleness of VideoQA models to linguistic variations in questions. We apply the attacks to each component of our pipeline and compare their robustness. This is important for investigating linguistic biases in the multimodal capability of VideoQA models (does the model really understand the question?) and for application in real-world scenarios. To the very best of our knowledge, this is a first attempt to investigate the vulnerability of VideoQA to adversarial attacks.

To summarize, in this paper we propose the following contributions: (1) A novel framework combining a cross-modal correlation module with a multimodal fusion transformer designed to model the interactions between spatio-temporal, visual and textual representations. Compared with previous architectures, our system uses an efficient, lightweight model, leveraging only one 2-layer transformer. (2) A more discriminative set of reciprocally conditioned visual and textual features. This is opposed to a simple concatenation of pre-trained embeddings typically used in early fusion strategies. (3) An in-depth analysis of the sensitivity of the proposed framework and associated components with respect to textual rephrasing of the questions that can frequently appear in practice.

The rest of the paper is organized as follows. In section II, we present the related work on video question answering. In section III, we present in details the proposed framework and describe the main modules involved. Finally, in Section IV, we experimentally analyze the related performances and compare them with state of the art methods. We also demonstrate the robustness of the proposed model through an ablation study. Finally, Section V concludes the paper and opens some perspectives of future work.

II. Related work

In recent years, video question answering has attracted much attention and has known an accelerated development. VideoQA is challenging as it requires shared understanding of visual and textual cues to determine the correct answer. The state of the art solutions typically include four main components: video embedding, text embedding, multimodal fusion and answer prediction (using a classifier). For video embedding, existing approaches represent video at the frame level using 2D CNNs, e.g. VGG [10] and ResNet [11], and/or at the clip level using 3D CNNs, e.g. S3D [12] and I3D [32]. Question embedding extracts token-level features using well-known NLP techniques such as GloVe [13] or BERT [14]. For cross-modality fusion, early techniques use monolithic models such as CNNs and RNNs. Zhao *et al*. [15] exploit a hierarchical double attention network to learn question-guided appearance and motion features with the help of Bi-GRU models. Yu *et al*. [16] propose a convolutional hierarchical decoder that computes a compatibility score between the two modalities by recursively evaluating the hidden matches. Monolithic models are relying on attention but are unsuitable to represent long-term dependencies. Another promising research direction for modeling evolving visual-textual interactions concerns the memory networks, which include an artificial memory component that can utilize even early information. Tapaswi *et al.* [17] adapt end-to-end memory networks for video question answering purposes. Kim *et al.*[18] incorporate attention mechanisms to prune out irrelevant temporal information from memory slots. Gao *et al.* [19] propose co-attention dynamic memory network to model appearance and motion interactions. Memory networks refine the answer gradually through multi-step reasoning and achieve competitive performances on relatively long videos.

However, transformer-based networks [2,4,20-23] have recently surpassed in terms of performances such approaches. Within this context, the visual and textual features are extracted using pre-trained (fixed) models, then fused in a multimodal transformer. The framework is pre-trained on large-scale video-text datasets and fine-tuned on downstream tasks (VideoQA, video captioning, video-text retrieval…). The works reported in [2,23] perform intermodal contextualization by computing question-guided attention over visual features and intra-modal interaction through self-attention. Lei *et al.*[20] and Yu *et al.* [21] solve the offline encoder problem by proposing an efficient sampling strategy during training. Zellers *et al.* [4] adopt end-to-end training by leveraging 2D models instead of 3D.

In contrast with such techniques, we propose a cross-correlation technique designed to reduce the heterogeneity between the video and text modalities. The proposed model is trained on a task-specific dataset: HowToVQA 69M [22] using a frozen visual backbone. We make the hypothesis that our model represents grounding features by jointly learning each modality representation under the constraint of the other.

III. Methodology

The synoptic scheme of the proposed method is illustrated in Fig.1. It contains two key components: (1) the cross-modal correlation module, and (2) the multimodal fusion module. As a pre-proccesing step, we start by extracting the video and text embeddings.

A. Pre-processing: feature extraction

To extract the visual representations, the video is uniformly sampled in *N* fixed length clips of 32 frames. We feed each clip to a S3D model, a 3D CNN architecture that aims at learning powerful video representations. The 3D model has been pre-trained on HowTo100M [5] using MIL-NCE technique [25]. We take the feature activations before the final fully connected layer and apply average pooling to obtain a 1024-dimension vector. Finally, a feed forward network (linear projection followed by GeLU activation function [26] and layer normalization) is used to project the feature vector. During training, the S3D model weights are frozen to improve efficiency. The spatio-temporal features are denoted by $F^v = [F_1^v, \dots, F_N^v] \in \mathbb{R}^{d*N}$, where d is the dimension of the projection space.

The text input is first tokenized using the WordPieces tokenizer [27], a sub-word segmentation algorithm with a 30,000 token vocabulary. The first token of the input question is a [CLS] token and the last token is the [SEP] token. We use [PAD] token to truncate the sentence with equal length. Each token is then fed to DistilBERT [28]. DistilBERT is an efficient, lightweight version of BERT, which is trained under low latency constraints. We use the activations of the last layer of DistilBERT to obtain a 768-dimensional feature vector which is then passed to a feed forward network, similarly to the video projection. The text embedding is denoted as $F^q = [F_1^q, \dots, F_T^q] \in \mathbb{R}^{d*T}$, where T is the number of tokens in the question.

B. Cross-modal module

Modeling video-text dynamics within and across modalities is an extremely challenging task. To mitigate this problem, we develop a cross-modal correlation module that efficiently accounts both intra and inter-modal relationships between modalities. Fig.2 represents the architecture of the proposed module. Inspired by the work in [29], we consider the cross-correlation matrix τ (Eq. (1)) that aims at modeling the relationships between the various visual and textual modalities involved:

$$\tau = F^{q^T} W F^v, \tag{1}$$

where $W \in \mathbb{R}^{d*d}$ is a learnable matrix.

A high coefficient of the correlation matrix τ means that the corresponding video and text features are highly relevant. We generate the cross-correlation video-question (resp. question-video) weights by column-wise softmax over τ and τ^T, respectively. This technique allows learning more discriminative representations for each individual modality, constrained by the other one. Formally, we compute the video-conditioned question features as:

$$F^{q-v} = F^q \, softmax(\tau^T) \tag{2}$$

Similarly, the question-conditioned video features are defined as:

$$F^{v-q} = F^v softmax(\tau) \tag{3}$$

To prevent information loss in the cross-correlation stage, we have adopted the dense skip connection technique. The reweighted features F^{q-v} and F^{v-q} are thus added to the original modality-specific representation.

$$\widehat{F^{q-v}} = \tanh(F^{q-v} + F^q) \tag{4}$$

$$\widehat{F^{v-q}} = \tanh(F^{v-q} + F^{v}) \quad (5)$$

The obtained features are further exploited in the multi-modal fusion module, as described in the following section.

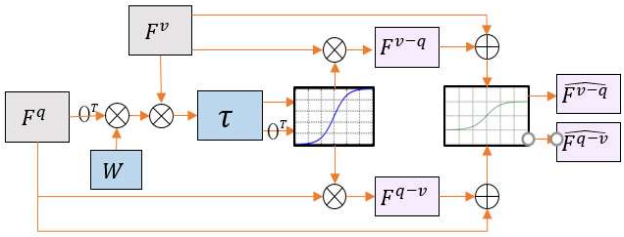

Fig. 2. Cross-modal correlation architecture

C. Transformed-based multimodal fusion

Different from recurrent neural networks, transformers are order-insensitive. For this reason, we add a positional encoding to explicitly retain the information regarding the word position as in [14]. We differentiate the two modalities using a learned embedding layer which is added to each token to specify whether it belongs to the video or text (see Fig.1). The representation of the video and the question are computed as follows.

$$\widehat{F_k^{q-v}} = dp(\widehat{F_k^{q-v}} + pos_k + mod_q) \quad (6)$$

and

$$\widehat{F_k^{v-q}} = dp(\widehat{F_k^{v-q}} + pos_k + mod_v) \quad (7)$$

where $mod_q \in \mathbb{R}^d, mod_v \in \mathbb{R}^d$ represent learnt modality embeddings; and $[pos_1, \dots, pos_{t+c}] \in \mathbb{R}^{d*T+N}$ are positional encodings. dp is the dropout layer.

The input to the transformer $F^{qv} \in \mathbb{R}^{d*T+N}$ is the concatenation of $\widehat{F_k^{q-v}}$ and $\widehat{F_k^{v-q}}$.

The transformer layers consist of an attention sublayer [1] followed by a position-wise feed-forward layer. The attention sublayers employ H attention heads. To obtain the sublayer output $O \in \mathbb{R}^{seq_{length}*d}$ $(seq_{leng} = T + N)$, we concatenate the results from each head and apply a linear projection. Each attention head operates on an input sequence $X \in \mathbb{R}^{H*seq\_length*d_{head}}$ and computes the attended feature $Z \in \mathbb{R}^{H*seq\_lengt\ *d_{head}}$ as follows.

$$z_i = \sum_{j=1}^{seq\_length} \alpha_{ij}(x_j W^V) \quad (8)$$

The weight coefficient α_{ij} is calculated using a softmax function.

$$\alpha_{ij} = \frac{exp\ e_{ij}}{\sum_{k=1}^{t+c} e_{ik}} \quad (9)$$

where

$$e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_{head}}} \quad (10)$$

where W^Q, W^K, $W^V \in \mathbb{R}^{d*d}$ are learnable matrices and d_{head} denotes a scaling factor.

The output of the transformer is then passed to an MLP (linear projection followed by GELU activation and layer normalization) with softmax to predict the correct answer from the vocabulary of predefined answers.

D. Rephrasing attacks

The objective of adversarial attacks is to fool the learned model by manipulating the input provided to it. This is not only important to test the vulnerability of DL models to security threats but also to verify its robustness in real-world scenarios. Adversarial attacks have been first introduced in the image domain for object recognition [30-32], then attracted many follow-up efforts in other domains including natural language processing (NLP). Text attacks are more challenging due to different reasons: (1) Small changes in the image are unperceivable by humans while text changes can be easily identified; (2) The semantics of the image are not changed by small perturbations. In contrast, even minor text manipulations can affect the general meaning of a sentence.

A successful attack should take such considerations into account, in order to be able to fool the DL model without changing the human judgement. Adversarial attacks can be categorized into two classes. A first one concerns the so-called *white box attacks*: in this setting, the attacker has access to the model information including input-output data, model architecture, parameters, loss functions and activation functions. The adversarial data is adjusted to maximize its influence on the classifier while keeping an imperceptible change. Most approaches use the gradient information of the loss with respect to the input to build the attack. In [33], authors use fast gradient sign method (FGSM) [31] by identifying the words with the most significant contribution to classification task. Specifically, they compute the cost gradient of training examples using backpropagation and assign the contribution of each item with respect to the magnitude of the cost gradient. Jacobian Saliency Map Adversary [34] (JSMA)-based methods [35-37] build adversarial perturbations using forward derivatives.

In the case of the second family of methods, called *black box attacks*, the attacker has only access to input-output data. This approach uses heuristic methods or iterative queries to perform the attack. In [38], authors distract the textual input by appending meaningless sentences at the end of the paragraph. Such perturbations are crafted by iteratively querying the model until the output changes. In [39], various strategies are applied to affect the model's performance such as random swap (transposing neighbor words), random deletion, stop-word dropout, paraphrasing as well as grammar and keyboard errors. In [40,41], the important tokens are identified based on a scoring system which measures the degree of perturbation of the model's output. The selected tokens are then modified using four techniques: delete, replace, swap and add. In [42,43], authors generate semantically equivalent adversaries (SEA) to fool the model. Such approaches generate paraphrases and compare the model's prediction with the original sentences. Other works [44,45] leverage generative adversarial networks (GANs) [46] to generate adversarial examples by searching for the neighbors of the input data in the latent space. The output of the adversarial attacks can be *targeted*, meaning that the attacker maps the output to a desired value, or *untargeted* in which case the attacker cares only about producing incorrect output. For multimodal attacks, there has been some work on image captioning [47], optical character recognition [48] and image question answering[49]. To the best of our knowledge,

this is the first work to consider adversarial attacks issues under the framework of video question answering methods.

Our objective is to verify the importance of the building elements of our pipeline and test their respective contribution to the model prediction. To allow a fair comparison, the same model-independent attacks are applied on the different models. For this reason, we apply untargeted black-box attack, meaning that we do not enforce any specific results. We use an automatic method to generate the rephrased questions without additional human intervention, which is more scalable in real-world environments. To this purpose, we have retained the BART approach [50], which is a sequence-to-sequence NLP model that uses a BERT -like encoder (i.e., bidirectional encoder) and a GPT-like decoder (i.e., left-to-right decoder). BART is pre-trained in an unsupervised manner using general objectives such as text corruption with random noise and text shuffling. The model is originally applied to sequence generation and machine translation tasks. The model is fine-tuned for text rephrasing purposes. The pre-trained model is directly used as a sequence-to-sequence model. At each time step, the model computes the probability of each word in the vocabulary to be the likely next word. Then, the next word is picked based on three decoding methods: (1) random sampling: we randomly choose the next word w_t according to its conditional probability distribution. (2) top-K sampling [51]: we only sample the K high probability words from the distribution. (3) top-p (nucleus) sampling: we sample from a set of words whose cumulative probability exceeds p.

For training, we use three datasets: Quora [52] (400k training samples), MSRP [53](13M training samples) and PAWS [54] (108k training samples). The original data is filtered to ensure more diversity as follows. First, the sentence pairs that present more than 80% unigram overlap are removed. This first step minimizes the chance to copy the original sentence. We use Siamese BERT [55] to remove the question pairs with low semantic similarity. For MSRP and Quora, we only select the sentences that are rephrases to each other. Finally, the trained model is applied on the test set of MSVD-QA.

TABLE I. EXAMPLES OF REPHRASED QUESTIONS USING AUTOMATIC TECHNIQUE

Original question	**Paraphrased question**
Who is on an ambulance stretcher	Who is riding an ambulance stretcher?
What are school aged children doing?	What is a group of teenagers doing?
How many elephants are spraying water on themselves?	How many elephants are spraying water on themselves with their trunks?
What is the best way to cut potato into pieces with a knife?	Who is cutting into pieces a potato with a knife?
What does a man pick up a card from?	What does a man pick a card up from?
What is climbing?	What is climbing?

Some examples of rephrased questions are provided in Table I. In order to compare the differences between the two datasets (original and rephrased) we compute the GLEU score [56] which is more suitable for single sentences. GLEU is a variant of the BLEU score that assigns more weight to n-grams that are changed from the source. Specifically, the GLEU score is the minimum of recall (ratio of the number of matching n-grams to the total number of n-grams in the original question) and precision (the ratio of the number of matching n-grams to the total n-grams in the rephrased question). The GLEU score range is between 0 (no matches) and 1 (all match). We have obtained a GLEU score of 0.5638.

IV. EXPERIMENAL EVALUATION

The experimental evaluation has been carried out on the publicly available dataset MSVD-QA[5].

A. Datasets

Under the framework of a *pre-training, then fine-tuning* paradigm, we have trained the model on the HowToVQA 69M task-specific dataset. HowToVQA 69M is today the largest videoQA available dataset, with over 69 million video question-answer triplets. The videos have been extracted from HowTo100M, which was originally designed for video captioning purposes. The question-answer pairs are automatically generated from the transcribed speech using two transformers. We randomly select 164148 training samples to reduce memory and computational requirements.

For fine-tuning, we have retained the popular MSVD-QA videoQA dataset, which represents a smaller dataset automatically derived from MSVD. It contains 1970 clips and 50505 question-answer pairs. MSVD-QA contains five categories of question, which are "What", "Who", and "When". The answer vocabulary contains 1852 training answers.

B. Implementation details

For pre-processing, we uniformly sample the video into $N = 20$ clips. Similarly, we set the maximum number of tokens in the question to $T = 20$. We project the video features and text features into a common embedding space of size d=512. For the multimodal transformer, a number of H=8 attention heads are retained. In this setting, the scaling factor d_{head} is the fraction of the embedding size over the number of heads $d_{head} = \frac{d}{H} = 64$. To train the rephrasing model BART, we select the high probability words based on top-K and p-sampling strategies. We set K=50 and p=0.95.

The loss function of the proposed model is the sum of the cross entropy loss and the masked language modeling (MLM) loss. The MLM objective is to predict a randomly masked word from a predefined vocabulary of 30K words. MLM loss is the negative log-likelihood for masked words. Specifically, we randomly select with a probability of 15% all WordPiece tokens in each question. Once the token is selected, the data generator replaces the token with a special token [MASK] 80% of the time, a random token 10% of the time, and the same token 10% of the time. The goal of this procedure is to influence the model to maintain a contextual representation of each input token, since it does not know which words will be predicted.

A cosine annealing learning rate schedule has been used, with initial values of 10^{-4} for pre-training and 10^{-5} for fine-tuning respectively. For optimization, we have adopted the Adam approach with batch size of 16 for pre-training and 32 for fine-tuning. The training process has been run on 2 NVIDIA GeForce RTX 2080 GPUs and for 20 epochs.

(a)

Original question: What are school aged children doing?
Rephrased question: What is a group of teenagers doing?
Ground truth: Perform
Original prediction: Perform
Prediction after rephrase: Perform

(b)

Original question: What is a man showing in a box?
Rephrased question: What is a man in a box?
Ground truth: Gun
Original prediction: Gun
Prediction after rephrase: Gun

(c)

Original question: What flees from an eagle?
Rephrased question: What escapes from an eagle?
Ground truth: Rabbit
Original prediction: Rabbit
Prediction after rephrase: Rabbit

(d)

Original question: What is the dog enjoyed doing?
Rephrased question: What do the dog like to do?
Ground truth: Play
Original prediction: Play
Prediction after rephrase: Play

(e)

Original question: Who is playing the guitar on stage in front of an audience?
Rephrased question: Who is playing guitar in front of an audience?
Ground truth: Man
Original prediction: Someone.
Prediction after rephrase: Play.

(f)

Original question: What is a man demonstrating his skills with in front of a crowd?
Rephrased question: What is a man doing in front of a crowd?
Ground truth: Sword.
Original prediction: Sword.
Prediction after rephrase: Ball.

Fig. 3. Examples of results of our approach on the MSVD-QA data set, with both original and rephrased questions.

The final model is selected according to the best performance on the validation set.

C. *Ablation study*

To investigate the effectiveness of each component of the pipeline, we have compared the performance of different baselines on both original and rephrased datasets.

More precisely, the following test baselines have been retained: (B1). early fusion strategy by concatenating the video and text representations of pre-trained models and then feeding them into a fully connected layer to predict the correct answer; (B2) cross-modal matching that learns intra-modal representations of each modality under the constraint of the other (Section B); (B3) multimodal transformer that neglects the cross-modal module. (B4) The proposed architecture trained from scratch on MSVD-QA. Let us note that baseline models B1 to B4 are trained from scratch on MSVD-QA for computational efficiency. (B5) Our model pre-trained on a subset of HowToVQA 69M then fine-tuned on MSVD-QA.

We use the accuracy metric as the answers do not exceed several words. The accuracy represents the ratio of the correct predictions with respect to the total number of input samples. The obtained results are summarized in Table II.

TABLE II. ABLATION STUDIES ON MSVD-QA. ACC1 REPRESENTS THE PERFORMANCE ON THE ORIGINAL DATASET. ACC2 REPRESENTS THE PERFORMANCE ON THE REPHRASED DATASET.

Methods	ACC1	ACC2
B1. Early fusion baseline	27.33%	21.31%
B2. Cross-modal module only	31.05%	25.81%
B3. Transformer module only	37.88%	33.47%
B4. Proposed model trained from scratch on MSVD-QA	38.96%	33.87%
B5. Proposed model pre-trained on HowToVQA then fine-tuned on MSVD-QA	**43.57%**	**39.42%**

The following conclusions can be drawn. (1) The lowest score is obtained by directly concatenating video and text representations. This behavior can be explained by the heterogeneous nature of the two modalities involved which in addition are pre-trained with different tasks/datasets. (2) Cross-correlation technique yields more grounded representations as features are learned under the constraint of the other modality, with a 3.72% improvement in accuracy. (3) The best results are obtained using the full pipeline, which integrates extensive inter-modal interactions. (4) Pre-training on large task-specific datasets effectively optimizes the weights of the proposed architecture. (5) Our approach is more robust to rephrasing attacks then the transformer-only architecture. This is due to learning-conditioned features as opposed to simple concatenation.

Fig. 3 shows some examples of results obtained with the proposed approach on the MSVD-QA data set, with both original and rephrased questions. Let us note that if the question is not clear, we can state that the model is able to extract meaningful information from the video (example (e)), even if the prediction is wrong.

D. Comparison with the state of the art

We have compared our approach to various state of the art methods on the MSVD-QA dataset [5]. Table III summarizes the accuracy of the different VideoQA models retained for comparison. More precisely, we have considered the following methods Co-Mem [19], HCRN[23] , B2A [57] and CoMVT [2].

The proposed method achieves the highest accuracy of 43.57%. In particular, it outperforms the state of the art CoMVT model by 4.61%, even if CoMVT is pre-trained on a larger, task-independent dataset (HowTo100M). CoMVT uses four transformer blocks to model intra- and inter-model dynamics, while we use a simple weight matrix followed by a 2-layer transformer. This demonstrates the importance of task-oriented pre-training and the effectiveness of our model, which minimizes the required computational effort.

TABLE III. COMPARATIVE EXPERIMENTAL RESULTS

Methods	Accuracy
Co-Mem [19]	31.7%
HCRN [23]	36.1%
B2A [57]	37.2%
CoMVT [2]	42.6%
Proposed model	**43.57%**

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a novel multimodal framework for video question answering. The proposed system is based on reciprocally constrained, cross-correlation conditioning of visual and textual features. Our system also integrates attention mechanisms using a multimodal, transformer-based approach to capture complex inter-modal dynamics. Ablation studies demonstrate the importance of each composing block of the approach. We have also proved the effectiveness of our pipeline by testing the robustness of the model to rephrasing attack. Finally, we have achieved 43.57% of accuracy on MSVD-QA dataset outperforming previous state of the art CoMVT methods with 4.61%.

For future work, we envisage to extend the video question-answering framework in order to incorporate natural language script input and audio features. We also intend to apply our model on real video-question platform to perform a subjective system evaluation with user feedback.

REFERENCES

[1] A. Vaswani et al., "Attention is all you need," arXiv [cs.CL], 2017.

[2] P. Hongsuck Seo, A. Nagrani, and C. Schmid, "Look Before you Speak: Visually Contextualized Utterances," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[3] J. Kim, M. Ma, T. Pham, K. Kim, and C. D. Yoo, "Modality shifting attention network for multi-modal video question answering," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[4] R. Zellers et al., "MERLOT: Multimodal neural script knowledge models," arXiv [cs.CV], 2021.

[5] D. Xu, Z. Zhao, J. Xiao, F. Wu, H. Zhang, X. He, and Y. Zhuang,''Video question answering via gradually refined attention over appearance and motion,'' in Proc. 25th ACM Int. Conf. Multimedia, Oct. 2017,pp. 1645–1653.

[6] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," arXiv [cs.CV], 2019.

[7] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; Look and answer: Overcoming priors for visual question answering," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[8] G. Kv and A. Mittal, "Reducing language biases in visual question answering with visually-grounded question encoder," in Computer Vision – ECCV 2020, Cham: Springer International Publishing, 2020, pp. 18–34.

[9] K. Kafle and C. Kanan, "An analysis of visual question answering algorithms," in 2017 IEEE International Conference on Computer Vision (ICCV), 2017.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv [cs.CV], 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[12] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, ''Efficient estimation of word representations in vector space,'' 2013, arXiv:1301.3781.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," arXiv [cs.CL], 2018.

[15] Z. Zhao, Q. Yang, D. Cai, X. He, and Y. Zhuang, "Video question answering via hierarchical spatio-temporal attention networks," in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017.

[16] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in Computer Vision – ECCV 2018, Cham: Springer International Publishing, 2018, pp. 487–503.

[17] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, ''MovieQA: Understanding stories in movies through questionanswering,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4631–4640.

[18] J. Kim, M. Ma, K. Kim, S. Kim, and C. D. Yoo, "Progressive attention memory network for movie story question answering," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[19] J. Gao, R. Ge, K. Chen, and R. Nevatia, "Motion-appearance co-memory networks for video question answering," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[20] J. Lei et al., "Less is more: CLIPBERT for video-and-language learning via sparse sampling," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[21] W. Yu, H. Zheng, M. Li, L. Ji, L. Wu, N. Xiao, and N. Duan,"Learning from inside: Self-driven siamese sampling and reasoning for video question answering,"Advances in Neural Information Processing Systems, vol. 34, 2021.

[22] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," arXiv [cs.CV], 2020.

[23] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[24] X. Li et al., "Beyond RNNs: Positional Self-attention with co-attention for video question answering," Proc. Conf. AAAI Artif. Intell., vol. 33, pp. 8658–8665, 2019.

[25] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, “End-to-end learning of visual representations from uncurated instructional videos,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[26] D. Hendrycks and K. Gimpel, “Gaussian Error Linear Units (GELUs),” arXiv [cs.LG], 2016

[27] Y. Wu et al., “Google’s Neural Machine Translation system: Bridging the gap between human and Machine Translation,” arXiv [cs.CL], 2016.

[28] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arXiv [cs.CL], 2019.

[29] J.-T. Lee, M. Jain, H. Park, and S. Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In ICLR, 2021.

[30] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in 2017 IEEE Symposium on Security and Privacy (SP), 2017.

[31] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” arXiv [stat.ML], 2014.

[32] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “DeepFool: A simple and accurate method to fool deep neural networks,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[33] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, “Deep Text Classification Can be Fooled,” in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018.

[34] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” in 2016 IEEE European Symposium on Security and Privacy (EuroS&P), 2016.

[35] N. Papernot, P. McDaniel, A. Swami, and R. Harang, “Crafting adversarial input sequences for recurrent neural networks,” in MILCOM 2016 - 2016 IEEE Military Communications Conference, 2016.

[36] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, “Adversarial perturbations against deep neural networks for malware classification,” arXiv [cs.CR], 2016.

[37] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, “Adversarial examples for malware detection,” in Computer Security – ESORICS 2017, Cham: Springer International Publishing, 2017, pp. 62–79.

[38] R. Jia and P. Liang, “Adversarial examples for evaluating reading comprehension systems,” in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017.

[39] Y. Belinkov and Y. Bisk, “Synthetic and natural noise both break neural machine translation,” arXiv [cs.CL], 2017.

[40] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, “Black-box generation of adversarial text sequences to evade deep learning classifiers,” in 2018 IEEE Security and Privacy Workshops (SPW), 2018.

[41] J. Li, S. Ji, T. Du, B. Li, and T. Wang, “TextBugger: Generating adversarial text against real-world applications,” in Proceedings 2019 Network and Distributed System Security Symposium, 2019.

[42] M. Iyyer, J. Wieting, K. Gimpel, and L. Zettlemoyer, “Adversarial example generation with syntactically controlled paraphrase networks,” arXiv [cs.CL], 2018.

[43] M. T. Ribeiro, S. Singh, and C. Guestrin, “Semantically equivalent adversarial rules for debugging NLP models,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018.

[44] Z. Zhao, D. Dua, and S. Singh, “Generating natural adversarial examples,” arXiv [cs.LG], 2017.

[45] H. Gao, H. Zhang, X. Yang, W. Li, F. Gao, and Q. Wen, “Generating natural adversarial examples with universal perturbations for text classification,” Neurocomputing, vol. 471, pp. 175–182, 2022.

[46] I.J.Goodfellow et.al.. Generative adversarial nets. volume 2, pages 2672–2680. NIPS’14 Proceed- ings of the 27th International Conference on Neural Information Processing Systems, 2014.

[47] H. Chen, H. Zhang, P.-Y. Chen, J. Yi, and C.-J. Hsieh, “Attacking visual language grounding with adversarial examples: A case study on neural image captioning,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018.

[48] C. Song and V. Shmatikov, “Fooling OCR systems with adversarial text images,” arXiv [cs.LG], 2018.

[49] X. Xu, X. Chen, C. Liu, A. Rohrbach, T. Darrell, and D. Song, “Fooling vision and language models despite localization and attention mechanism,” in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[50] M. Lewis et al., “BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

[51] A. Fan, M. Lewis, and Y. Dauphin, “Hierarchical Neural Story Generation,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018.

[52] N. Ansari and R. Sharma, “Identifying semantically duplicate questions using data science approach: A Quora case study,” arXiv [cs.IR], 2020.

[53] W. B. Dolan and C. Brockett, “Automatically constructing a corpus of sentential paraphrases,” in Proceedings of the Third International Workshop on Paraphrasing (IWP2005), 2005.

[54] Y. Zhang, J. Baldridge, and L. He, “PAWS: Paraphrase Adversaries from Word Scrambling,” arXiv [cs.CL], 2019.

[55] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019.

[56] Y. Wu et al., “Google’s Neural Machine Translation system: Bridging the gap between human and Machine Translation,” arXiv [cs.CL], 2016.

[57] J. Park, J. Lee, and K. Sohn, “Bridge to answer: Structure-aware graph interaction network for video question answering,” in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.