Attend What You Need: Motion-Appearance Synergistic Networks for Video Question Answering

Ahjeong Seo[1], **Gi-Cheon Kang**[1,2], **Joonhan Park**[3,*], **Byoung-Tak Zhang**[1,2]
[1]Seoul National University
[2]AI Institute for Seoul National University (AIIS)
[3] Hanyang University
{ajseo,gckang,jhpark,btzhang}@bi.snu.ac.kr

Abstract

Video Question Answering is a task which requires an AI agent to answer questions grounded in video. This task entails three key challenges: (1) understand the intention of various questions, (2) capturing various elements of the input video (*e.g.,* object, action, causality), and (3) cross-modal grounding between language and vision information. We propose **M**otion-**A**ppearance **S**ynergistic **N**etworks (MASN), which embed two cross-modal features grounded on motion and appearance information and selectively utilize them depending on the question's intentions. MASN consists of a motion module, an appearance module, and a motion-appearance fusion module. The motion module computes the action-oriented cross-modal joint representations, while the appearance module focuses on the appearance aspect of the input video. Finally, the motion-appearance fusion module takes each output of the motion module and the appearance module as input, and performs question-guided fusion. As a result, MASN achieves new state-of-the-art performance on the TGIF-QA and MSVD-QA datasets. We also conduct qualitative analysis by visualizing the inference results of MASN. The code is available at https://github.com/ahjeongseo/MASN-pytorch.

1 Introduction

Recently, research in natural language processing and computer vision has made significant progress in artificial intelligence (AI). Thanks to this, vision-language tasks such as image captioning (Xu et al., 2015), visual question answering (VQA) (Antol et al., 2015; Goyal et al., 2017), and visual commonsense reasoning (VCR) (Zellers et al., 2019) have been introduced to the research community, along with some benchmark datasets. In particular, video question answering (video QA) tasks (Xu et al., 2016; Jang et al., 2017; Lei et al., 2018; Yu et al., 2019; Choi et al., 2020) have been proposed with the goal of reasoning over higher-level vision-language interactions. In contrast to QA tasks based on static images, the questions presented in the video QA dataset vary from frame-level questions regarding the appearance of objects (*e.g.,* what is the color of the hat?) to questions regarding action and causality (*e.g.,* what does the man do after opening a door?).

There are three crucial challenges in video QA: (1) understand the intention of various questions, (2) capturing various elements of the input video (*e.g.,* object, action, and causality), and (3) cross-modal grounding between language and vision information. To tackle these challenges, previous studies (Li et al., 2019; Jiang et al., 2020; Huang et al., 2020) have mainly explored this task by jointly embedding the features from the pre-trained word embedding model (Pennington et al., 2014) and the object detection models (He et al., 2016; Ren et al., 2016). However, as discussed in (Gao et al., 2018), the use of the visual features extracted from the object detection models suffers from motion analysis since the object detection model lacks temporal modeling. To enforce the motion analysis, a few approaches (Xu et al., 2017; Gao et al., 2018) have employed additional visual features (Tran et al., 2015) (*i.e.,* motion features) which were widely used in the action recognition domain, but their reasoning capability is still limited. They typically employed recurrent models (*e.g.,* LSTM) to embed a long sequence of the visual features. Due to the problem of long-term dependency in recurrent models (Bengio et al., 1993), their proposed methods may fail to learn dependencies between distant features.

In this paper, we propose Motion-Appearance

[*] Work done during an internship at AI Institute for Seoul National University (AIIS).

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6167–6177
August 1–6, 2021.

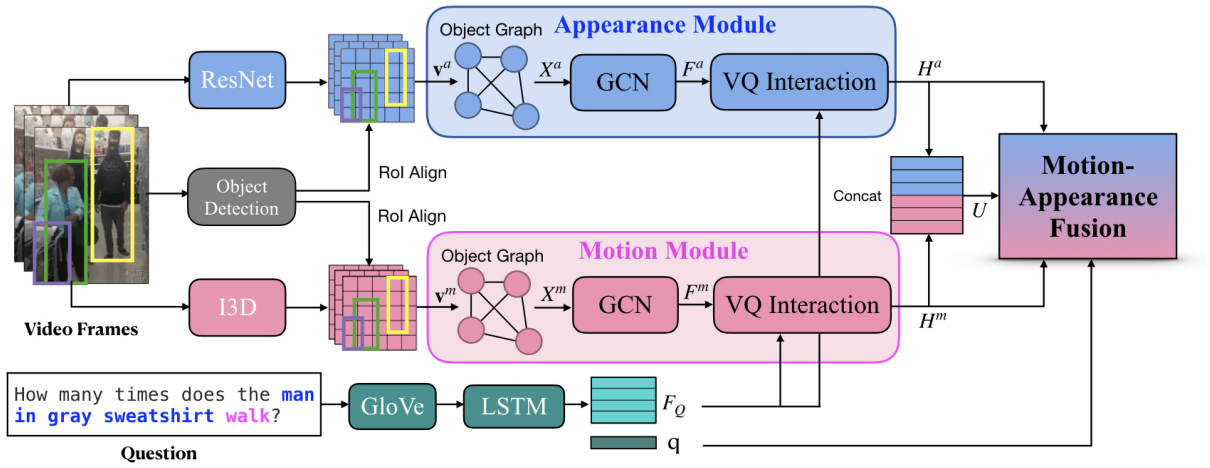

Figure 1: An overview of MASN. Each extracted feature from ResNet and I3D is fed into the Appearance and Motion modules. Both modules have the same structure with a GCN and VQ interaction submodule. The results from each module are then concatenated and fused in the Motion-Appearance Fusion module. The output from the fusion module is used to derive answers. For question features, the word-level representation F_Q is integrated with the visual features in the VQ interaction submodule. The last hidden units $\mathbf{q}$ from the bi-LSTM are used to combine appearance and motion features.

Synergistic Networks (MASN) for video question answering which consist of three kinds of modules: the motion module, the appearance module, and the motion-appearance fusion module. As shown in Figure 1, the motion module and the appearance module aim to embed rich cross-modal representations. These two modules have the same architecture except that the motion module takes the motion features extracted from I3D as visual features and the appearance module utilizes the appearance features extracted from ResNet. Each of these modules first constructs the object graphs via graph convolutional networks (GCN) to compute the relationships among objects in each visual feature. Then, the vision-question interaction module performs cross-modal grounding between the output of the GCNs and the question features. The motion module and the appearance module each yield cross-modal representations of the motion and the appearance aspects of the input video respectively. The motion-appearance fusion module finally integrates these two features based on the question features.

The main contributions of our paper are as follows. First, we propose Motion-Appearance Synergistic Networks (MASN) for video question answering based on three modules, the motion module, the appearance module, and the motion-appearance fusion module. Second, we validate MASN on the large-scale video question answering datasets TGIF-QA, MSVD-QA, and MSRVTT-QA. MASN achieves the new state-of-the-art performance on TGIF-QA and MSVD-QA. We perform ablation studies to validate the effectiveness of our proposed methods. Finally, we conduct a qualitative analysis of MASN by visualizing inference results.

2 Related Work

Visual Question Answering (VQA) is a task that requires both understanding questions and finding clues from visual information. VQA can be classified into two categories based on the type of the visual source: image QA and video QA. In image QA, earlier works approach the task by applying attention between the question and the spatial dimensions of the image (Yang et al., 2016; Anderson et al., 2018; Kim et al., 2018a; Kang et al., 2019). In video QA, since a video is represented as a sequence of images over time, recognizing the movement of objects or causality in the temporal dimension should also be considered along with the details from the spatial dimension (Jang et al., 2017; On et al., 2020). There have been some attempts (Xu et al., 2017; Gao et al., 2018; Fan et al., 2019) to extract motion and appearance features and integrate them on a spatio-temporal dimension via memory networks. Li et al. (2019), Huang et al. (2020), Jiang et al. (2020) proposed better performing models using attention in order to overcome the long-range dependency problem in memory networks. However, they do not represent motion in-

formation sufficiently since they only use features pre-trained on image or object classification. To better address this, we model spatio-temporal reasoning on multiple visual information (*i.e.,* ResNet, I3D) while also solving the long-range dependency problem that occurred in previous studies.

Action Classification is a task of recognizing actions, which are composed of interactions between actors and objects. Therefore, this task has much in common with video QA, in that the model should perform spatio-temporal reasoning. For better spatio-temporal reasoning, Tran et al. (2015) introduced C3D, which extends the 2D CNN filters to the temporal dimension. Carreira and Zisserman (2017) proposed I3D, which integrates 3D convolutions into a state-of-the-art 2D CNN architecture, which now acts as a baseline in action classification tasks (Murray et al., 2012; Girdhar et al., 2018). Feichtenhofer et al. (2019) introduced SlowFast, a network which encodes images in two streams with different frame rates and temporal resolutions of convolution. This study based on a two-stream architecture inspired us in terms of assigning different inputs to each encoder module. However, our method differs from the former studies in two aspects: (1) we utilize language features as well as vision features, and (2) we expand the two-stream structure to solve more than motion-oriented tasks.

Attention Mechanism explicitly calculates the correlation between two features (Bahdanau et al., 2015; Lin et al., 2017), and has been widely used in a variety of fields. For machine translation, the Transformer architecture first introduced by Vaswani et al. (2017), utilizes multi-head self-attention that captures diverse aspects in the input features (Voita et al., 2019). For video QA, Kim et al. (2018b); Li et al. (2019) use self and guided-attention to encode temporal dynamics in video and ground them in the question. For multi-modal alignment, Tsai et al. (2019) apply the Transformer to merge cross-modal time series between vision, language, and audio features. We utilize the attention mechanism to capture various relations between appearance and motion and to aggregate them.

3 Model

In this section, we introduce a detailed description of our MASN network. First, we explain how to obtain appearance and motion features in Section 3.1. Then, we describe the Appearance and Motion modules, which encode visual features and combine them with the question in Section 3.2. Finally, the Motion-Appearance Fusion module modulates the amount of motion and appearance information utilized and integrates them based on question context.

3.1 Visual and Linguistic Representation

We first extract appearance and motion features from the video frames. For the appearance representation, we use ResNet (He et al., 2016) pre-trained on an object and its attribute classification task as a feature extractor. For the motion representation, we use I3D (Carreira and Zisserman, 2017) pre-trained on the action classification task. We obtain local features representing object-level information without background noise and global features representing each frame's context for both appearance and motion features.

Appearance Representation. For local features, given a video containing T frames, we obtain N objects from each frame using Faster R-CNN (Ren et al., 2016) that applies RoIAlign to extract the region of interest from ResNet's convolutional layer. We denote the appearance-object set as $\mathcal{R}^a = \{\mathbf{o}^a_{t,n}, \mathbf{b}_{t,n}\}^{t=T,n=N}_{t=1,n=1}$, where $\mathbf{o}$, $\mathbf{b}$ indicate object feature and bounding box location, respectively. Therefore, there are $K = N \times T$ objects in a single video. Following previous works, we extract the feature map from ResNet-152's *Conv5* layer and apply a linear projection (Jiang et al., 2020; Huang et al., 2020). We denote global features as $\mathbf{v}^a_{global} \in \mathbb{R}^{T \times d}$, where d is the size of the hidden dimension.

Motion Representation. We obtain a feature map from the last convolutional layer in I3D (Carreira and Zisserman, 2017) whose dimension is (time, width, height, feature) = ($\lfloor \frac{t}{8} \rfloor$, 7, 7, 2048). That is, each set of 8 frames is represented as a single feature map with dimension $7 \times 7 \times 2048$. For local features, we apply RoIAlign (He et al., 2017) on the feature map using object bounding box location $\mathbf{b}$. We define the motion-object set as $\mathcal{R}^m = \{\mathbf{o}^m_{t,n}, \mathbf{b}_{t,n}\}^{t=T,n=N}_{t=1,n=1}$. We apply average pooling in the feature map and linear projection to obtain global features $\mathbf{v}^m_{global} \in \mathbb{R}^{T \times d}$.

Location Encoding. To reason about relations between objects as in Section 3.2, it is required to consider each object's spatial and temporal location. As appearance and motion features share identical operations until the Motion-Appearance

Fusion module, we combine superscript a and m for simplicity. Following L-GCN (Huang et al., 2020), we add a location encoding and define local features as:

$$\mathbf{v}_{local}^{a/m} = \mathrm{FFN}([\mathbf{o}^{a/m}; \mathbf{d}^{s}; \mathbf{d}^{t}]) \quad (1)$$

where $\mathbf{d}^{s} = \mathrm{FFN}(\mathbf{b})$ and $\mathbf{d}^{t}$ is obtained by position encoding according to each frame's index. Here $\mathbf{o}^{a/m}$ denotes the object features mentioned above while FFN denotes a feed-forward network. Analogous to local features, position encoding information $\mathbf{d}^{t}$ is added to global features as well. We then concatenate object features with global features to reflect the frame-level context in objects and obtain the visual representation $\mathbf{v}^{a/m} \in \mathbb{R}^{K \times d}$:

$$\mathbf{v}^{a/m} = \mathrm{FFN}([\mathbf{v}_{local}^{a/m}; \mathbf{v}_{global}^{a/m}]) \quad (2)$$

Linguistic Representation. We apply the pre-trained GloVe to convert each question word into a 300-dimensional vector, following previous work (Jang et al., 2017). To represent contextual information in a sentence, we feed the word representations into a bidirectional LSTM (bi-LSTM). Word-level features and last hidden units from the bi-LSTM are denoted by $F^{q} \in \mathbb{R}^{L \times d}$, and $\mathbf{q} \in \mathbb{R}^{d}$ respectively. L denotes the number of words in a question.

3.2 Motion and Appearance Module

In this section, we explain the modules generating high-level visual representations and integrate them with linguistic representations. Each module consists of (1) an **Object Graph**: spatio-temporal reasoning between object-level visual features, and (2) **VQ interaction**: calculating correlations between objects and words and obtaining cross-modal feature embeddings. Since the modules share the same architecture, we describe each module's components only once with a shared superscript to avoid redundancy.

3.2.1 Object Graph Construction

In this section, we define object graphs $\mathcal{G}^{a/m} = (\mathcal{V}^{a/m}, \mathcal{E}^{a/m})$ to capture spatio-temporal relations between objects. $\mathcal{V}$, $\mathcal{E}$ denotes the node and edge set of the graph. As equation 2 provides visual features $\mathbf{v}^{a/m}$, we define these as the graph input $X^{a/m} \in \mathbb{R}^{K \times d}$. We denote the graph as $\mathcal{G}^{a/m}$. The nodes of graph $\mathcal{G}^{a/m}$ are given by $v_i^{a/m} \in X^{a/m}$, and edges are given by $(v_i^{a/m}, v_j^{a/m})$, representing a relationship between the two nodes. Given the constructed graph $\mathcal{G}$, we perform graph convolution (Kipf and Welling, 2016) to obtain the relation-aware object features. We obtain the similarity scores of nodes by calculating the dot-product after projecting input features to the interaction space and define the adjacency matrix $A^{a/m} \in \mathbb{R}^{K \times K}$ as follows:

$$A^{a/m} = \mathrm{softmax}((X^{a/m}W_1)(X^{a/m}W_2)^{\top}) \quad (3)$$

We denote the two-layer graph convolution on input X with adjacency matrix A as:

$$\begin{aligned} \mathrm{GCN}(X; A) &= \mathrm{ReLU}(A\ \mathrm{ReLU}(AXW_3)\ W_4) \\ F &= \mathrm{LayerNorm}(X + \mathrm{GCN}(X; A)) \end{aligned} \quad (4)$$

We omit superscripts in the graph convolution equation for simplicity. We add a skip connection for residual learning between self-information X and smoothed-information with neighbor objects.

3.2.2 Vision-question (VQ) Interaction

We compute both appearance-question and motion-question interaction to obtain correlations between language and each of the visual features. As we encode visual feature $F^{a/m}$ and question feature F^{q} in Equation 4 and Section 3.1, we calculate every pair of relations between two modalities using the bilinear operation introduced in BAN (Kim et al., 2018a) as follows:

$$H_i = \mathbb{1} \cdot \mathrm{BAN}_i(H_{i-1}, V; \mathcal{A}_i)^{\top} + H_{i-1} \quad (5)$$

where $H_0 = F^{q}$, $\mathbb{1} \in \mathbb{R}^{L}$, $1 \le i \le g$ and $\mathcal{A}$ denotes the attention map. $F^{a/m}$ is substituted for V respectively in our method. In the equation above, calculating the result $\mathrm{BAN}(H, V; \mathcal{A}) \in \mathbb{R}^{d}$ and adding it to the H is repeated in g times. Afterwards, H represents the combined visual and language features in the question space incorporating diverse aspects from the two modalities (Yang et al., 2016).

3.3 Motion-Appearance Fusion

In this section, we introduce the Motion-Appearance Fusion module which is our key contribution. Depending on what the question ultimately asks about, the model is supposed to decide which features are more relevant among appearance and motion information, or a combination of both. To do this, we produce appearance-centered, motion-centered, and all-mixed features and aggregate them depending on question context. Based on the previous step, we obtain cross-modal combined

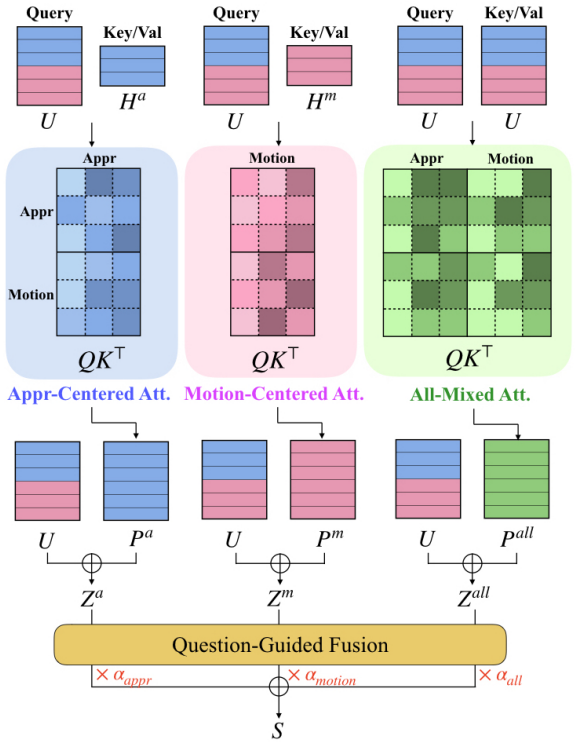

Figure 2: Motion-Appearance Fusion module. The blue-colored elements in a matrix denote appearance-question, and the pink ones indicate motion-question combined features. Matrices above $QK^\top$ represent an attention score maps from each kind of attention. The final output S in the figure is the weighted-sum matrix of all three attended features.

features H^a and H^m in terms of appearance and motion. We concatenate these two matrices and define U as:

$$U = \begin{bmatrix} H^a \\ H^m \end{bmatrix}, \; U \in \mathbb{R}^{2L \times d} \tag{6}$$

Motion-Appearance-centered Attention. We first define regular scaled dot-product attention to attend features to diverse aspects:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V \tag{7}$$

where Q, K, V denotes the query, key, and value, respectively. To obtain motion-centered, appearance-centered and mixed attention, we substitute U with the query, and H^a, H^m, U with the key and value in the equation 7 as:

$$\begin{aligned} P^a &= \text{Attention}(U,\ H^a,\ H^a) \\ P^m &= \text{Attention}(U, H^m, H^m) \\ P^{all} &= \text{Attention}(U,\ U,\ U) \\ Z^{a/m/all} &= \text{LayerNorm}(P^{a/m/all} + U) \end{aligned} \tag{8}$$

where $P \in \mathbb{R}^{2L \times d}$ and $Z \in \mathbb{R}^{2L \times d}$.

As in the first line of the equation 8, we add projected appearance features P^a on each appearance and motion feature to obtain Z^a, since the matrix U is the concatenation of H^a and H^m. Therefore, we argue that Z^a contains appearance-centered information. Similarly, $Z^{m/all}$ contains motion-centered and all-mixed features, respectively. We argue that the Motion-Appearance-centered attention fuses appearance and motion features in various proportions and these three matrices work like multi-head attention sharing the task of capturing diverse information, and become synergistic when combined.

Question-Guided Fusion. For question-guided fusion, we first define $\mathbf{z}^{a/m/all}$ as the sum of matrix $Z^{a/m/all} \in \mathbb{R}^{2L \times d}$ over sequence length $2L$. We obtain attention scores between each $\mathbf{z}^{a/m/all}$ and question context vector $\mathbf{q}$:

$$\alpha^{a/m/all} = \text{softmax}(\frac{\mathbf{q}(\mathbf{z}^{a/m/all})^\top}{\sqrt{d_z}}) \tag{9}$$

where $\mathbf{q}$ denotes the last hidden vector. The attention score $\alpha^{a/m/all}$ can be interpreted as the importance of each matrix Z based on question context. We obtain the question-guided fusion matrix O as:

$$\begin{aligned} S &= \alpha^a Z^a + \alpha^m Z^m + \alpha^{all} Z^{all} \\ O &= \text{LayerNorm}(S + \text{FFN}(S)) \end{aligned} \tag{10}$$

where $O \in \mathbb{R}^{2L \times d}$ is obtained by linear transformation and a residual connection after weighted sum. We aggregate information by attention over the sequence length of O:

$$\begin{aligned} \beta_i &= \text{softmax}(\text{FFN}(O_i)) \\ \mathbf{f} &= \sum_{i=1}^{2L} \beta_i O_i \end{aligned} \tag{11}$$

The final output vector $\mathbf{f} \in \mathbb{R}^d$ is used for answer prediction.

3.4 Answer Prediction and Loss Function

The video QA task can be divided into counting, open-ended word, and multiple-choice tasks (Jang et al., 2017). Our method trains the model and predicts the answer based on the three tasks similar to previous work.

The counting task is formulated as a linear regression of the final output vector $\mathbf{f}$. We obtain the

final answer by rounding the result and we minimize Mean Squared Error (MSE) loss.

The open-ended word task is essentially a classification task over the whole answer set. We calculate a classification score by applying a linear classifier and softmax function on the final output $\mathbf{f}$ and train the model by minimizing cross-entropy loss.

For the multiple-choice task, like in previous work (Jang et al., 2017), we attach an answer to the question and obtain M candidates. Then, we obtain the score for each of the M candidates by a linear transformation to the output vector $\mathbf{f}$. We minimize the hinge loss within every pair of candidates, $max(0, 1 + s_n - s_p)$, where s_n and s_p are scores from incorrect and correct answers respectively.

4 Experiments

In this section, we evaluate our proposed model on three Video QA datasets: TGIF-QA, MSVD-QA, and MSRVTT-QA. We first introduce each dataset and compare our results with the state-of-the-art methods. Then, we report ablation studies and include visualizations to show how each module in MASN works.

4.1 Datasets

TGIF-QA (Jang et al., 2017) is a large-scale dataset that consists of 165K QA pairs collected from 72K animated GIFs. The length of video clips is very short, in general. TGIF-QA consists of four types of tasks: Count, Action, State transition (Trans.), and FrameQA. Count is an open-ended question to count how many times an action repeats. Action is a task to find action repeated at certain times, and Transition aims to identify a state transition over time. Both types are multiple-choice tasks. Lastly, FrameQA is an open-ended question that can be solved from just one frame, similar to image QA.

MSVD-QA & MSRVTT-QA (Xu et al., 2017) are automatically generated from video descriptions. It consists of 1,970 video clips and 50K and 243K QA pairs, respectively. The average video lengths are 10 seconds and 15 seconds respectively. Questions belong to five types: what, who, how, when, and where. The task is open-ended with a pre-defined answer sets of size 1,000 and 4,000, respectively.

Methods	Count	Action	Trans.	FrameQA
ST-VQA	4.28	60.8	67.1	49.3
Co-Mem	4.10	68.2	74.3	51.5
PSAC	4.27	70.4	76.9	55.7
STA	4.25	72.3	79.0	56.6
HME	4.02	73.9	77.8	53.8
HGA	4.09	75.4	81.0	55.1
L-GCN	3.95	74.3	81.1	56.3
QueST	4.19	<u>75.9</u>	81.0	**59.7**
HCRN	<u>3.82</u>	75.0	<u>81.4</u>	55.9
MASN	**3.75**	**84.4**	**87.4**	<u>59.5</u>

Table 1: State-of-the-art comparison on the TGIF-QA dataset. Mean ℓ_2 loss for Count, and accuracy (%) for others. Best results in bold, underlined results denote the second best.

Methods	MSVD-QA	MSRVTT-QA
ST-VQA	31.3	30.9
GRA	32.0	32.5
Co-Mem	31.7	32.0
HME	33.7	33.0
HGA	34.7	<u>35.5</u>
QuesT	<u>36.1</u>	34.6
HCRN	<u>36.1</u>	**35.6**
MASN	**38.0**	35.2

Table 2: State-of-the-art comparison on the MSVD-QA and MSRVTT-QA datasets. All values represent accuracy (%). Best results in bold, underlined results denote the second best.

4.2 Implementation Details

We first extract frames with 6 fps for all datasets. In the case of **appearance features**, we sample 1 frame out of 4 to avoid information redundancy. We apply Faster R-CNN (Ren et al., 2016) pre-trained on Visual Genome (Krishna et al., 2017) to obtain local features. The number of extracted objects is $N = 10$. For global features, we use ResNet-152 pre-trained on ImageNet (Deng et al., 2009). In the the case of **motion features**, we apply I3D pre-trained on the Kinetics action recognition dataset (Kay et al., 2017). For the input of I3D, we concatenate a set of 8 frames around the sampled frame mentioned above. In terms of training details, we employ Adam optimizer with learning rate as 10^{-4}. The number of BAN glimpse g is 4. We set the batch size as 32 for the Count and FrameQA tasks and 16 for Action and Trans. tasks.

Methods		Count	Action	Trans.	FrameQA
Appr. Module		3.94	82.9	86.2	58.6
Motion Module		3.84	82.5	86.2	51.0
Appr. Module + Motion Module		3.82	83.4	86.8	58.6
Appr. Module + Motion Module + Fusion (Ours)		**3.75**	**84.4**	<u>87.4</u>	**59.5**
Single-Att. Fusion	Appr.	3.78	82.8	86.3	58.9
	Motion	3.79	83.1	87.0	59.1
	All	3.78	83.6	<u>87.4</u>	<u>59.3</u>
Dual-Att. Fusion	Appr. + Motion	<u>3.77</u>	83.6	<u>87.4</u>	59.2
	Appr. + All	<u>3.77</u>	83.6	**87.5**	59.0
	Motion + All	3.80	<u>84.1</u>	86.5	59.1

Table 3: Ablation study on the TGIF-QA dataset. Mean ℓ_2 loss for Count, and accuracy (%) for others. Appr. and Att. stand for Appearance and Attention. Best results in bold, underlined results denote the second best.

4.3 Comparison with State-of-the-arts

We compare MASN with state-of-the-art (SoTA) models on the aforementioned datasets.

TGIF-QA. Compared with ST-VQA (Jang et al., 2017), Co-Mem (Gao et al., 2018), PSAC (Li et al., 2019), STA (Gao et al., 2019), HME (Fan et al., 2019), and recent SoTA models: HGA, L-GCN, QueST, HCRN (Jiang and Han, 2020; Huang et al., 2020; Jiang et al., 2020; Le et al., 2020), MASN shows the best results for three tasks: Count, Trans., and Action, outperforming the baseline methods by a large margin as shown in Table 1. In the case of FrameQA, the performance is similar to QueST. However, considering that there exists some tradeoff between the performance of Count and FrameQA since Count focuses on identifying temporal information and FrameQA focuses on spatial information, MASN shows the best overall performance on the entire task.

MSVD-QA & MSRVTT-QA. As shown in Table 2, MASN outperforms the best baseline methods, QuesT and HCRN by approximately 2% on MSVD-QA, and shows competitive results on MSRVTT-QA. Since these datasets are composed of wh-questions, such as what or who, the question sets seemingly resemble FrameQA in TGIF-QA, as they tend to focus on spatial appearance features. This means that MASN is able to capture spatial details well based on the spatiotemporally mixed features.

4.4 Ablation Study

Analyzing the impact of motion module and appearance module. We investigate the effect of each module as seen in Figure 1. In Table 3, the 1^{st} and 2^{nd} row represent the result of using only the Appearance and Motion module, respectively. The 3^{rd} row shows the result of just concatenating appearance and motion features from each module and flattening them, by substituting the input X for O in equation 11. Most existing SOTA models utilize only ResNet features for spatio-temporal reasoning based on the difference of vectors over time. Using only the Appearance module is similar to most of these existing methods, which can catch spatio-temporal relations relatively well. On the other hand, we found that the accuracy on FrameQA when only using the Motion module is about 7% lower than when using the Appearance module. This means the Motion module is limited in its ability to capture the appearance details. However, comparing the 1^{st} and 3^{rd} row in Table 3, the performance in the Action and Trans. tasks increase consistently when the Motion module is added compared to using only the Appearance module. This indicates that the Motion module is a meaningful addition. Lastly, compared to the 1^{st}, 2^{nd} and 3^{rd} row, when integrating the output from both modules there is a further overall performance improvement. This indicates a synergistic effect occurs when integrating both the appearance and motion feature after obtaining them as high-level features.

Analyzing the impact of fusion module. We show ablation studies inside the fusion module represented in Table 3. The 4^{th} row indicates the performance of our proposed MASN architecture. The results in the 'Single-Attention Fusion' row use only one type of attention among appearance-centered, motion-centered, and all-mixed as seen in equation 8. The results in the 'Dual-Attention

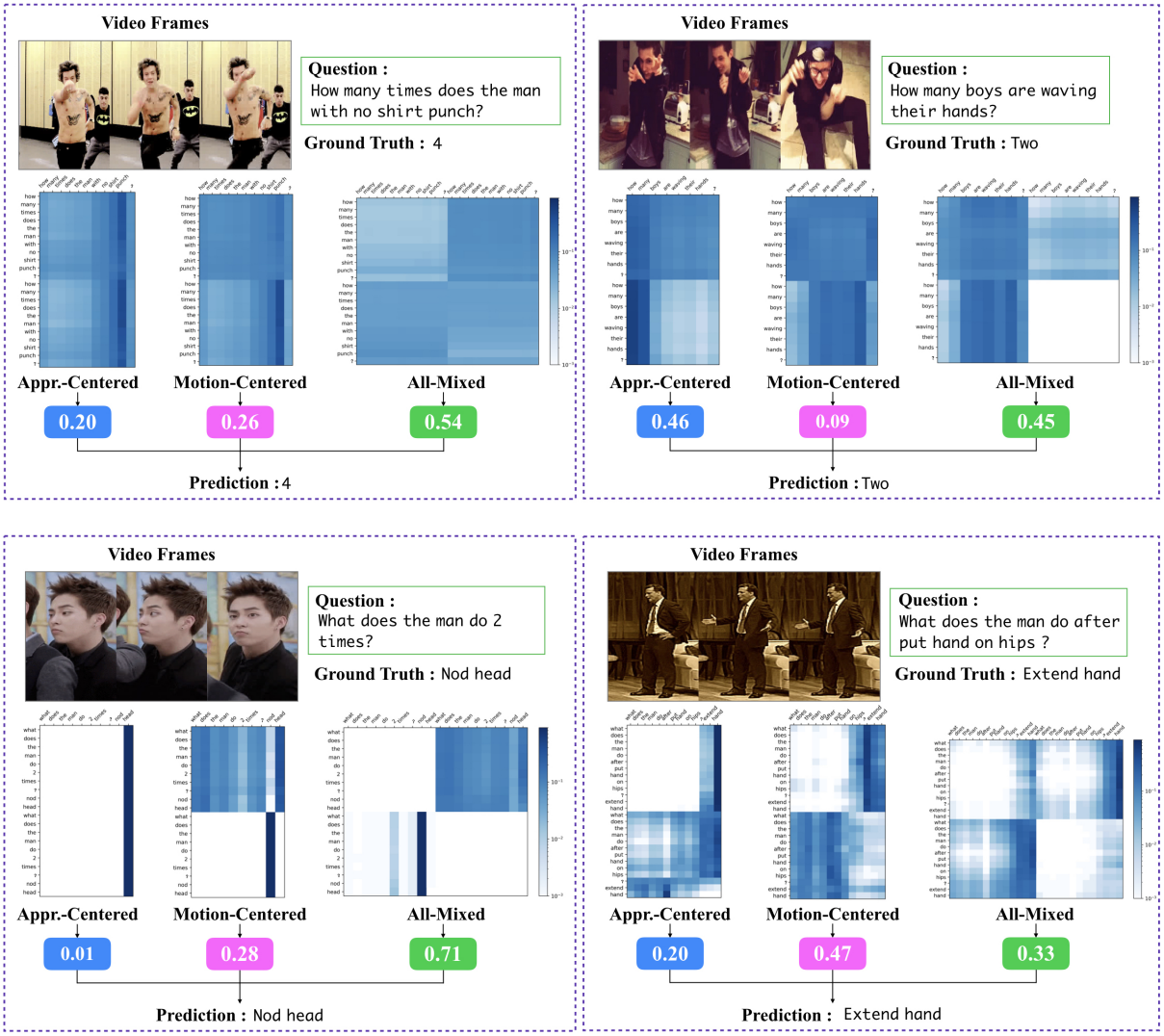

Figure 3: Qualitative results on TGIF-QA dataset. From the left, Count and FrameQA are shown in 1st row and Action, Trans. in 2nd row. Each visualized attention map is log-scaled. Scores below attention maps represent α from the equation 9.

Fusion' row utilize two among the three types of attention mentioned above. Due to the nature of video, when a question such as "How many times does the man in the white shirt put his hand on the head?" is given, the model is supposed to find the motion information "put" while catching the appearance information "man in white shirt" or "hand on head", and finally mixing them in different proportions depending on the context of question. Comparing the result of the 3^{rd} (without fusion) row and MASN first, MASN shows better performance across tasks. This means mixing appearance and motion features in various proportions using the Motion-Appearance-centered Fusion module and computing the weighted fusion via the Question-Guided Fusion module contributes to the performance. When comparing the general performance with the number of attention types in fusion module, using single, dual, and triple attention (MASN) shows increasingly better performance in the same order. This indicates that focusing on different aspects and integrating each attended feature performs better than calculating attention at once. Additionally, comparing the result of using only appearance or motion-centered attention in 'Single' with both of them in 'Dual', we find that using both features shows better performance, which means they play complementary roles for each other. Similarly, we argue the reason for the performance increase in FrameQA in the 'Motion' row of 'Single-Att. Fusion' is due to the fact that the model can find relevant appearance information better based

on motion information.

4.5 Qualitative Results

We give examples of each attention score matrix from Motion-Appearance Fusion module in Figure 3. We draw two conclusions from the Figure: (1) each attention map catches different relations similarly to multi-head attention, (2) each attention map is used to a different extend depending on the type of task. For example, in FrameQA, the appearance-centered's attention map captures which appearance trait to find focusing on 'how many'. On the other hand, the motion-centered's and all-mixed's attention map attend on 'waving' or 'hands' to catch motion-related information. In Action, similar to FrameQA, the appearance-centered's attention map attends on 'head' which is the object of action, while the motion-centered's attention map catch 'nod' which is related to movement. However, in the case of the Count task, the two attention weights are not as sparse as scores in the other tasks. We think this dense attention map causes the inconsistency in the performance increase between Count task and Action and Trans. task, although questions for all of these three tasks ask for motion information.

5 Conclusion

In this paper, we proposed a Motion-Appearance Synergistic Networks to fuse and create a synergy between motion and appearance features. Through the Motion and Appearance modules, MASN manages to find motion and appearance clues to solve the question, while modulating the amount of information used of each type through the Fusion module. Experimental results on three benchmark datasets show the effectiveness of our proposed MASN architecture compared to other models.

Acknowledgement The authors would like to thank Ho-Joon Song, Yu-Jung Heo, Bjorn Bebensee, Seonil Son, Kyoung-Woon On, Seongho Choi and Woo-Suk Choi for helpful comments and editing. This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (2015-0-00310-SW.StarLab/25%, 2017-0-01772-VTT/25%, 2018-0-00622-RMI/25%, 2019-0-01371-BabyMind/25%) grant funded by the Korean government.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Yoshua Bengio, Paolo Frasconi, and Patrice Simard. 1993. The problem of learning long-term dependencies in recurrent networks. In *IEEE international conference on neural networks*, pages 1183–1188. IEEE.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Seungchan Lee, Minsu Lee, and Byoung-Tak Zhang. 2020. Dramaqa: Character-centered video story understanding with hierarchical qa. *arXiv preprint arXiv:2005.03356*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1999–2007.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211.

Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585.

Lianli Gao, Pengpeng Zeng, Jingkuan Song, Yuan-Fang Li, Wu Liu, Tao Mei, and Heng Tao Shen. 2019.

Structured two-stream attention network for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6391–6398.

Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. 2018. A better baseline for ava. *arXiv preprint arXiv:1807.10066*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *AAAI*, pages 11021–11028.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766.

Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. 2020. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *AAAI*, pages 11101–11108.

Pin Jiang and Yahong Han. 2020. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI*, pages 11109–11116.

Gi-Cheon Kang, Jaeseo Lim, and Byoung-Tak Zhang. 2019. Dual attention networks for visual reference resolution in visual dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2024–2033.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018a. Bilinear attention networks. *arXiv preprint arXiv:1805.07932*.

Kyung-Min Kim, Seong-Ho Choi, Jin-Hwa Kim, and Byoung-Tak Zhang. 2018b. Multimodal dual attention memory for video story question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 673–688.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9972–9981.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE.

Kyoung-Woon On, Eun-Sol Kim, Yu-Jung Heo, and Byoung-Tak Zhang. 2020. Cut-based graph learning networks to discover compositional structure of sequential video data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5315–5322.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 2048–2057.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.