Video Captioning with Commonsense Knowledge Anchors

by

Huiliang Shao

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2022 by the
Graduate Supervisory Committee:

Yezhou Yang, Chair
Suren Jayasuriya
Chaowei Xiao

ARIZONA STATE UNIVERSITY

May 2022

ABSTRACT

It is not merely an aggregation of static entities that a video clip carries, but also a variety of interactions and relations among these entities. Challenges still remain for a video captioning system to generate natural language descriptions focusing on the prominent interest and aligning with the latent aspects beyond observations. This work presents a **C**ommonsense knowledge **A**nchored **V**ideo c**A**ptio**N**ing (dubbed as CAVAN) approach. CAVAN exploits inferential commonsense knowledge to assist the training of video captioning model with a novel paradigm for sentence-level semantic alignment. Specifically, commonsense knowledge is queried to complement per training caption by querying a generic knowledge atlas ATOMIC, and form the commonsense-caption entailment corpus. A BERT based language entailment model trained from this corpus then serves as a commonsense discriminator for the training of video captioning model, and penalizes the model from generating semantically misaligned captions. With extensive empirical evaluations on MSR-VTT, V2C and VATEX datasets, CAVAN consistently improves the quality of generations and shows higher keyword hit rate. Experimental results with ablations validate the effectiveness of CAVAN and reveals that the use of commonsense knowledge contributes to the video caption generation.

DEDICATION

I dedicate this work to my family, friends, labmates and advisors, who have been supporting me to pull through.

ACKNOWLEDGMENTS

First, I would like to offer my deepest appreciation to our committee chair and advisor, Dr. Yezhou Yang for giving me the opportunity to eplore such interesting field at APG group. I appreciate his patience with our long term dissertation relationship. His professional advice and knowledge throughout the course of research has been invaluable for the completion of the project. I would also like to thank my thesis committee members Dr. Suren Jayasuriya and Dr.Chaowei Xiao for their time and efforts in reviewing this thesis.

A special thanks to our lab mates, Zhiyuan Fang, for his constant motivation, inputs and advice throughout my time at the lab. Without his guidance, much of this work would not have been possible. I am also thankful to all the other lab members for their friendship and collaborations.

Finally I thank Arizona State University and the graduate program for allowing students to further their education. The staff and administration have always been friendly and responsive.

TABLE OF CONTENTS

CHAPTER Page

APPENDIX

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

1.1 Purpose Statement

This work was implemented to satisfy degree requirements for Masters of Science in Computer Science and course requirements for independent study with Professor Yezhou Yang. Video Captioning, as a popular task in the intersection area of Computer Vision and Natural Language Processing, aims at generating textual descriptions from video content. Challenge still remains for current captioning systems to describe observed daily events into narrative that semantically aligns with their contextual knowledge, i.e, probable causes, effects and attributes. The main purpose of the project is to design a novel training schema of video captioning model to generate more accurate and natural descriptions that aligns well with the latent aspects beyond observations. Additionally, the development of the model structure within the system contributes to better understand the video content by focusing on the prominent interest.

1.2 Intended Audience

The intended audiences for the work are the members of the graduate committee Dr. Yezhou Yang [chairperson], Dr. Suren Jayasuriya, and Dr. Chaowei Xiao; and anyone who are devoted to the vision and language field (including, but not limited to, Captioning, Visual Question Answering, Visual Reasoning) either extending this work or using this as a reference in their own work.

1.3 Problem Definition

1.3.1 Sentence-level Semantic Alignment

Human beings with extensive life experiences could describe observed daily events into narrative that semantically aligns with their contextual knowledge. For instance, given the video clips shown in Figure 1, one can identify the agent and the patient are "*people*" and "*food*" respectively by leveraging recognition, then supplement them with latent relations carrying interactions between the agent and the patient with multiple possibilities. The description could be as succinct as "*people are eating food*", or a verbose one, "*people are talking about food while eating*". Beyond straightforwardly narrating objects/entities of interest, an accumulation of good sense and sound judgement in practical matters connects them with latent relations, thus forming descriptions carrying prominent entities as well as suggesting probable causes, effects and attributes. In other words, *we say not only what we discern, but also reflects what we think and feel as well.*

Motivated by the example in Figure 1, we argue that a video captioning system benefits from aligning descriptions semantically *w.r.t.* an inferable context (causes, effects and attributes). Such a need has been recognized by a few image captioning work (Vinyals et al. 2015; Karpathy and Fei-Fei 2015; Yang et al. 2011). These approaches attempt to fill the gap between the perceivable entities and their latent and even obscure relationships by exploiting visual representation learning with direct supervisions. (Fu et al. 2016; C. Liu et al. 2016; Lu et al. 2017; You et al. 2016; Pedersoli et al. 2017) adopt the spatial attention mechanism with a goal to learn more descriptive visual representations. Similarly, (L. Gao et al. 2017; Zhang and Peng 2019; Chen et al. 2018; Yang, Han, and Wang 2017) adopt the temporal attention module for extracting informative frames in video captioning. However, stacking blocks solely in the visual encoder could ease the symptoms, but it does not fundamentally address the pain point of lacking commonsense aspects during the decoding phase.

Advancements made in image/video sequence-to-sequence translation domain reveal the benefits of adopting the evaluation metrics, *e.g.*, CIDEr (Vedantam, Zitnick, and Parikh 2014), BLEU(Papineni et al. 2002) and SPICE(Anderson et al. 2016) scores, as additional loss, together with a traditional word-level cross-entropy loss. More recently, reinforcement-based text generations, *e.g.*, policy gradient (X. Wang et al. 2018; S. Liu et al. 2017), actor-critic (L. Zhang et al. 2017; Z. Ren et al. 2017) formulate reward functions incorporating the phrase-matching metrics. Adopting the evaluation metrics as reward/loss to maximize/minimize boosts the performances measure by the same set of evaluation metrics. Even so, the performance gain are mostly ascribed to a distributional consistency among training and testing sets. By a brutal integration of the phrase-level evaluation metric based reward function could trigger severe overall semantic misalignment. Figure 1 shows such a failure case, where comparing to the ground-truth annotation: "*person is introducing the food*", a caption like "*person eating while talking*" achieves a higher SPICE score than its semantic inverse: "*person talking while eating*", even though the latter one is semantically more correct according to the ground-truth. In essence, the captioning performance training and evaluation done by the aforementioned metrics are unanimously constricted by the ground-truth annotations from the datasets, neglecting the latent and probably

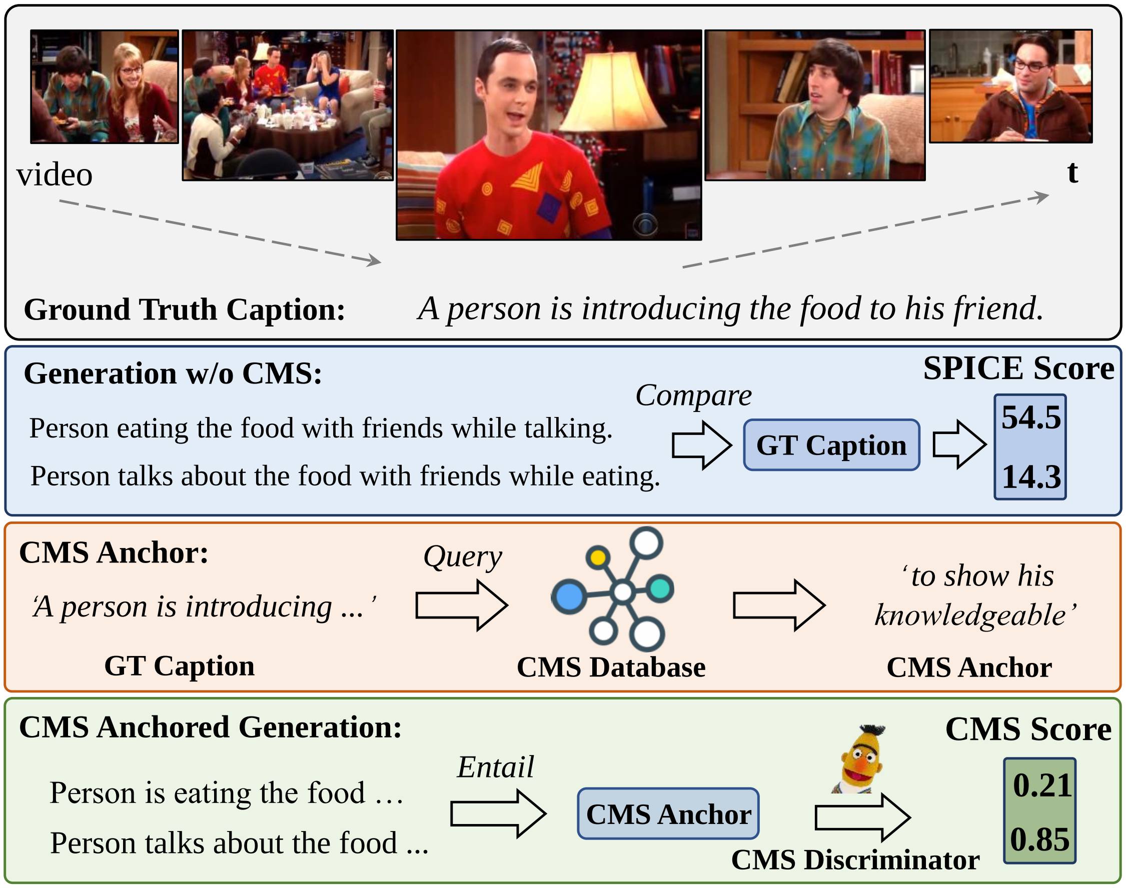

video
t
Ground Truth Caption:
A person is introducing the food to his friend.
Generation w/o CMS:
Person eating the food with friends while talking.
Person talks about the food with friends while eating.
Compare
GT Caption
SPICE Score
54.5
14.3
CMS Anchor:
'A person is introducing ... '
GT Caption
Query
CMS Database
'to show his knowledgeable'
CMS Anchor
CMS Anchored Generation:
Person is eating the food ...
Person talks about the food ...
Entail
CMS Anchor
CMS Discriminator
CMS Score
0.21
0.85

Figure 1. Commonsense Anchor

Note: We present CAVAN to address the semantic alignment for video captioning tasks using commonsense alignment (CMS-A). Different from traditional generation methods where the generations are only supervised by ground-truth annotations using token-level or phrase-matching metrics (*e.g.*, Cross-Entropy, CIDEr and BLEU). CMS-A leverages commonsense knowledge as anchors to constrain the overall semantics of the generated captions from deviating the current latent context.

inferable context that is not explicitly expressed by caption annotations, *i.e.*, cause and effect, entity and its attributes at sentence-level.

1.3.2 Multi-modal Visual Feature Fusion

To fulfill video captioning tasks, a robust overall video encoding is required to incorporate multi-modal visual features with higher-order interactions. Previous works (Yao et al. 2015; Yu et al. 2015; B. Wang et al. 2018; P. Pan et al. 2015; Y. Pan et al. 2015) always extract appearance features of video frames and motion features of video segments to represent global information of video contents. The most recent works (Z. Zhang et al. 2020; B. Pan et al. 2020; Zhang and Peng 2019; Hu et al. 2019) pre-train an object detector to extract salient object features from video keyframes. To fuse the global and local features, existing research either apply simple concatenation (N. Xu et al. 2018), or a polynomial feature fusion (Jiyang Gao et al. 2017) while ignore the spatio-temporal relations. Z. Zhang et al. 2020 propose a hierarchical decoder with a temporal-spatial attention module to generate global and local context feature and fuse them simply by a concatenation operation, which fails to exploit higher-order interactions among visual features. It is attractive to develop a fusion mechanism that has the capability to selectively capitalize on visual information and exploit both spatio-temporal relations and higher-order interactions between the input multi-modal features.

1.4 Contribution

In this work, we propose a novel model supervised by a sentence-level metric ensuring semantic alignment exploiting commonsense knowledge. Specifically, we first design a fusion module that reasons over multi-model visual features and dynamically aggregates them to obtain high-level semantic feature, which is conducive to infer the otherwise neglected sentence-level context. We then leverage a commonsense knowledge atlas to query semantic anchors carrying the inferable context, and adopt a sentence level entailment score comparing generated caption with the retrieved anchors as a semantic consistency measure. We present the **C**ommonsense knowledge **A**nchored **V**ideo c**A**ptio**N**ing (dubbed as CAVAN), where the commonsense entailment loss is introduced for the first time to supplement the existing caption generation supervisions. To the best of our knowledge, this work is the first that leverages the complementary commonsense knowledge thus imposes additional contextual constraints for video captioning training, and ultimately generates captions with better aligned sentence-level semantic.

We compile a complementary set of commonsense knowledge by querying caption annotations from the ATOMIC dataset (Sap et al. 2019) and a human curate commonsense annotations of captions from the V2C dataset (Fang et al. 2020), then retrieving a set of probable causes, effects and attributes. With the augmented and paired (caption, commonsense knowledge) data, we train a generic natural language entailment model based on BERT (Devlin et al. 2018) to serve as a discriminator during training by evaluating the entailment score of each generated caption (see Figure 2). Empirically, we test and observe that our CAVAN model achieves significant improvements over the baseline models and achieve competitive results with previous state-of-the-art video captioning methods. We further provide in-depth analysis of each critical sub-modules within CAVAN by extensive ablation experiments. We summarize our contributions as:

- CAVAN is the first to leverage commonsense knowledge to assist the training of video captioning model, and shows quality of improvement for generations.

- We carefully design a novel fusion module to reason over and capture the higher-order interactions between multi-modal features.
- CAVAN achieves state-of-the-art performances on both V2C, MSR-VTT and VATEX testing beds for video captioning task over a standard set of automated metrics.
- Our ablations on CAVAN comprehensively analyze the effect of incorporating different types of knowledge and modules, providing guiding insights for future research.

Chapter 2

RELATED WORK

2.1 Video Captioning

Traditional captioning systems (Venugopalan et al. 2015; Yao et al. 2015; Karpathy and Fei-Fei 2015; K. Xu et al. 2016) are trained typically with a teacher-forcing (Bengio et al. 2015) manner and evaluated using discrete and non-differential metrics. However, such training schema suffers from exposure bias (Ranzato et al. 2015) and the inconsistency between the optimizing function and evaluation metrics. Recent work (S. Liu et al. 2016; L. Zhang et al. 2017; Bahdanau et al. 2017; Junlong Gao et al. 2019) introduce Reinforcement-Learning (RL) techniques based on policy gradient to tackle these issues. Specifically, Ranzato *et al.* (Ranzato et al. 2015) adopt REINFORCE algorithm to sequence training with RNNs via treating the task metrics as optimization objectives. Later, Rennie *et al.* (Rennie et al. 2017) directly optimize CIDEr metric with a self-critical sequence training (SCST) approach that harmonizes the model with respect to its test-time inference procedure. Though optimizing towards the automatic metrics yields impressive benchmark results, these metrics tend to neglect the essential need of semantic alignment. In our work, we further incorporate a sentence-level semantic score into the reinforced objectiveness following the REINFORCE training strategy.

2.2 Higher-order Interactions among Multi-modal Visual Features

Yao *et al.* (Yao et al. 2015) is the first to introduce a C3D visual encoder with a attention mechanism to dynamically model the video's global temporal structure. Yu et al. 2015; Aafaq et al. 2019 make efforts to design attention mechanisms to effectively capture spatio-temporal dynamics of the video content. More recently, graph based visual representations have been exploited in video captioning. B. Pan et al. 2020; Z. Zhang et al. 2020 captures more detailed interaction information to learn discriminative spatio-temporal representations via building visual relation graphs. However, most of the previous work only concentrate on the exploration of 1^{st} order feature interactions, which is lacking in efficacy. In this work, we propose a novel fusion module to reason over multi-modal visual representations and learn higher-order feature interactions among them.

2.3 Commonsense Knowledge in Visual Understanding

It is becoming popular to make use of commonsense knowledge to mine the underlying semantics for visual understanding (Aditya, Yang, and Baral 2019; Fader, Zettlemoyer, and Etzioni 2014; Shah et al. 2019; Hou et al. 2020; P. Wang et al. 2018). Previous work (Zhou, Sun, and Honavar 2019; Hou et al. 2019) take advantage of external knowledge to augment the visual information, thus improving the quality of machine generated captions. For fine-grained video understanding, recent work aim to obtain inferable context beyond appearances. Fang *et al.* (Fang et al. 2020) present a generative model for commonsense video captioning to describe the latent aspects of an visual scene. More recently, Lei *et al.* (Lei et al. 2020) incorporate commonsense into text representations, proving helpful to address an interesting next event prediction task. Unlike the above work where commonsense knowledge solely serves as a guidance to improve the visual/text representations, we directly apply it to regulate the learning of the inherent visual semantics.

2.4 Semantic Alignment across Modalities.

As the semantic inconsistency between vision and language inevitably exists, a few approaches (You, Luo, and Zhang 2018; Fang et al. 2019; Wu et al. 2019; Z. Wang et al. 2020) attempt to learn such inter-modal correspondence. Karpathy *et al.* (Karpathy and Li 2014) learn the latent alignment between the two modalities through a common embedding space and a structured objective. Dognin *et al.* (Dognin et al. 2019) present a co-attention discriminator to score the similarity of the visual and linguistic representations and enforce semantic alignments among them. Recently, instead of directly aligning visual features and linguistic tokens, Guo *et al.* (Guo et al. 2019) hierarchically align the tokens with visual relations. The above methods learns effective yet limited semantic consistency from existing data. Few efforts exploit the extra semantic alignment with the inferential context from commonsense knowledge.

Chapter 3

COMMONSENSE KNOWLEDGE RETRIEVAL PIPELINE

3.1 Commonsense Knowledge Base: Atomic

We query from a external knowledge atlas ATOMIC (Sap et al. 2019) to complement each video caption with 3 types of complementary commonsense descriptions (intention/attribute/effect) as commonsense anchors for training. More concretely, ATOMIC is an atlas of everyday commonsense knowledge and consists of 880k triplets of annotations that contain causes, effects, attributes of human activities/events as an if-then relations. Given an observed event, ATOMIC provides unobserved related causes and effects: what might happened just before, what might happen next as a result, and how different events are chained through causes and effects. For instance, as depicted in Figure 2, when observing the event *"X repels Y's attack."*, ATOMIC provides plausible facts in relation to the events. As for the pre-conditions prior to the event, X might wanted to save themselves. Regarding the plausible motivations, X might have been trained hard enough to rebel Y's attack. As a result of the event, X might gain an enemy while Y, on the other hand, might wants to attack X again. It can be inferred from the event that X is brave, strong and skilled.

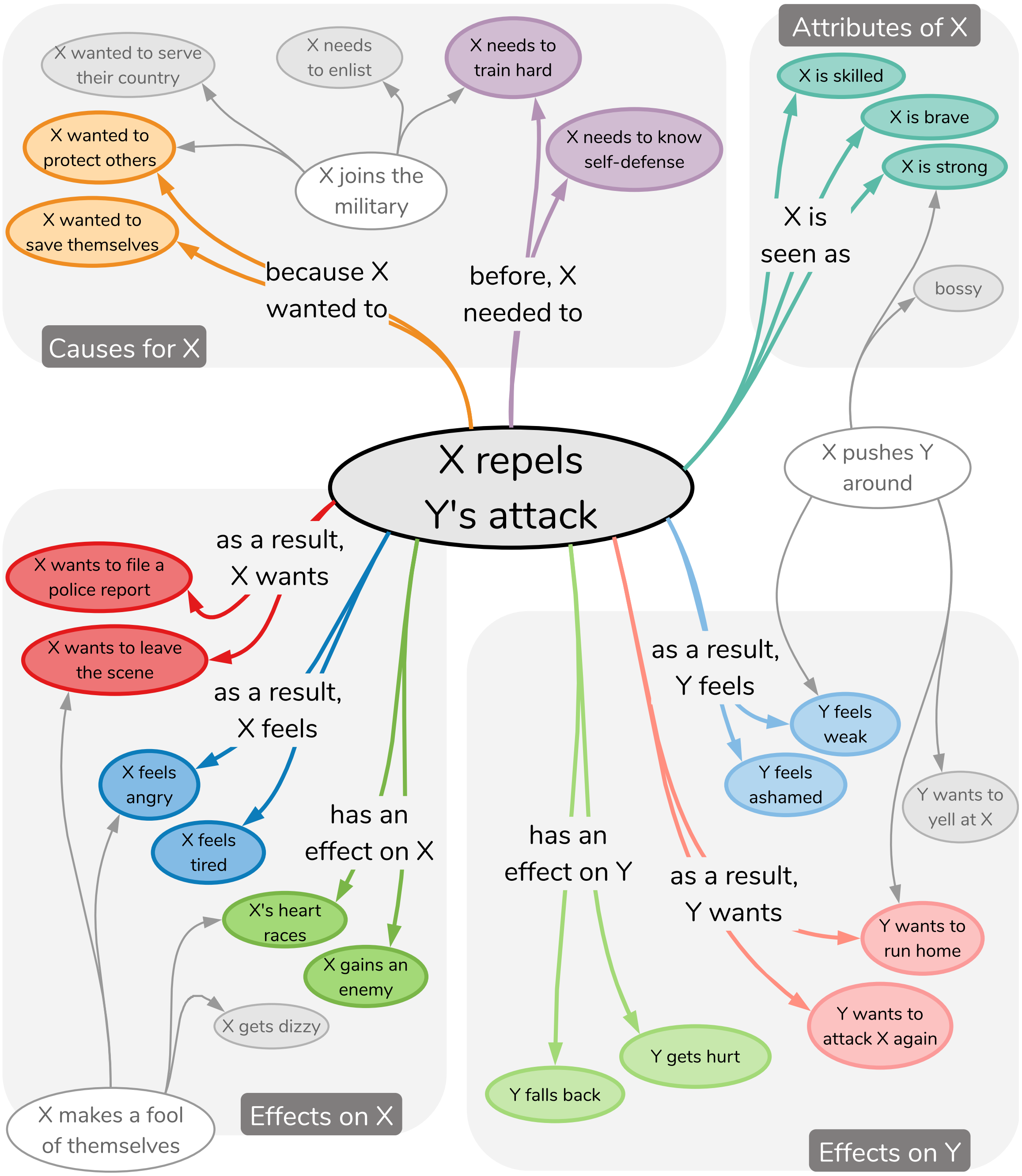


Figure 2. A subset of ATOMIC

Source:**sap_ATOMIC_2019**

Note: ATOMIC is an atlas of machine commonsense for everyday events, causes, and effects.

3.2 Step 1: Event and Caption Embeddings Generation

To augment each caption with commonsense knowledge, we need to extract out the most similar ATOMIC event *w.r.t* each caption and take the plausible inference provided by ATOMIC as complementary knowledge. It's required to project and encode both captions and events in the same embedding space prior to comparing their similarity. We extract out the key nouns and verbs of the each event in ATOMIC and encode them into word vectors based on pretrained GloVe embeddings[1]. Then we take the addition of word vectors for key nouns and verbs as final embeddings of each event. Similarly, we get the final encodings for each ground-truth caption by adding up word embeddings of key nouns and verbs.

3.3 Step 2: Commonsense-Caption Pairing

Based on the pre-computed embeddings in Sec. 3.2, each caption is paired with the most similar event by comparing their cosine similarities. The surrounding inferential facts (intention/attribute/effect) related to the paired event are retrieved as complementary commonsense anchors for each caption.

[1] GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Files with the pre-trained vectors Glove can be found in many sites like Kaggle or in the previous link of the Stanford University. We will use the glove.6B.100d.txt file containing the glove vectors trained on the Wikipedia and GigaWord dataset.

3.4 Step 3: Refinement Based on Ranking Score

Following Fang et al. 2020, we pre-train a Bert discriminator to get a more reasonable subsets of commonsense descriptions for each type of knowledge associated with the events for each caption. To be more specific, a Bert discriminator is pretrained on the entailment task which predicts a binary relation between two textual descriptions - whether the meaning of one text fragment is inferred from the meaning of the other text fragment.

For the training data, we choose event sentence and its corresponded commonsense description as positive pair, and another random commonsense sentence from the ATOMIC as a negative pair. In total, we have 230,624 event-commonsense pairs constructed, with 70% for training, and 30% for testing. Our discriminator achieves 85% accuracy on the testing split.

We then select the top-3 most plausible commonsense descriptions for each type of knowledge associated with the events by ranking scores produced by the pre-trained Bert discriminator.

3.5 Commonsense Retrieval Example

An Example of retrieved commonsense knowledge is shown in Figure 3. Given the ground-truth caption, "*A woman is practicing some movements in dancing room.*", we query three types of commonsense knowledge from ATOMIC: As for the pre-conditions of practice dancing, the woman might need to perform on stage. In terms of the motivation behind the event, the woman might want to be a better dancer. For the

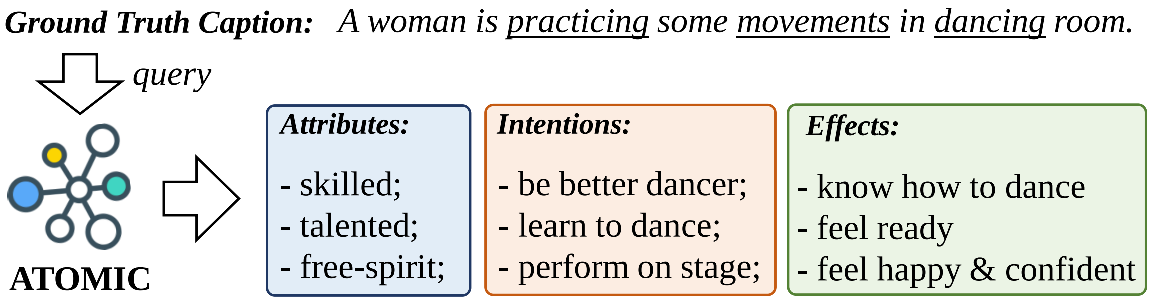


Figure 3. An example of retrieved commonsense knowledge

Note: Inferential commonsense knowledge retrieved from ATOMIC includes several types, *e.g.*, intentions, effects and attributes of the agents.

characteristics of the woman, she can be described as talented, skilled. As a result, she might become more confidence and feel ready for her debut on stage.

Chapter 4

METHODOLOGY

4.1 Overview of the Framework

CAVAN's backbone is an encoder-decoder architecture based on the transformer self-attention modules (Vaswani et al. 2017). A two-branch encoder takes the input of global and object features respectively, and produces attentively aggregated visual representations. Notably, we develop a novel module that dynamically reasons over attended features and alternatively fusing them based on the high-level interactions across modalities. A transformer decoder then generates the caption taking the visual representations from the specifically designed fusion module, and is supervised by both traditional video captioning losses (*i.e.*, smoothed cross-entropy) and the newly introduced commonsense entailment reinforcement loss (as shown in Figure 4).

4.2 Video Encoder

Given a sequence of video frames, a couple pre-trained networks are employed to extract both global (key frames or video snippets) and entity-level features (local regional features for objects) to form a holistic representations. Specifically, we obtain the per-frame features from ImageNet (Deng et al. 2009) pre-trained 2D recognition network by sampling one key frame from every 32 frames, $V^f = [v_1^f \dots v_T^f]$, with T denotes the temporal length of videos. For motion signals, we encode every non-overlapping 32 frames by a 3D activity recognition network (Carreira and Zisserman

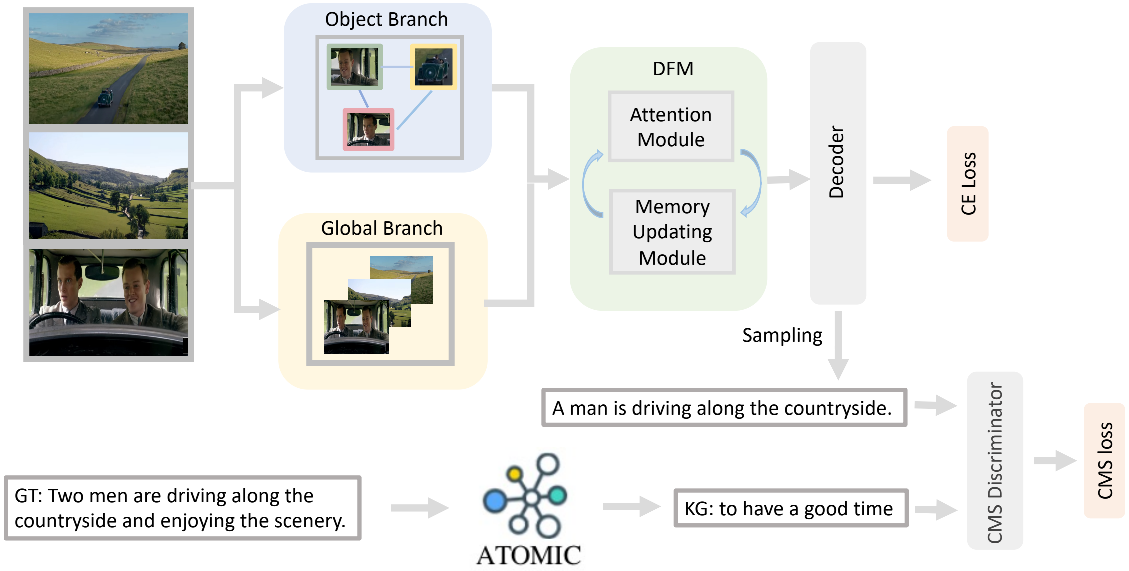


Figure 4. Overview of our proposed framework.

Note: CAVAN consists of transformer-based encoders and decoders, a dynamic fusion module, and a commonsense discriminator. Our model adopts a two-branch structure that generates attended object and global representations respectively. A fusion module is then adopted to fuse the outputs of two branches for decoding. The final predicted probability distribution is under the supervision of traditional cross-entropy loss. Meanwhile, a commonsense entailment loss is applied to guide the semantic alignment between current decoding descriptions and commonsense knowledge queried from ATOMIC Sap et al. 2019.

2017), and yields $V^m = [v_1^m \dots v_T^m]$. Following recent work in video captioning (Z. Zhang et al. 2020; B. Pan et al. 2020), we extract features of the class-agnostic object proposals sampled from keyframes of the input video. Then typical candidates proposals are obtained by clustering on the sampled candidates proposals and represented by the cluster centers. Let $V^o = [v_1^o \dots v_N^o]$ denote the features of typical object proposals, where N is the number of object proposals.

We directly adopt the transformer-based visual encoder for encoding global and object features separately. Specifically, the object branch passes the features of candidate proposals V^o and generates enhanced local representations $L = [l_1 \dots l_N] \in \mathcal{R}^{N \times d}$ with interaction message between objects. The global branch takes the concatenation of appearance features V^f and motion features V^m as inputs to produce a global embedding $G = [g_1 \dots g_T] \in \mathcal{R}^{T \times d}$ of a temporal sequence, which provides additional global context that may be missing in the object branch.

4.3 Dynamic Fusion Module

Effective video captioning calls for a robust overall video encoding. It is critical for such encoding to incorporate representations with higher-order interactions. Existing research either apply simple concatenation (N. Xu et al. 2018), or a polynomial feature fusion (Jiyang Gao et al. 2017). Apart from that, **D**ynamic **M**emory **N**etworks (DMN) (Kumar et al. 2015) has been applied in tasks across domains that require higher-order interactions among features, and is shown to be effective in VQA (Li, Su, and Zhu 2017).

In CAVAN, we propose the **D**ynamic **F**usion **M**odule which builds on an attention module and a memory update module (dubbed as DFM). The attention module is responsible for producing global contextual representations from global features with relevance inferred by typical object features and previous memory status. Then the memory update module renews its internal episodic memory based on the global contextual message, which has the ability to retrieve new global context that were considered to be irrelevant during previous iteration.

Formally, given the refined global features $G = [g_1 \dots g_N]$ and object representations $L = [l_1 \dots l_N]$ from visual encoders, an episodic memory $M = [m_1 \dots m_N]$ is initialized as $M^{(0)} = L$ and iteratively refined by an attention mechanism until the final step I is reached.

Attention Component: For the n_{th} object proposal, the attention is implemented by allowing the interaction between object feature vector $l_n \in L$ and both the global features $G = [g_1 \dots g_N]$ and previous memory states $m_n^{(i-1)} \in M^{(i-1)}$. The context $c_n^{(i)}$ is obtained by applying soft attention procedure as:

$$\begin{aligned} z_n^{(i)} &= \Big[\; G \odot l_n \; ; \; G \odot m_n^{(i-1)} \; ; \; |G - l_n| \; ; \\ & \quad |G - m_n^{(i-1)}| \; \Big]; \\ \alpha^{(i)} &= \textsc{Softmax}(W_2(\textsc{tanh}(W_1 z_n^{(i)} + b_1)) + b_2); \\ c_n^{(i)} &= \sum_{t=1}^{T} \alpha_t^{(i)} \cdot g_t, \end{aligned} \tag{4.1}$$

where $\odot$ denotes element-wise multiplication; $| \cdot |$ is the the element-wise absolute value; $[;]$ represents concatenation operation. $\alpha_t^{(i)}$ is the t_{th} element of $\alpha^{(i)}$ which denotes the normalized attention weight for g_t at i_{th} iteration. W_1, W_2, b_1 and b_2 are the parameters in the linear operation.

Memory Updating Component: The memory vector is updated as

$$m_n^{(i)} = \textsc{ReLU}(W_3[m_n^{(t-1)}; c_n^{(i)}; l_n] + b_3), \tag{4.2}$$

where W_3, b_3 are the parameters for the linear layer. m_n^i is the memory vector for n_{th} object proposal at the t_{th} iteration.

By the I_{th} iteration, the memory vector $m_n^{(I)}$ that memorizes the most relevant context from global features for n_{th} object proposal, is fused with the object vector l_n to generate globally contextualized object representations $\tilde{l}_n$ for decoding.

$$\tilde{l}_n = \textsc{ReLU}(W_4[l_n; m_n^{(I)}] + b_4), \tag{4.3}$$

where W_4, b_4 are the linear parameters.

4.4 Language Decoder

We design the language decoder by compiling a stack of transformer attention blocks using self-attention module. During training, it takes as input of the encoded word embedding and their corresponding positional encoding Vaswani et al. 2017 and attend to visual representations from the fusion module. The training criterion is based on cross-entropy loss $\mathcal{L}_{CE}$:

$$\mathcal{L}_{CE} = -\sum_{t=1}^{T} \phi(w_t^*)' \cdot \log(P(w_t)), \tag{4.4}$$

where T denotes the total training step of the ground-truth captions; $\mathbf{P}(w_t)$ represents the probability distribution across the vocabulary at time t; $\phi(w_t^*)$ is the one-hot vector of ground-truth word at time t.

4.5 Commonsense Entailment Loss

Supervising captioning model learning with existing short textual annotations largely limits training efficacy. The semantic carried by caption only is often with weak expressive power without latent inferable context. Instead, we leverage inferable commonsense knowledge to complement each video caption and treat them as additional constrains to regularize the generating process. In practice, we acquire the commonsense knowledge description k, *w.r.t.* the video caption by either retrieving from knowledge base (MSR-VTT + ATOMIC) or directly from human annotations (V2C). We discuss the commonsense knowledge retrieval procedure in Chapter 3.

Given a textual sequence $w^s = \{w_1^s \dots w_T^s\}$ sampled from language decoder, we regularize the generation by an entailment reward leveraging the commonsense description k. Intuitively, we encourage the model to caption by entailing the commonsense knowledge. To enable optimizing over non-differentiable metrics, previous efforts adopt the policy gradient approach (Ranzato et al. 2015; Rennie et al. 2017) and treat the task as a reinforcement learning one, with the testing metrics as the reward function. Particularly, Pasunuru *et al.* (Pasunuru and Bansal 2017) implement an entailment-enhanced score from a pre-trained model as the reward. Formally, with the policy gradient strategy, model like an active agent which generates word (as action) and the learning process is supervised by minimizing the negative expected reward function:

$$\mathcal{L}(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[r(w^s)], \tag{4.5}$$

where p_θ is the policy and θ is the model parameters.

CAVAN exploits commonsense-caption entailment score as a reward for training. We adopt a BERT model as the commonsense (CMS) discriminator $\mathcal{D}_{cms}$, which returns the entailment score for the caption and commonsense description pair. Following Fang et al. 2020, we pre-train the BERT model on ATOMIC dataset using the next sentence prediction task, whose input is an event description sentence and its associated commonsense description. Then, this BERT model is frozen and applied to our entailment score computation as offline. Further details for BERT pre-training are given in Section **??**. $\mathcal{D}_{cms}$ computes a probability (as SE score) for whether the sampled caption (w^s) entails the commonsense anchor:

$$r_{cms}(w^s) = \mathcal{D}_{cms}(w^s, k). \tag{4.6}$$

Here, the commonsense-caption entailment score essentially encodes whether the generated caption semantically aligns with the caption w.r.t. the sentence-level meaning. Applying $r_{cms}(w^s)$ to *e.q.* (4.5) yields a commonsense entailment loss $\mathcal{L}_{cms}$.

The gradient is estimated as follows:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{cms}(\theta) = -\mathbb{E}_{w^s \sim p_\theta}[(r_{cms}(w^s) - r_{cms}(\hat{w})) \\ \nabla_\theta \log p_\theta(w^s)],\end{aligned} \tag{4.7}$$

where $\hat{w}$ is the generated sequence obtained by the current model using greedy decoding. The corresponding entailment reward $r_{cms}(\hat{w})$ is seen as a baseline to reduce the variance of the gradient estimate without changing the expected gradient.

For our experiments, we also adopt the commonsense-caption entailment score as an extra evaluating metric on the testing split. Note that, the queried commonsense knowledge and $\mathcal{D}_{cms}$ are only needed to form supervision signal $\mathcal{L}_{cms}$, but not required during inference.

4.6 Training Loss

Putting all the loss terms together for an end-to-end training yields an overall optimization target:

$$\mathcal{L} = \mathcal{L}_{CE} + \beta \cdot \mathcal{L}_{cms}, \tag{4.8}$$

where β is a trade-off hyper-parameter weighting each loss term. During the training process, we freeze the CMS discriminator and compute the $r_{rms}(w^s)$ with an inference mode.

Chapter 5

EVALUATION

Evaluating natural language generation systems is a complex task. For this reason, a number of different metrics have been proposed for tasks such as captioning and machine translation etc. BLEU@4 (Papineni et al. 2002), ROUGE-L (Lin 2004), CIDEr (Vedantam, Zitnick, and Parikh 2014) and METEOR (Banerjee and Lavie 2005) are the widely used and popular metrics for captioning tasks. Previous metrics are mainly used to evaluate texts at corpus level, which fails to take into account the sentence-level semantic congruity and alignment. Thus we propose a Semantic Entailment score (dubbed as SE score) measuring the sentence-level alignment between candidate and reference sentences.

5.1 BLEU@4

BLEU, which stands for Bilingual Evaluation Understudy, is one of the most popular metrics to measure the quality of machine-generated texts based on their correspondence with human descriptions. BLEU scores quantify the overlap between predicted uni-gram (single word) or n-grams (sequence of n adjacent words) and a set of one or more ground-truth sentences. A description that has exact match of words and their order with ground-truth texts will get a high score on BLEU metric. However, BLEU metric is barely designed on corpus-level instead of sentence-level. BLEU score can be calculated as:

$$\log BLEU = \min(1 - l_r/l_p, 0) + \sum_{n=1}^{N} w_n \log_{p_n} \tag{5.1}$$

where l_r/l_p is the ratio between the lengths of reference sentences and predicted descriptions. w_n are positive weights, and p_n is the geometric average of the n-gram precisions. The second term calculates the exact matching score while the first term penalizes the sentence that has less words than the reference sentences.

5.2 ROUGE-L

ROUGE-L is a evaluation metric for text summaries, which computes recall and precision scores of the longest common subsequences (LCS) between the generated and each reference sentence. To be more specific, the subsequences refers to the common words that are in sequence but not strictly consecutive. The intuition behind the metric is that the longer LCS between reference sentences and predicted texts indicates high similarity between the two summaries. Unlike BLEU, pre-defined n-gram length is not required since this is automatically incorporated by LCS. ROUGE-L score is computed to find how similar summary A of length m is to summary B of length n:

$$ROUGE - L = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}} \tag{5.2}$$

where A is the ground-truth sentence and B is the candidate sentence. R_{lcs} and P_{lcs} are the recall and precision scores of the longest common subsequences (LCS) between the generated and each reference sentence. R_{lcs} and P_{lcs} can be obtained as:

$$\begin{aligned} R_{lcs} &= \frac{LCS(A,B)}{m} \\ P_{lcs} &= \frac{LCS(A,B)}{n} \end{aligned} \tag{5.3}$$

where $LCS(A,B)$ denotes the length of longest common subsequence between reference sentences and candidate sentences. ROUGE-L is equal to 1 When A and B is exactly the same *i.e.*, $LCS(A,B) = 1$, otherwise 0 in case A and B have no commonalities, *i.e.*, $LCS(A,B) = 0$.

5.3 METEOR

METEOR metric measures the alignment between generated and reference sentences. Each sentence is seen as a set of uni-grams and the alignment is done by mapping the uni-grams of generated and reference sentences. A uni-gram in predicted sentence should either map to a uni-gram in generated sentence or to zero. Meteor score is calculated using a uni-gram based F-score:

$$F_{mean} = \frac{10PR}{R + 9P} \tag{5.4}$$

where P, R denotes the unigram-based precision score and recall score respectively. The precision and recall score can be calculated as follows:

$$\begin{aligned} P &= \frac{m_c r}{m_c t} \\ R &= \frac{m_c r}{m_r t} \end{aligned} \tag{5.5}$$

where m_{cr} represents the number of uni-grams appears co-occurring in both generated and reference sentences. m_{ct} and m_{rt} are the number of uni-grams in generated sentences and reference sentence, respectively.

For METEOR metric, a penalty weight is placed on the uni-gram based F-score by using higher order similarities. The penalty is calculated by grouping the uni-grams into a minimum number of chunks that includes adjacent uni-grams in both candidate and reference sentences. The penalty P is computed as:

$$P = 1/2(\frac{N_c}{N_u}) \tag{5.6}$$

where N_c denotes the number of chunks and N_u corresponds to the number of uni-grams grouped together. METEOR score is computed by applying penalty weight P

to the uni-gram based F-score:

$$METEOR = F_mean(1 - P) \tag{5.7}$$

In case where multiple reference sentences are given, the maximum METEOR score of a generated and reference sentence pair is taken.

5.4 CIDEr

Recently, CIDEr metric has been proposed especially for captioning task. It evaluates the consensus between generated descriptions and one or more reference sentences by stemming. Specifically, all the words from candidate as well as reference sentences are stemmed into their root forms. For examples, *'sees', 'saw, 'seen','seeing'* should be converted to its stem word *'see'*. According to CIDEr, each sentence is decomposed into a set of n-gram including 1 to 4 words. By measuring the co-existence frequency of n-grams in both sentences, n-grams that are very common among the reference sentences of all the given visual data are assigned lower weight, as they are likely to be less informative about the visual content and more biased towards lexical structure of the sentences. The weight for each n-gram is computed using Term Frequency Inverse Document Frequency (TF-IDF), where the first term TF places higher weights on the frequently occurring n-grams in the reference sentences of the single sample, whereas IDF reduces the weight of n-grams that commonly appear across the whole dataset. The $CIDEr_n$ score for n-grams of length n is computed based on calculating the average cosine similarity between the predicted sentence c_i and the reference sentences s_{ij}:

$$CIDEr_n(c_i, s_{ij}) = \frac{1}{m} \sum_j \frac{\mathbf{g}^n(c_i)\mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\|\|\mathbf{g}^n(s_{ij})\|} \tag{5.8}$$

where m denotes the number of reference sentences for the single sample s_i. $\mathbf{g}^n(c_i)$ is a vector representing all n-grams with length n for predicted sentences and $\|\mathbf{g}^n(c_i)\|$ is the magnitude of $\mathbf{g}^n(c_i)$. Same is true for $\|\mathbf{g}^n(s_{ij})\|$, which is the vector representation for reference sentences.

The CIDEr is obtained by combining higher order n-grams of varying lengths to capture grammatical properties as well as richer semantics:

$$CIDEr(c_i, s_{ij}) = \sum_{n}^{N} w_n CIDEr_n(c_i, s_{ij}) \tag{5.9}$$

where w_n is the weight and uniformly set to $\frac{1}{N}$, which proves to work for best. N is empirically set to 4.

5.5 Semantic Entailment Score

Since previous evaluation metrics are mainly based on corpus-level measurements, it is not fair to measure the sentence-level semantic congruity and correctness by the pre-mentioned automatic metrics. Therefore, we propose an entailment based semantic score (SE score) to evaluate the sentence-level semantic alignment between generated captions and retrieved commonsense knowledge. Technically, commonsense knowledge that carries inferential context is queried from ATOMIC dataset for each ground-truth caption. Then a pre-trained discriminator is applied to compute SE score between generated captions and commonsense knowledge that is paired with the corresponding ground-truth captions. Formattly, given generated captions w_s and commonsense knowledge k:

$$SE(w_s) = \frac{1}{N} \sum_{n=1} ND_{cms}(w_s, k) \tag{5.10}$$

where D_{cms} is the pre-trained discriminator which is frozen and applied to SE score computation as offline. N represents the total number of generated samples. More technical details can be found in Chapter 3.

5.6 Human Evaluation

Considering low correlation with human judgments of automated evaluation metrics, human evaluations are critical for measuring the performance of machine-generated captions. It can either be done by crowd-sourced, *e.g.*, AMT workers, or specialist judges, as in some competitions. Such human evaluations can be achieved by two common measurements: *Relevance Rating* and *Grammar Correctness*. For *Relevance Rating* task, experts are required to choose subjective scores judging video-content relevance, with highest score to '*Most Relevance*' and lowest score to '*Least Relevance*'. In terms of *Grammar Correctness*, without video content, the sentences are graded directly based on the grammatical correctness.

Chapter 6

EXPERIMENTS

6.1 Overview

The aforementioned motivation and technical contributions suggest an empirical validation of the effectiveness using CAVAN. To this end, we conduct experiments and ablation studies on two benchmarks, MSR-VTT (J. Xu et al. 2016) and V2C (Fang et al. 2020) dataset. We evaluate the performance on CAVAN with standard caption evaluation metrics: BLEU@4 (Papineni et al. 2002), METEOR (Lin 2004), ROUGE-L (Banerjee and Lavie 2005), CIDEr (Vedantam, Zitnick, and Parikh 2014), and our newly proposed commonsense-caption entailment score (SE) (see Section 4.5). In order to further validate the effectiveness of CAVAN, we further conduct experiments on VATEX (X. Wang et al. 2019) dataset, one of the largest multi-lingual video-caption dataset, and observe similar performance improvements on it.

6.2 Dataset and Augmentation

MSR-VTT (J. Xu et al. 2016) as a large-scale video description dataset, contains 10,000 video clip with 200,000 clip-sentence pairs in total. Each video is annotated with 20 English descriptions. It covers the most comprehensive categories and diverse visual content. Following the official split, we use 6,513 videos for training, 457 videos for validation and 2,990 videos for testing.

VATEX (X. Wang et al. 2019) as a large-scale multilingual video description dataset, contains 41,250 videos with 825,000 captions in both English and Chinese. It covers diverse human activities and a variety of video content. The dataset is collected by reusing a subset of the Kinetics-600 dataset with additional human annotations. Each video is paired with 10 English and 10 Chinese diverse captions. In CAVAN, we only use English captions for monolingual video captioning task. Following the official split, we use 25,991 videos for training, 3,000 videos for validation and 6,000 videos for testing.

Following the pipeline introduced in Chapter 3, we augment each video caption in MSR-VTT (J. Xu et al. 2016) and VATEX dataset (X. Wang et al. 2019) with 3 types of complementary commonsense descriptions (intention/attribute/effect) retrieved from **ATOMIC** dataset as commonsense anchors for training.

As the queried knowledge from ATOMIC unavoidably comes with noises and incorrect annotations, we further move to use V2C (Fang et al. 2020) dataset with more reliable commonsense knowledge for CAVAN.

V2C (Fang et al. 2020) is a video description dataset adapted from a subset of MSR-VTT (J. Xu et al. 2016). It contains 9,725 videos, 121,651 captions with each surrounded by 3 types of commonsense descriptions, *i.e.*, intention, attribute and effect. We use the standard splits with 6,819 videos for training, and 2,906 videos for testing.

6.3 Implement Details

To obtain the visual representations, we encode videos using multiple pre-trained visual models. For global video representations, we use the I3D (Carreira and Zisserman 2017) network pre-trained on Kinetics dataset (Kay et al. 2017) for motion feature extraction, and the ImageNet pre-trained (Deng et al. 2009) InceptionResNetV2 (Szegedy, Ioffe, and Vanhoucke 2016) for appearance feature of frames. As for object features on entity-level, we utilize a ResNet152 backbone based Faster-RCNN (S. Ren et al. 2015) pretrained on VisualGenome (Krishna et al. 2016). For caption pre-processing, all captions are truncated to a maximum of 24 words, and converted into lower case with punctuation removed. We replace all words with less than 2 word counts into ⟨UNK⟩ token in the vocabulary.

Our BERT based CMS discriminator consists of 12 transformer blocks, 12 attention heads, and is with 768 hidden dimensions. For the entailment pre-training, we choose the event sentence and its corresponded commonsense description as positive pair, and another random commonsense sentence from the ATOMIC as a negative pair. In total, we have 230,624 event-commonsense pairs constructed, with 70% for training, and 30% for testing. Our discriminator achieves 85% accuracy on the testing split.

We use 3 transformer blocks in the visual encoders and decoders, with the hidden dimensions to be 768 and 8 attention heads. We find optimum result by setting the weighting loss term β=0.5. Following the strategy used by Devlin et al. 2018, we train the model using Adam optimizer with the initial learning rate 1e-4, β_1=0.9, β_2=0.999, L2 weight decay 0.01. We use the warm-up strategy for the first 5 epochs, and the learning rate is updated by the cosine scheduler. We set the batch size as 32 and train it for 50 epochs. The reinforcement losses $\mathcal{L}_{cms}$ is not applied until 15 epochs. During testing, we use greedy decoding to generate sentences. Our experiments are implemented on single GTX1080Ti GPU using PyTorch toolbox (Paszke et al. 2019).

6.4 Experimental Results

6.4.1 Results on MSR-VTT Dataset

We show performances of CAVAN in Table 1 and compare them with state-of-the-art methods. To translate content-rich videos into human language, current methods not only extract multi-modal visual features *i.e.*motion, appearance and object features, but also bridge the semantic gap to generate accurate captions by introducing external knowledge. For comparison, we list the feature extractors and the sources of external knowledge in Table 1.

CAVAN outperforms all of the earlier methods on four metrics except ORG-TRL. We summarize the following reasons for this: (1) ORG-TRL carefully designs a relation graph to encode the cross-object interactions later aggregated with global features via a temporal-spatial attention module. Since the main focus lies on the novel commonsense supervision, CAVAN puts less effort on visual encoding. (2) Different video pre-processing and feature extraction methods make it harder to get a completely fair comparison and have a great impact on the results: the baseline models only using appearance and motion features for CAVAN and ORG-TRL achieve 40.9 and 41.9 on BLEU@4 metric respectively. Despite the performance gap for baseline results, CAVAN still gets as competitive improvement as ORG-TRL in comparison with their own baseline models, which validates the effectiveness of our proposed methods.

It is worth noting that CAVAN gets outstanding result on CIDEr metric because semantically aligning with commonsense knowledge encourages accurate and informative details to be involved in the output descriptions, which coincides with the mechanism of CIDEr. In addition, we propose to evaluate the generated captions using the BERT produced semantic score on testing split, which heuristically measures the caption quality and its semantic alignment to commonsense knowledge.

Table 1. Results on MSR-VTT public testing split

Note: We compare CAVAN with previous state-of-the-art models on MSR-VTT (J. Xu et al. 2016) public testing split using various evaluation metrics. “External K.” represents the source of external knowledge. “SE” denotes the average entailment scores of generated captions with their corresponded commonsense knowledge using a generic commonsense discriminator model (BERT) pre-trained on ATOMIC dataset.

Method	**Motion**	**Appearence**	**Object**	**External K.**	**B@4**	**M**	**R**	**C**	**SE**
RecNet(B. Wang et al. 2018)	-	Inception-V4	-	-	39.1	26.6	59.3	42.7	-
PickNet(Chen et al. 2018)		ResNet152	-	-	41.3	27.7	59.8	44.1	-
MARN(Pei et al. 2019)	C3D	ResNet-101	Faster-RCNN	-	40.4	28.1	60.7	47.1	-
OA-BTG(Zhang and Peng 2019)	-	ResNet-200	MASK-RCNN	-	41.4	28.2	-	46.9	-
GRU-EVE(Aafaq et al. 2019)	C3D	InceptionResnetV2	YOLO	-	38.3	28.4	60.7	48.1	-
MGSA(Chen and Jiang 2019)	C3D	InceptionResnetV2	Faster-RCNN	-	42.4	27.6	-	47.5	-
ORG-TRL(Z. Zhang et al. 2020)	C3D	InceptionResnetV2	Faster-RCNN	TBC&WiKi	**43.6**	**28.8**	**62.1**	50.9	-
STG-KD(B. Pan et al. 2020)	I3D	ResNet-100	Faster-RCNN	-	40.5	28.3	60.9	47.1	
Baseline(ours)	I3D	InceptionResNetV2	-	-	40.9	27.6	60.5	47.3	47.8
CAVAN	I3D	InceptionResNetV2	Faster-RCNN	ATOMIC	43.0	**28.8**	61.6	**51.0**	**49.2**

6.4.2 Results on VATEX Dataset

We also conduct experiments on VATEX (X. Wang et al. 2019) dataset which contains 41,250 videos with 825,000 captions in both English and Chinese. It covers diverse human activities and a variety of video content. We show performances of CAVAN using different modules in Tab. 2 and compare them with previous methods. To demonstrate the effectiveness of TSC block and the CMS entailment loss, we list results with and without commonsense knowledge and different features. First, the baseline model applies only appearance and motion features with only single global branch. After we fuse the object-level features using the DFM module (see baseline+DFM), performances of the our model are improved from 30.5 to 31.7 on B@4 metric. This clearly indicates that the object-level features aggregated by DFM module help boosting the results. ORG (Z. Zhang et al. 2020) is a counterpart baseline of (baseline+DFM), which also exploits motion, appearance and object-level features. Then, we combine the commonsense entailment loss to the baselines to compare the effectiveness of $\mathcal{L}_{cms}$. Specifically, when equipped with the $\mathcal{L}_{cms}$, B@4 of baseline is increased from 30.5 to 31.1. Similar trend is also observed on all other metrics, verifying that the use of commonsense anchor brings comprehensive benefits to captioning task. The final performance of CAVAN is shown at the last row where both object-level features and commonsense entailment loss are utilized.It is also worth noting that, TRL and TRL+ORG (Z. Zhang et al. 2020) also make uses of external linguistic knowledge by distilling word probabilities inferred from BERT to the language decoder.

Table 2. Performance on VATEX public testing split

Note: We compare CAVAN with previous state-of-the-art models on VATEX public testing split using various evaluation metrics. “External K.” represents the source of external knowledge. “SE” denotes the average entailment scores of generated captions with their corresponded commonsense knowledge using a generic commonsense discriminator model (BERT) pre-trained on ATOMIC dataset.

Method	**Motion**	**Appear.**	**Entity**	**External K.**	**B@4**	**M**	**R**	**C**	**SE**
Shared Enc X. Wang et al. 2019	I3D	-	-	-	28.9	21.9	47.4	46.8	-
Shared Enc-Dec X. Wang et al. 2019	I3D	-	-	-	28.7	21.9	47.2	45.6	-
ORG Z. Zhang et al. 2020	C3D	Incep.	F-RCNN	-	31.5	21.9	48.7	48.8	-
TRL Z. Zhang et al. 2020	C3D	Incep.	-	TBC&Wiki	31.5	22.1	48.7	49.3	-
TRL+ORG Z. Zhang et al. 2020	C3D	Incep.	F-RCNN	TBC&Wiki	32.1	22.2	48.9	49.7	-
baseline	I3D	Incep.	-	-	30.5	21.6	47.2	47.6	40.6
baseline + DFM	I3D	Incep.	F-RCNN	-	31.7	22.3	48.2	49.6	41.3
baseline + CMS	I3D	Incep.	-	ATOMIC	31.1	21.9	47.8	48.5	41.9
CAVAN	I3D	Incep.	F-RCNN	ATOMIC	**32.3**	**22.4**	**48.4**	**50.4**	**42.5**

6.4.3 Results on V2C Dataset

We report video captioning results on V2C (Fang et al. 2020) dataset in Table 3 using CAVAN. Different with other video captioning models, V2C (Fang et al. 2020) model learns to generate both the ground-truth caption and its complementary commonsense description as a two-stage generating task. Comparing with V2C (Fang et al. 2020) which also uses intention-type knowledge, CAVAN shows a great improvement on B@4 score: 4.0 higher on B@4 metric. The consistent improvements on both MSR-VTT (J. Xu et al. 2016) and V2C (Fang et al. 2020) datasets corroborate that CAVAN is not dataset specific, it is applicable for video captioning task as a generic and novel training schema.

Table 3. Video captioning results on V2C testset using intention-type of knowledge.

Method	**K. Type**	**B@4**	**M**	**R**	**C**	**SE**
V2C Fang et al. 2020	-	34.2	-	-	-	-
V2C Fang et al. 2020	INT.	34.6	-	-	-	-
CAVAN (w/o CMS)	-	38.0	26.6	59.1	57.3	48.3
CAVAN	INT.	**38.6**	**26.8**	**59.4**	**58.7**	**49.6**

Chapter 7

ABLATION STUDY

7.1 Effectiveness of Components

To demonstrate the effectiveness of the proposed DFM module and commonsense entailment loss, we design control experiments. First, baseline model applies only appearance and motion features with only global branch. After we fuse the object-level features with global representations using the DFM module (see baseline+DFM), performances of the our model are dramatically improved, which clearly indicates that the enhanced object-level features aggregated by DFM module help boosting the results. Also, we combine the commonsense entailment loss to the baselines(see baseline+CMS) to compare the effectiveness of $\mathcal{L}_{cms}$. Specifically, when equipped with the $\mathcal{L}_{cms}$, CIDEr of the baseline is obviously increased from 47.3 to 48.4. Similar trend is also observed on all other metrics, verifying that the use of commonsense anchor brings comprehensive benefits to captioning task. We notice that both object-level feature and commonsense knowledge make improvements on SE score. This is because object-level features provide more semantic information and commonsense knowledge put more semantic constrains for the generation. The final performance of CAVAN is shown at the last row of Table 4 where both DFM and commonsense entailment loss are utilized.

Table 4. Effect of each component on MSR-VTT dataset.

Model	**B@4**	**M**	**R**	**C**	**SE**
Baseline	40.9	27.6	60.5	47.3	47.8
Baseline + DFM	42.5	28.6	61.2	49.6	48.2
Baseline + CMS	41.5	27.9	60.9	48.4	48.8
Baseline + DFM + CMS	**43.0**	**28.8**	**61.6**	**51.0**	**49.2**

7.2 Effects of Types of Knowledge

We investigate the benefit of using different types of knowledge annotated in V2C (Fang et al. 2020). The results are presented in Table 5. We can observe that each type of knowledge all can produce positive impact on the caption generation. Among them, using intention-type knowledge gives the best performance for CAVAN. This observation also aligns with the conclusion in (Fang et al. 2020), where the intention-type descriptions lead best generation scores. We analyze that this relates to the annotating bias in ATOMIC dataset, where intentions of human activities are more likely to be annotated correctly.

Table 5. Comparison of performances using different types of commonsense knowledge in CAVAN on V2C testing split.

K. Type	**B@4**	**M**	**R**	**C**	**SE**
-	38.0	26.6	59.1	57.3	48.3
ATT.	38.3	26.7	59.3	57.9	48.9
EFF.	38.4	26.8	59.4	58.3	49.4
INT.	**38.6**	**26.8**	**59.4**	**58.7**	**49.6**

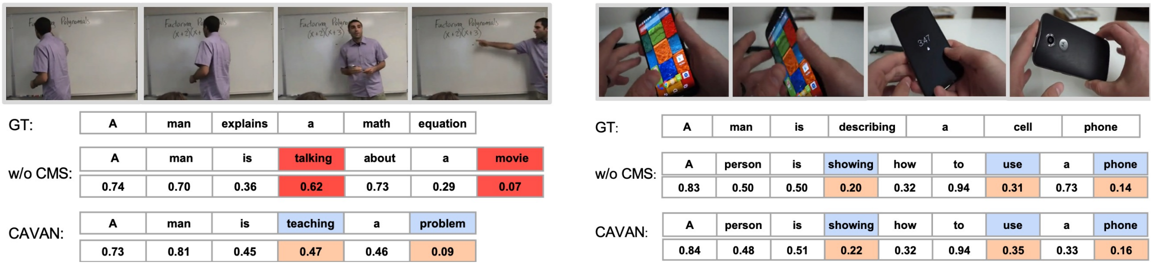
7.3 Human Evaluation

Human evaluation is critical to verify the quality of queried commonsense knowledge and the performance of CAVAN. We conduct human evaluations by crowdsourcing ratings from workers on Amazon Mechanical Turk (AMT). To evaluate the quality of retrieved commonsense-knowledge, the workers are provided with the ground-truth caption and retrieved knowledge and asked to rate whether the retrieved knowledge entails the caption from a scale of 1-5 (the higher the better, 1 denotes irrelevant and 3 means valid.). We get an average score 3.6, 3.3, 3.1 for retrieved intention, attribute and effect respectively on MSR-VTT (J. Xu et al. 2016), this verifies the extracted commonsense-knowledge from ATOMIC is highly relevant to the video content. To validate the performance of CAVAN, given the videos and generations from CAVAN, the AMTurkers are required to watch and rate how well the generated caption describes the video content from 1-5. The skilled workers report that CAVAN achieves 3.65 on average versus 3.50 from the state-of-the-art captioning methods.

7.4 Discussions of CMS in Generations

Figure 5 shows examples of generations from CAVAN comparing with the baseline model without commonsense knowledge. Comparing with model without CMS constrain, ours generation aligns better with the ground-truth annotation semantically. As illustrated in the left example, the model without CMS constrain generates amusing descriptions, whose keywords marked as red are totally misaligned with the ones in ground-truth caption. CAVAN however has the capacity to rectify wrong semantics and hit the correct keywords marked in blue. Moreover, even when both models generate semantically correct descriptions, the probabilities of keywords marked in orange are improved after applying CMS produces (right example in Figure 5). More examples can be found in the appendix.

To further verify the effect of CMS supervision, we show the hit rate of top-20



GT:	A	man	explains	a	math	equation	
w/o CMS:	A	man	is	talking	about	a	movie
	0.74	0.70	0.36	0.62	0.73	0.29	0.07
CAVAN:	A	man	is	teaching	a	problem	
	0.73	0.81	0.45	0.47	0.46	0.09	

GT:	A	man	is	describing	a	cell	phone		
w/o CMS:	A	person	is	showing	how	to	use	a	phone
	0.83	0.50	0.50	0.20	0.32	0.94	0.31	0.73	0.14
CAVAN:	A	person	is	showing	how	to	use	a	phone
	0.84	0.48	0.51	0.22	0.32	0.94	0.35	0.33	0.16

Figure 5. Qualitative Examples

Note: Caption generation examples using CAVAN on V2C dataset with intention-type knowledge (KG). GT represents the ground-truth captions. w/o CMS denotes the model (baseline+DFM) without CMS constrains.

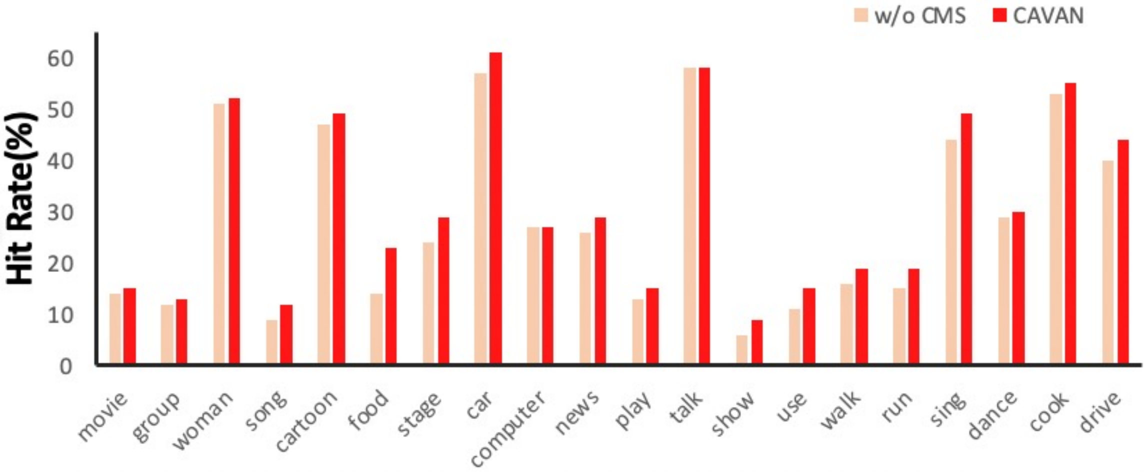


Figure 6. Hit rate

Note: Keywords hit rate on MSR-VTT dataset with or without (intention-type) commonsense knowledge.

most frequent keywords in Figure 6. Concretely, we extract out keywords from each reference sentence based on their TF-IDF score. If the keywords appears in the generated captions, it is then considered to be hit or otherwise missed. We observe from Figure 6 that by applying CMS constrain improves the hit rate of keywords, which is consistent with the outstanding performance on CIDEr metric that also builds on TF-IDF weighting.

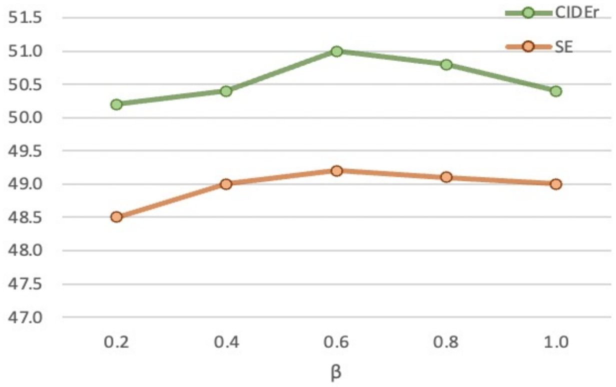


Figure 7. Analysis of different weighting value for $\mathcal{L}_{cms}$ on MSR-VTT dataset.

7.5 Effect of Weighting Parameter β

The performance on CIDEr and SE metric with different values of the weighting parameter β is shown in Fig. 7. The weight *beta* the degree of commonsense constrain: if the value of β is too low, it plays a subtle role in increasing the semantic alignment while if the value of β is too high, the model will be overwhelmed by too much constrain and generate captions deviating from the content of the video itself.

7.6 Vocabulary Coverage of Generated Captions on V2C Dataset

As shown in Fig. 8, the designed CMS supervision leads to a clear improvement in vocabulary coverage. The vocabulary coverage for baseline without CMS knowledge is only 8.39% while for CAVAN with 3 types of CMS knowledge, it is obviously increased to 11.10%, 9.76%, 11.10% respectively. Since the mechanism of proposed commonsense-caption entailment metric is to punish the sentences semantically misaligned with the reference sentence, meaningless words are punished and more informative words

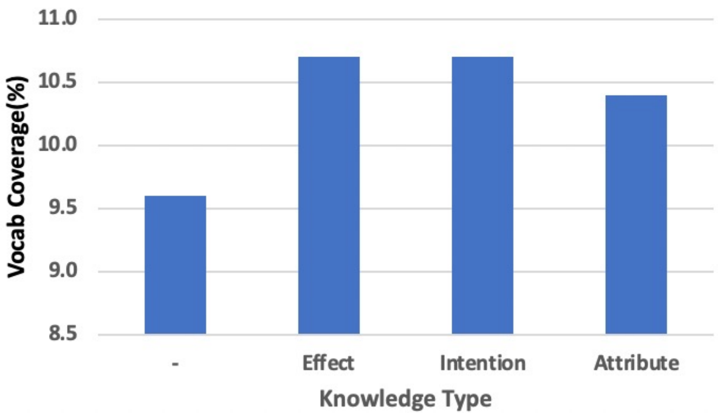


Figure 8. Vocabulary coverage of generated captions on V2C dataset.

are selected to be involved in the generations, which accounts for the increase in vocabulary coverage.

Chapter 8

CONCLUSION AND FUTURE DIRECTION

8.1 Conclusion

We present CAVAN, a novel training schema for video captioning leveraging commonsense knowledge as anchors during model learning. CAVAN efficiently captures higher-order interactions from multi-modal visual features using a carefully designed fusion model, namely DFM, for a better understanding of video content. Moreover, CAVAN is among the first which proposes to measure sentence-level semantics using inferential-knowledge, and incorporate it over an end-to-end training as a supervision signal. We conduct extensive experiments to verify the effectiveness of CAVAN on both MSR-VTT (J. Xu et al. 2016), V2C (Fang et al. 2020) and VATEX (X. Wang et al. 2019) dataset, where CAVAN achieves new state-of-the-art results respectively. The observed success of CAVAN confirms the exciting research avenue by adopting commonsense knowledge for high level cognitive vision tasks, including but not limited to image/video captioning, Visual Question Answering, visual navigation, etc.

8.2 Future Direction

In this work, commonsense knowledge is applied as semantic anchors to form a sentence-level supervision signal that guides the caption generation. We barely make use of the commonsense knowledge in the decoding phase while encoding phase is also an essential part for captioning task. It’s a good direction for us to project commonsense knowledge into an embedding space as semantic features and further explore how to utilize the newly semantic features for generating semantically aligned captions. More extensive experiments will be conducted to figure out the best way to encode the semantic features and incorporate with visual features before decoding.

REFERENCES

Aafaq, Nayyer, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. 2019. "Spatio-Temporal Dynamics and Semantic Attribute Enriched Visual Encoding for Video Captioning." *CoRR* abs/1902.10322. arXiv: 1902.10322. http://arxiv.org/abs/1902.10322.

Aditya, Somak, Yezhou Yang, and Chitta Baral. 2019. "Integrating knowledge and reasoning in image understanding." *arXiv preprint arXiv:1906.09954.*

Anderson, Peter, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. "SPICE: Semantic Propositional Image Caption Evaluation." *CoRR* abs/1607.08822. arXiv: 1607.08822. http://arxiv.org/abs/1607.08822.

Bahdanau, Dzmitry, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. "An Actor-Critic Algorithm for Sequence Prediction." *ArXiv* abs/1607.07086.

Banerjee, Satanjeev, and Alon Lavie. 2005. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments" (January).

Bengio, Samy, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. *Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks.* arXiv: 1506.03099 [cs.LG].

Carreira, João, and Andrew Zisserman. 2017. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset." *CoRR* abs/1705.07750. arXiv: 1705.07750. http://arxiv.org/abs/1705.07750.

Chen, S., and Yu-Gang Jiang. 2019. "Motion Guided Spatial Attention for Video Captioning." In *AAAI.*

Chen, Yangyu, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2018. "Less is more: Picking informative frames for video captioning." In *Proceedings of the European Conference on Computer Vision (ECCV),* 358–373.

Deng, Jia, 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,* 248–255. Ieee.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *CoRR* abs/1810.04805. arXiv: 1810.04805. http://arxiv.org/abs/1810.04805.

Dognin, Pierre, Igor Melnyk, Youssef Mroueh, Jerret Ross, and Tom Sercu. 2019. "Adversarial semantic alignment for improved image captions." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 10463–10471.

Fader, Anthony, Luke Zettlemoyer, and Oren Etzioni. 2014. "Open question answering over curated and extracted knowledge bases." In *KDD '14.*

Fang, Zhiyuan, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. "Video2commonsense: Generating commonsense descriptions to enrich video captioning." *Empirical Methods in Natural Language Processing.*

Fang, Zhiyuan, Shu Kong, Charless Fowlkes, and Yezhou Yang. 2019. "Modularized textual grounding for counterfactual resilience." In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 6378–6388.

Fu, Kun, Junqi Jin, Runpeng Cui, Fei Sha, and Changshui Zhang. 2016. "Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts." *IEEE transactions on pattern analysis and machine intelligence* 39 (12): 2321–2334.

Gao, Jiyang, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. "Tall: Temporal activity localization via language query." In *Proceedings of the IEEE international conference on computer vision,* 5267–5275.

Gao, Junlong, Shiqi Wang, Shanshe Wang, Siwei Ma, and Wen Gao. 2019. "Self-critical n-step Training for Image Captioning." *CoRR* abs/1904.06861. arXiv: 1904.06861. http://arxiv.org/abs/1904.06861.

Gao, Lianli, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. 2017. "Video captioning with attention-based LSTM and semantic consistency." *IEEE Transactions on Multimedia* 19 (9): 2045–2055.

Guo, Longteng, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. 2019. "Aligning linguistic words and visual semantic units for image captioning." In *Proceedings of the 27th ACM International Conference on Multimedia,* 765–773.

Hou, Jingyi, Xinxiao Wu, Yayun Qi, Wentian Zhao, Jiebo Luo, and Yunde Jia. 2019. "Relational Reasoning using Prior Knowledge for Visual Captioning." *CoRR* abs/1906.01290. arXiv: 1906.01290. http://arxiv.org/abs/1906.01290.

Hou, Jingyi, Xinxiao Wu, Xiaoxun Zhang, Yayun Qi, Yunde Jia, and Jiebo Luo. 2020. "Joint Commonsense and Relation Reasoning for Image and Video Captioning."

Proceedings of the AAAI Conference on Artificial Intelligence 34 (April): 10973–10980. https://doi.org/10.1609/aaai.v34i07.6731.

Hu, Yaosi, Zhenzhong Chen, Zhengjun Zha, and Feng Wu. 2019. "Hierarchical Global-Local Temporal Modeling for Video Captioning." *Proceedings of the 27th ACM International Conference on Multimedia.*

Karpathy, Andrej, and Li Fei-Fei. 2015. "Deep visual-semantic alignments for generating image descriptions." In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 3128–3137.

Karpathy, Andrej, and Fei-Fei Li. 2014. "Deep Visual-Semantic Alignments for Generating Image Descriptions." *CoRR* abs/1412.2306. arXiv: 1412.2306. http://arxiv.org/abs/1412.2306.

Kay, Will, 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.*

Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, et al. 2016. "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations." *CoRR* abs/1602.07332. arXiv: 1602.07332. http://arxiv.org/abs/1602.07332.

Kumar, Ankit, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2015. "Ask Me Anything: Dynamic Memory Networks for Natural Language Processing." *CoRR* abs/1506.07285. arXiv: 1506.07285. http://arxiv.org/abs/1506.07285.

Lei, Jie, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. "What is More Likely to Happen Next? Video-and-Language Future Event Prediction." *Empirical Methods in Natural Language Processing.*

Li, Guohao, Hang Su, and Wenwu Zhu. 2017. "Incorporating External Knowledge to Answer Open-Domain Visual Questions with Dynamic Memory Networks." *CoRR* abs/1712.00733. arXiv: 1712.00733. http://arxiv.org/abs/1712.00733.

Lin, Chin-Yew. 2004. "ROUGE: A Package for Automatic Evaluation of Summaries." In *ACL 2004.*

Liu, Chenxi, Junhua Mao, Fei Sha, and Alan Yuille. 2016. "Attention correctness in neural image captioning." *arXiv preprint arXiv:1605.09553.*

Liu, Siqi, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2016. “Optimization of image description metrics using policy gradient methods.” *CoRR* abs/1612.00370. arXiv: 1612.00370. http://arxiv.org/abs/1612.00370.

———. 2017. “Improved image captioning via policy gradient optimization of spider.” In *Proceedings of the IEEE international conference on computer vision,* 873–881.

Lu, Jiasen, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. “Knowing when to look: Adaptive attention via a visual sentinel for image captioning.” In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 375–383.

Pan, Boxiao, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. 2020. “Spatio-Temporal Graph for Video Captioning With Knowledge Distillation.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* June.

Pan, Pingbo, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2015. “Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning.” *CoRR* abs/1511.03476. arXiv: 1511.03476. http://arxiv.org/abs/1511.03476.

Pan, Yingwei, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2015. “Jointly Modeling Embedding and Translation to Bridge Video and Language.” *CoRR* abs/1505.01861. arXiv: 1505.01861. http://arxiv.org/abs/1505.01861.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. “BLEU: A Method for Automatic Evaluation of Machine Translation.” In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics,* 311–318. ACL ’02. Philadelphia, Pennsylvania: Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135.

Pasunuru, Ramakanth, and Mohit Bansal. 2017. “Reinforced video captioning with entailment rewards.” *arXiv preprint arXiv:1708.02300.*

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. “Pytorch: An imperative style, high-performance deep learning library.” In *Advances in neural information processing systems,* 8026–8037.

Pedersoli, Marco, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. “Areas of attention for image captioning.” In *Proceedings of the IEEE international conference on computer vision,* 1242–1250.

Pei, Wenjie, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. 2019. "Memory-Attended Recurrent Network for Video Captioning." *CoRR* abs/1905.03966. arXiv: 1905.03966. http://arxiv.org/abs/1905.03966.

Ranzato, Marc'Aurelio, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. "Sequence level training with recurrent neural networks." *arXiv preprint arXiv:1511.06732.*

Ren, Shaoqing, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *CoRR* abs/1506.01497. arXiv: 1506.01497. http://arxiv.org/abs/1506.01497.

Ren, Zhou, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. "Deep reinforcement learning-based image captioning with embedding reward." In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 290–298.

Rennie, Steven J, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. "Self-critical sequence training for image captioning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 7008–7024.

Sap, Maarten, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. "Atomic: An atlas of machine commonsense for if-then reasoning." In *Proceedings of the AAAI Conference on Artificial Intelligence,* 33:3027–3035.

Shah, Sanket, Anand Mishra, N. Yadati, and P. Talukdar. 2019. "KVQA: Knowledge-Aware Visual Question Answering." In *AAAI.*

Szegedy, Christian, Sergey Ioffe, and Vincent Vanhoucke. 2016. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning." *CoRR* abs/1602.07261. arXiv: 1602.07261. http://arxiv.org/abs/1602.07261.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." In *Advances in neural information processing systems,* 5998–6008.

Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. 2014. "CIDEr: Consensus-based Image Description Evaluation." *CoRR* abs/1411.5726. arXiv: 1411.5726. http://arxiv.org/abs/1411.5726.

Venugopalan, Subhashini, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. *Sequence to Sequence – Video to Text.* arXiv: 1505.00487 [cs.CV].

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. "Show and tell: A neural image caption generator." In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 3156–3164.

Wang, Bairui, Lin Ma, Wei Zhang, and Wei Liu. 2018. "Reconstruction Network for Video Captioning." *CoRR* abs/1803.11438. arXiv: 1803.11438. http://arxiv.org/abs/1803.11438.

Wang, P., Q. Wu, C. Shen, A. Dick, and A. van den Hengel. 2018. "FVQA: Fact-Based Visual Question Answering." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (10): 2413–2427. https://doi.org/10.1109/TPAMI.2017.2754246.

Wang, Xin, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018. "Video captioning via hierarchical reinforcement learning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 4213–4222.

Wang, Xin, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. "VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research." *CoRR* abs/1904.03493. arXiv: 1904.03493. http://arxiv.org/abs/1904.03493.

Wang, Zhe, Zhiyuan Fang, Jun Wang, and Yezhou Yang. 2020. *ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language.* arXiv: 2005.07327 [cs.CV].

Wu, H., J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, and W. Ma. 2019. "Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* 6602–6611. https://doi.org/10.1109/CVPR.2019.00677.

Xu, Jun, Tao Mei, Ting Yao, and Yong Rui. 2016. "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language." IEEE International Conference on Computer Vision / Pattern Recognition (CVPR), June. https://www.microsoft.com/en-us/research/publication/msr-vtt-a-large-video-description-dataset-for-bridging-video-and-language/.

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.* arXiv: 1502.03044 [cs.LG].

Xu, Ning, An-An Liu, Yongkang Wong, Yongdong Zhang, Weizhi Nie, Yuting Su, and Mohan Kankanhalli. 2018. "Dual-stream recurrent neural network for video captioning." *IEEE Transactions on Circuits and Systems for Video Technology* 29 (8): 2482–2493.

Yang, Yezhou, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. "Corpus-guided sentence generation of natural images." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing,* 444–454.

Yang, Ziwei, Yahong Han, and Zheng Wang. 2017. "Catching the temporal regions-of-interest for video captioning." In *Proceedings of the 25th ACM international conference on Multimedia,* 146–153.

Yao, Li, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. *Describing Videos by Exploiting Temporal Structure.* arXiv: 1502.08029 `[stat.ML]`.

You, Q., J. Luo, and Z. Zhang. 2018. "End-to-End Convolutional Semantic Embeddings." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 5735–5744. https://doi.org/10.1109/CVPR.2018.00601.

You, Quanzeng, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. "Image captioning with semantic attention." In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 4651–4659.

Yu, Haonan, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2015. "Video Paragraph Captioning using Hierarchical Recurrent Neural Networks." *CoRR* abs/1510.07712. arXiv: 1510.07712. http://arxiv.org/abs/1510.07712.

Zhang, Junchao, and Yuxin Peng. 2019. "Object-aware aggregation with bidirectional temporal graph for video captioning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 8327–8336.

Zhang, Li, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. 2017. "Actor-critic sequence training for image captioning." *arXiv preprint arXiv:1706.09601.*

Zhang, Ziqi, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. 2020. "Object Relational Graph with Teacher-Recommended Learning for Video Captioning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 13278–13288.

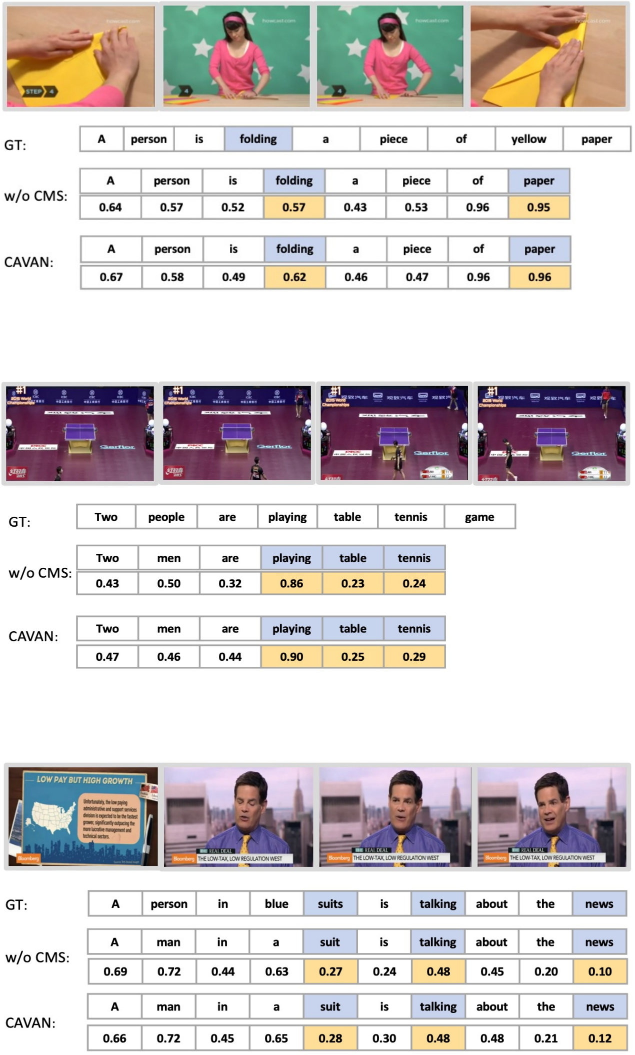
Zhou, Yimin, Yiwei Sun, and Vasant G. Honavar. 2019. “Improving Image Captioning by Leveraging Knowledge Graphs.” *CoRR* abs/1901.08942. arXiv: 1901.08942. http://arxiv.org/abs/1901.08942.

APPENDIX A

QUALITATIVE EXAMPLES

In Appendix, we provide more qualitative results on MSR-VTT (J. Xu et al. 2016) dataset. For MSR-VTT dataset, we compare the captions generated by baseline + DFM and CAVAN in Fig. 9. The keywords in the generated sentence, corresponding with the ones in ground-truth captions, are marked blue with their probabilities marked orange while the predicted keywords marked red are semantically misaligned with the reference sentences.

A.1 Visualization Results on MSR-VTT Dataset

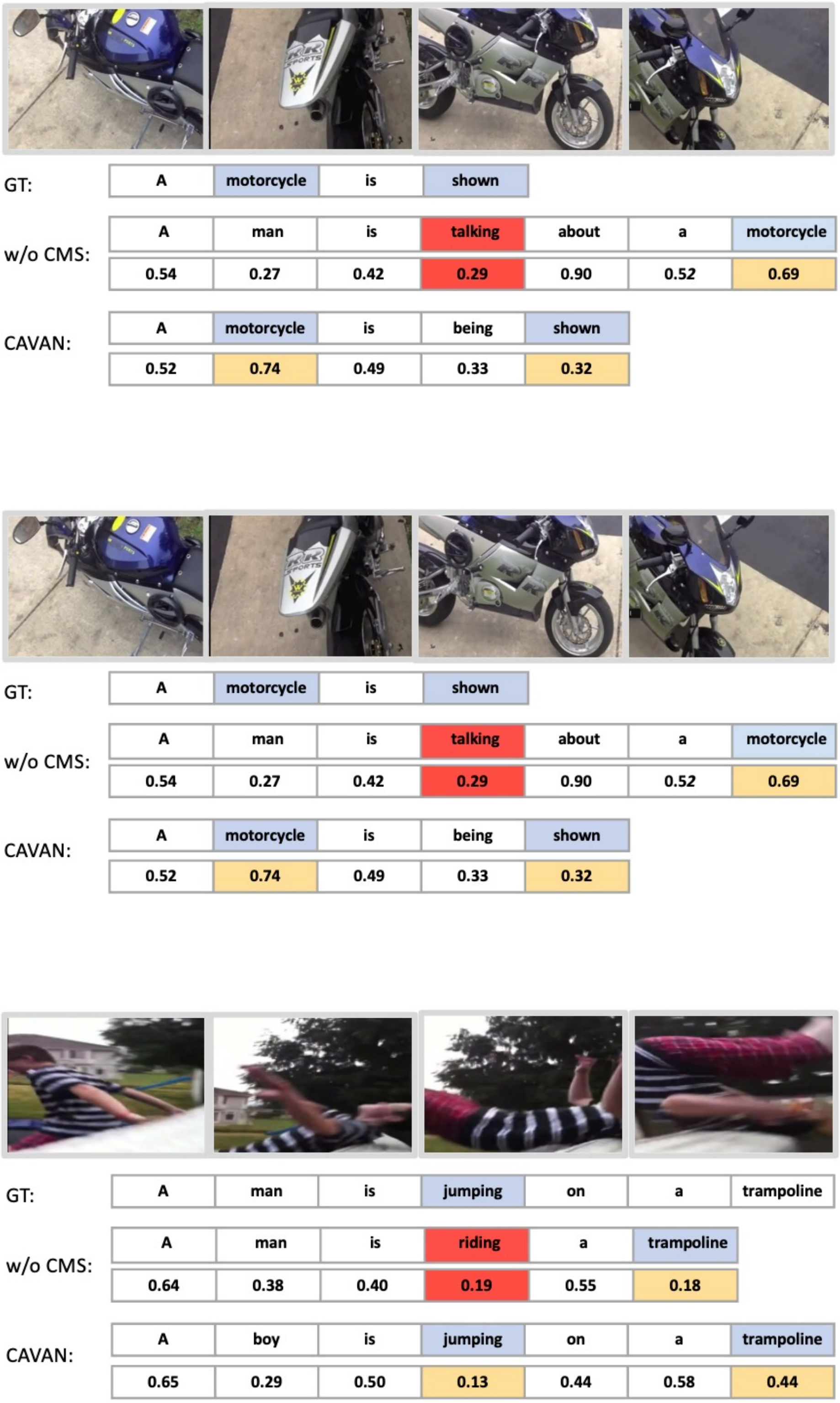


GT:	A	person	is	folding	a	piece	of	yellow	paper
w/o CMS:	A	person	is	folding	a	piece	of	paper	
	0.64	0.57	0.52	0.57	0.43	0.53	0.96	0.95	
CAVAN:	A	person	is	folding	a	piece	of	paper	
	0.67	0.58	0.49	0.62	0.46	0.47	0.96	0.96	

GT:	Two	people	are	playing	table	tennis	game
w/o CMS:	Two	men	are	playing	table	tennis	
	0.43	0.50	0.32	0.86	0.23	0.24	
CAVAN:	Two	men	are	playing	table	tennis	
	0.47	0.46	0.44	0.90	0.25	0.29	

GT:	A	person	in	blue	suits	is	talking	about	the	news
w/o CMS:	A	man	in	a	suit	is	talking	about	the	news
	0.69	0.72	0.44	0.63	0.27	0.24	0.48	0.45	0.20	0.10
CAVAN:	A	man	in	a	suit	is	talking	about	the	news
	0.66	0.72	0.45	0.65	0.28	0.30	0.48	0.48	0.21	0.12

Figure 9. Qualitative examples on MSR-VTT dataset.



GT:	A	motorcycle	is	shown			
w/o CMS:	A	man	is	talking	about	a	motorcycle
	0.54	0.27	0.42	0.29	0.90	0.52	0.69
CAVAN:	A	motorcycle	is	being	shown		
	0.52	0.74	0.49	0.33	0.32		

GT:	A	motorcycle	is	shown			
w/o CMS:	A	man	is	talking	about	a	motorcycle
	0.54	0.27	0.42	0.29	0.90	0.52	0.69
CAVAN:	A	motorcycle	is	being	shown		
	0.52	0.74	0.49	0.33	0.32		

GT:	A	man	is	jumping	on	a	trampoline
w/o CMS:	A	man	is	riding	a	trampoline	
	0.64	0.38	0.40	0.19	0.55	0.18	
CAVAN:	A	boy	is	jumping	on	a	trampoline
	0.65	0.29	0.50	0.13	0.44	0.58	0.44

Figure 10. Qualitative examples on MSR-VTT dataset.