A hierarchical approach to vision-based language generation: from simple sentences to complex natural language

Simion-Vlad Bogolin* **Ioana Croitoru*** **Marius Leordeanu**
Institute of Mathematics of the Romanian Academy
University "Politehnica" of Bucharest
vladbogolin@gmail.com ioana.croi@gmail.com marius.leordeanu@imar.ro

Abstract

Automatically describing videos in natural language is an ambitious problem, which could bridge our understanding of vision and language. We propose a hierarchical approach, by first generating video descriptions as sequences of simple sentences, followed at the next level by a more complex and fluent description in natural language. While the simple sentences describe simple actions in the form of (subject, verb, object), the second-level paragraph descriptions, indirectly using information from the first-level description, presents the visual content in a more compact, coherent and semantically rich manner. To this end, we introduce the first video dataset in the literature that is annotated with captions at two levels of linguistic complexity. We perform extensive tests that demonstrate that our hierarchical linguistic representation, from simple to complex language, allows us to train a two-stage network that is able to generate significantly more complex paragraphs than current one-stage approaches.

1 Introduction

Automatic video to text translation is a very difficult, still unsolved problem, at the intersection of natural language processing, machine learning and computer vision which has been widely researched in the last years (Aafaq et al., 2019). Its applications are varied and far reaching, with impact in virtually all aspects of our lives. One of the main challenges is that for a given visual input, there are infinitely many ways to describe it in natural language. Every human utterance is unique (Chomsky and Lightfoot, 2002; Pinker, 2003), meaning that, while the classic machine learning paradigm assumes, for a given input, a certain "label" from a finite set, language is much more complex and versatile.

Recurrent neural networks (RNN) establish the current basis for research in video to language translation. Using Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) to produce the final text description has become the standard approach. The most common architecture is the sequence-to-sequence model (Venugopalan et al., 2015; Donahue et al., 2015), which usually produces good results. However, it is still unclear what the neural network learns, w.r.t language representation, semantics and diversity. Recent works (Shetty et al., 2017) have clearly pointed out that automated systems have a significantly less diverse vocabulary than humans. It has also been shown that as the sentence becomes longer, the relation to the actual video content drops (Shetty et al., 2017).

Our goal is to study whether a dual, intermediate linguistic description at the level of short sentences, describing simple actions in the form of (subject, verb, object) could help in the generation of more complex and fluent language. Common sense and the challenges encountered in current research, strongly suggest that the transition from vision to language could benefit from a more gradual approach: we should first detect the actors and objects in the scene, then understand and describe their actions in short sentences and only after, put everything together into a larger, coherent story, described in fluent natural language. Note that the end description should not necessarily contain the initial simple sentences.

Thus, we propose a dual-stage video-to-language representation: at the first stage we describe the video as a sequence of events, in the form of simple (subject, verb, object) sentences, which are also well

*Equal contributions

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 28th International Conference on Computational Linguistics, pages 2436–2447
Barcelona, Spain (Online), December 8-13, 2020

localized in time and space. With a slight abuse of terminology, we refer to such simple sentences that describe simple SVO (subject, verb, object) actions as SVO sentences, even though they may sometimes contain more than three words. Then, the second-level description sums up the video content within a coherent paragraph, which describes the visual content in a more elaborate way, by adding to the first level information, causal and semantic relationships between actors and events (without having to repeat the initial SVO sentences). In order to train our models to fully capture the role of the simple sentence descriptions within the more complex video to natural language translation problem, we introduce a novel and relatively large Videos-to-Paragraphs dataset (with appropriate annotations at both linguistic levels), which contains indoor videos from a well-contained ”universe”. The scenes, actors and their activities are in the context of what typically happens in a classroom.

Then, we conduct extensive experiments, which show that our two-stage video to language generation system (Fig. 2) greatly benefits from the intermediate simple sentence representations, which stands between the pure visual interpretation (at the level of single objects or actions as simple ”action labels”) and the more complex interpretations in natural language (at the level of paragraphs).

2 Related work

The task of video captioning can be split in two stages: understanding the video and generating the text. For generating the text, the most used approach is based on RNNs, which uses a video embedding to generate the text (Venugopalan et al., 2015; Donahue et al., 2015). For creating the embedding, the literature varies from average pooling over frame-based features (Venugopalan et al., 2014) to using other various RNNs (Venugopalan et al., 2015; Donahue et al., 2015) and attention models (Yao et al., 2015). Recently, reinforcement learning was also employed to improve pre-trained models by defining policies and rewards specific to image (Chen et al., 2017; Liu et al., 2017; Ranzato et al., 2015) or video (Pasunuru and Bansal, 2017b; Pasunuru and Bansal, 2017a) captioning metrics. Interestingly enough, recent literature has also shown that the actual encoder architecture affects the results less than the high-level, pretrained image, video and audio features used at input (Duta et al., 2018). This suggests that a more complex, stage-wise encoder-decoder pathway could be needed between vision and language, which further justifies our approach.

The idea of generating textual descriptions of simpler events has also been studied (Krishna et al., 2017; Li et al., 2018; Wang et al., 2018a). However, those methods finish at that first level of short sentences, without producing longer and more complex paragraphs, beyond the simple concatenation of the sentences. Moreover, the published datasets do not provide annotations at both levels of simpler sentences and complex paragraphs.

Another recent idea that is related to ours (Wang et al., 2018b) is to ”divide and conquer” the video content, by first breaking the long caption into many small segments, then employ a sequence to sequence model to generate the final caption. They introduce a transformed version of the Charades dataset (Sigurdsson et al., 2016). Actors were given a script and were asked to record a video that follows it as closely as possible. Different than us, they identify actions by class labels (which is a standard approach), not describe such actions by short sentences (which is an intermediate vision-to-language translation approach). Moreover, their final captions are significantly shorter than ours (ours: 40.03 vs theirs: 24.13 words on average per caption). Thus, they focus on the intermediate prediction of ”action labels”, combined with a reinforcement learning approach. We generate more sophisticated textual descriptions by going through an initial stage of descriptions in simple sentences and, thus, take a longer step towards bridging the gap between vision and natural language.

The lack of other datasets in the literature that provide annotations at both levels of simple sentences and more complex language, limits, for now, the experimental comparisons with methods such as (Wang et al., 2018b). However, by introducing our new Videos-to-Paragraphs dataset, with full annotations at both levels, we are able to compare to published state of the art methods by retraining them on our dataset.

Our approach is to the best of our knowledge novel. However, there are other methods that proposed to have an intermediate representation (Yu et al., 2016), but, different than our first-level SVO represen-

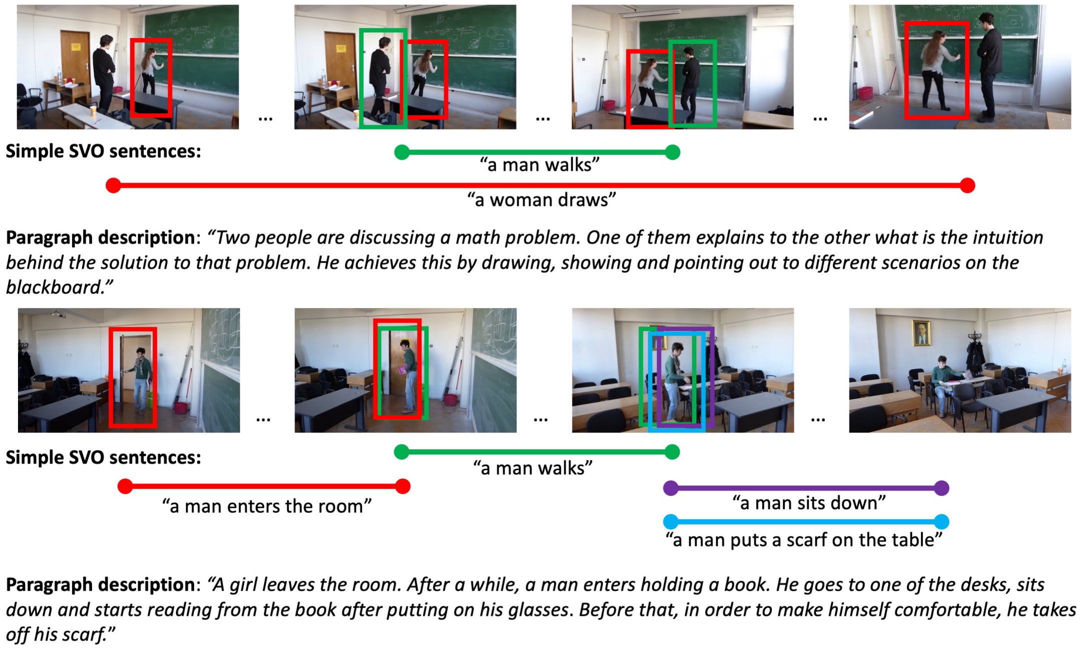

...
...
...
Simple SVO sentences:
"a man walks"
"a woman draws"
Paragraph description: "Two people are discussing a math problem. One of them explains to the other what is the intuition behind the solution to that problem. He achieves this by drawing, showing and pointing out to different scenarios on the blackboard."
...
...
...
Simple SVO sentences:
"a man walks"
"a man enters the room"
"a man sits down"
"a man puts a scarf on the table"
Paragraph description: "A girl leaves the room. After a while, a man enters holding a book. He goes to one of the desks, sits down and starts reading from the book after putting on his glasses. Before that, in order to make himself comfortable, he takes off his scarf."

Figure 1: Examples from Videos-to-Paragraphs Dataset. We present annotations for two different videos. For each we present a few SVO (subject, verb, object) events along with: their start frame, end frame, corresponding space-time bounding box that contains the event, and its simple sentence description. At the bottom we present the paragraph-level video description. The paragraphs contain longer and more complex sentences than the SVOs.

tation, that intermediate representation is a hidden RNN state that does not have an explicit, explainable meaning. The two levels we propose are valid linguistic representations at different levels of spatiotemporal context. One is limited to local SVOs, the other is a compact paragraph at the whole video level. Then, other published works predict only simple semantic representations of video content (similar to our proposed SVOs) (Xu et al., 2015; Zanfir et al., 2016).

3 Our Videos-to-Paragraphs Dataset

Below we describe our Videos-to-Paragraphs dataset and provide details regarding annotations and various statistics of the dataset.

3.1 Dataset description

The dataset consists of 510 videos captured in an indoor school environment. There are videos from various shots from two different classrooms and some shots on the hallway in between. Each video clip has about 30 seconds and the videos were filmed with two different cameras: a fixed one and a moving one, focusing on the central actors. Thus, we aim to better understand how learning video-to-text generalizes depending on one camera setting or the other. The protagonists perform various activities, which usually involve interacting with different objects. All activities are centered around the people and their interaction with other people and objects. Examples of such simple "atomic" actions are: enter/leave a room, drink water, sit up/sit down, pick up/put down objects (e.g. pen, laptop, notebook, mobile phone etc), talk to/hug/shake hands with/present to another person, open the door/the window, and others, that usually take place in a classroom environment. We designed scenarios that are as realistic as possible, such that a given video could contain many other objects and actions taking place in the background. These basic, atomic actions are the ones described in small (subject, verb, object) sentences. The main feature that distinguishes our dataset from others in the literature is that we limited all scenes and scenarios to a self-contained, well-covered "universe" of actors, actions and plausible event, in order to better capture and study the relationship between vision and language within a specific context. Without this compact, self-contained set of videos, the gap between vision to language is simply too large to bridge, and the task is more prone to overfitting - an observation which often can be made in the current literature.

3.2 Annotations

For each video we have several two-layered text annotations from multiple annotators. Actors were asked to behave naturally, as if in a real-life classroom scenario, without a given script, by acting freely. A list of 39 possible actions and 19 different object categories, were given only as examples at the beginning. Annotators were only asked to limit the complexity of the short sentences (SVO) as much as possible, without any other strict rules.

Simple SVO sentences: they describe simple, atomic actions that are well located in space (with a bounding box annotation for each frame) and time (start and end frame for each action). The simple SVO sentences at level one could sometimes contain more than three (subject, verb, object) words necessary to describe a simple action. Also, each action bounding box has annotated ”links” to other bounding boxes, if they are semantically related.

Higher, video level paragraph description: a paragraph consisting of several sentences (3.6 sentences on average), which explains concisely and fluently what happens in the video. The paragraphs describe the video content at a holistic level, by incorporating the information from simple sentences within a larger context. They provide information that is not included in the simpler sentences, regarding how the atomic events relate to each other in time and space. Being closer to natural human language, they often provide higher level interpretations and explanations regarding why the people in the video perform certain actions. Note that the annotators were left free to write the paragraphs as they considered appropriate. As mentioned, the annotation at the first level of atomic actions is more strict, being limited to simple SVO sentence.

We provide some annotation examples in Fig. 1. For the first example in Fig. 1 the full list of SVOs is: 1 - *”A woman draws”*, 2 - *”A person shows another person something on the blackboard”*, 3 - *”Two persons are talking”*, 4 - *”A man walks”*. We also annotate links between SVOs that are considered to be semantically related, providing a higher level of dependency, not between objects but between events/actions. In the given example, we have a link between SVO 2 and SVO 1, which constrains SVO 2 to be conditioned on SVO 1. These events connecting links enable the construction of a hierarchy or a graph of events.

The annotators consist of 20 graduate students in artificial intelligence, who volunteered for this task. The videos were randomly assigned to students so that each video has at least 2 annotations. For the annotation process, we developed a specific, modified version of a public annotation tool (Shen, 2016).We modified the tool to facilitate drawing and defining SVOs, adding links and the overall paragraph description.

3.3 Videos-to-Paragraphs Dataset statistics

Our dataset consists of 510 videos (245 filmed with a fixed camera and 265 with a mobile camera that follows the scene), each having about 30 seconds. From the 510 videos, we use 438 for training, 20 for validation and 52 for testing. We have collected 1048 annotations, so that each video has at least 2 annotations. In these 1048 annotations we have 9036 annotated SVOs. Thus we have, on average: 8.62 SVOs per annotation, with 5.24 words per SVO and 4.13 sec. covered by a SVO. A SVO covers on average 14% of the video, while about 81.4% from a video is covered by all the annotated SVOs for that particular video. Many SVOs (68% of them) are overlapping with others temporally, as different actions may happen simultaneously, sometimes even done by the same person (eg: talking to someone, sitting down and putting down something). A paragraph-level description has on average 3.66 sentences (longer than the initial-level SVO sentences) and 40.03 words.

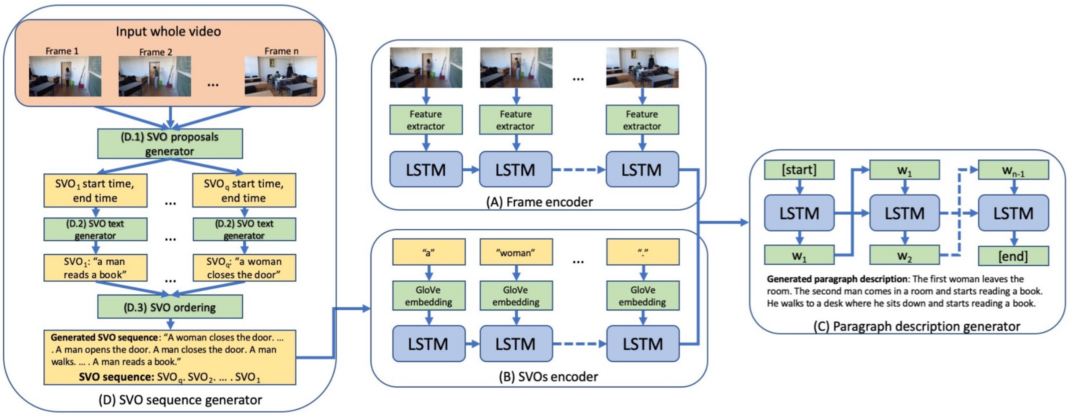

Input whole video
Frame 1
Frame 2
...
Frame n
(D.1) SVO proposals generator
SVO_1 start time, end time
...
SVO_q start time, end time
(D.2) SVO text generator
...
(D.2) SVO text generator
SVO_1: "a man reads a book"
...
SVO_q: "a woman closes the door"
(D.3) SVO ordering
Generated SVO sequence: "A woman closes the door. ... . A man opens the door. A man closes the door. A man walks. ... . A man reads a book."
SVO sequence: SVO_q. SVO_2. ... . SVO_1
(D) SVO sequence generator
Feature extractor
Feature extractor
...
Feature extractor
LSTM
LSTM
LSTM
(A) Frame encoder
"a"
"woman"
...
"."
GloVe embedding
GloVe embedding
GloVe embedding
LSTM
LSTM
LSTM
(B) SVOs encoder
[start]
w_1
w_{n-1}
LSTM
LSTM
LSTM
w_1
w_2
[end]
Generated paragraph description: The first woman leaves the room. The second man comes in a room and starts reading a book. He walks to a desk where he sits down and starts reading a book.
(C) Paragraph description generator

Figure 2: Our system consists of several modules that work together to form two encoding pathways, one that processes the video frames (module *(A) Frame encoder*) and the other that processes the SVOs (module *(B) SVOs encoder* and module *(D) SVOs sequence generator*). At the end, a RNN generator (module *(C) Paragraph description generator*) outputs the final description. For generating a SVO sequence, we first determine the start and end times of potential SVOs (module *(D.1) SVO proposal generator*). Then, for each SVO time window we generate a textual description (module *(D.2) SVO text generator*), then sort the SVOs in temporal order (module *(D.3) SVO ordering*) and obtain the SVO sequence which is fed to the *(B) SVO encoder* module. Please see the full method's pseudo-code in the supplementary material.

4 Proposed approach

Our key idea is to generate natural language descriptions from video input in a hierarchical, stage-wise manner. Starting from describing atomic actions in simple sentences, then we generate a more compact and coherent final text, in more complex language. Intuitively, we relate this approach to visual object recognition, in which we recognize whole objects by first detecting their parts and sub-parts. We view language generation from a similar, stage-wise perspective. In fact, by describing the world at many different layers of semantic abstraction in space and time, we expect language to be more complex than the visual recognition of physical objects. Thus, we propose a hierarchical representation of the vision-language world, in which objects first do simple things described in simple sentences, before relating these simple events and creating larger stories. These stories contain more complex connections, descriptions and explanations, requiring more complex language. We present an overview of our system in Fig. 2. There are two pathways: one that receives the whole video, processes it and creates a video embedding and the other that processes SVOs and creates a SVOs embedding. In the end, features from both pathways are combined and fed to a "Paragraph description generator" module. Next we define each component.

4.1 Frame encoder

The frame encoder (module A in Fig. 2), a standard LSTM (Venugopalan et al., 2015) receives frame-level Inception v4 (Szegedy et al., 2017) features pre-trained on ImageNet (Deng et al., 2009), for a given input video. It produces a video embedding that is then fed into the description generator.

4.2 SVOs encoder

The SVOs encoder (module B in Fig. 2) receives an input paragraph formed by concatenating all the SVOs descriptions for a given video and outputs an embedding vector, for the whole paragraph (see Sec. 4.4). The SVOs encoder is a LSTM that receives one word at a time (Fig 2). Note that we add to the vocabulary the end sentences delimiter ".", such that the SVOs encoder can process an entire paragraph. Before being fed to the LSTM, each word is transformed into a feature vector using GloVe (Pennington et al., 2014). In the end, we obtain one embedding vector that represents the whole sequence of SVOs.

4.3 Paragraph description generator

The description generator (module C in Fig. 2) is a classical textual decoder (Venugopalan et al., 2015), initialized with the concatenation of the SVOs encoding and the frame encoding. The goal is to have a sequence to sequence model that also receives SVOs features (obtained from the SVOs encoder), as well as the frame features.

4.4 SVO sequence generator

The purpose of this module (module D in Fig. 2) is to generate multiple SVOs descriptions, at different times, from a video and eventually output a sequence of SVOs in termporal order. The SVO sequence is encoded and then fed to the final paragraph description generator module.

4.4.1 SVO proposals generator

For the SVO proposals generator (module D.1 in Fig. 2) we first test a simple idea: generate a fixed number of SVOs per video, uniformly distributed over time. So, we split each video in 8 sub-videos representing the SVOs, then generate a text for each (Sec. 4.4.2). The second idea is to train a confidence network that, given a sub-video, outputs a score which indicates whether or not an atomic event (SVO) occurs on those frames. We use an LSTM network to process the frames (similar to the ”Frames encoder” module) and add at the end a fully connected layer that outputs the confidence score. We train this network to predict the time-wise intersection over union between a sub-video and ground truth SVOs from the training videos. Finally, we generate many SVOs proposals, compute the score given by the confidence network, then apply non-maxima suppression to obtain the final SVO sequence ordered in time.

4.4.2 SVO text generator

This module (D.2 in Fig. 2) receives as input a video (or sub-video) and outputs a single short sentence (a SVO). We test two ideas: 1) receive as input a full-size sub-video (consider only start and end times, without bounding box) and 2) take an input sub-video (also limited to the given bounding box).

4.4.3 SVO ordering

This module (D.3, Fig. 2) aims to generate a sequence of SVOs that are correlated with the way the events occurs in the video. In order to obtain the final description, we look at the links between the SVOs ensuring that if a SVO depends on another, then it will appear before it, in the final sequence. The second option is to sort the SVOs by their starting time. Please see the pseudo-code of our full algorithm in the supplementary material.

5 Experimental setup

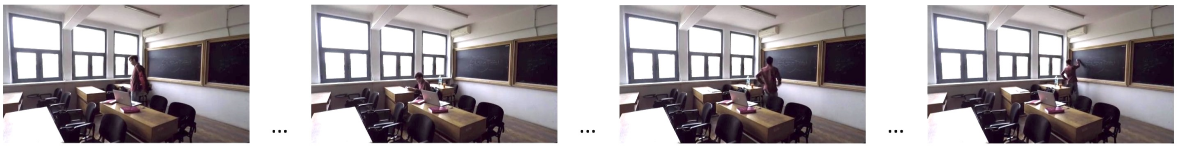
We validate experimentally the benefit of the two-stage text representation and generation approach for video to language translation. Thus, we design experiments that highlight the relevance of key elements of our method and use the standard language metrics in the literature: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015) and METEOR (Banerjee and Lavie, 2005).

5.1 Implementation details

All models were trained in PyTorch (Paszke et al., 2019) using the Adam optimizer (Kingma and Ba, 2014) for about 1000 epochs using a batch size of 64 on a Nvidia GTX1080Ti. The base learning rate is 4e-4 and has a decay rate of 0.8 at each 200 epochs. The SVO text generator is trained separately using ground truth SVO annotations and then frozen in all other experiments. All the other components are trained end-to-end.

5.2 Comparison to existing methods

Even though we introduce a new annotation scheme, it is crucial to see how we stand against other published methods. So, we compare against a strong baseline (S2VT system (Venugopalan et al., 2015)) and a recent state-of-the-art method (Duta et al., 2018) (see Tab. 1). Please note that for a fair comparison,



GT: *"A man sits down on a chair in order to read a book. After a while he looks at the blackboard, sits up and takes a chalk. He is probably going to write something on the blackboard."*

Seq-GT-SVO: *"A man takes off his jacket. A man puts his jacket on the chair. A man sits down. A man reads a book. A man stands up. A man writes on the blackboard."*

PG-GT-SVO: *"A man takes off his jacket and puts the jacket on the table. Then he puts down the book and starts reading and browsing it."*

(Duta et al. 2018): *"A man closes the window and leaves the room then closes the window and sits down."*

PG-NMS-Pred-SVO: *"A man takes off his backpack and puts it on the desk while sitting down and starts reading. After a while he sits up and starts writing something on the blackboard."*

Figure 3: Qualitative video to language generation results. Note the quality differences between the descriptions at the level of SVO sequences and those at the level of paragraphs. The generated paragraphs are more concise and coherent. They tell more with less words, using longer sentences and overall more complex language. Also note that our full model (PG-NMS-Pred-SVO) generates longer and more diverse descriptions than the state of the art method for vision-to-language translation of (Duta et al., 2018), which is trained on the same videos, having the same target ground truth, with no restriction regarding the length or structure of the generated text. Seq-GT-SVO are the ground truth annotations for the SVOs concatenated without using any model. PG-GT-SVO receives as input the ground truth SVOs and generates the paragraph, while our final model PG-NMS-Pred-SVO does not have access to the ground truth SVOs during inference and uses predicted SVOs. Please see additional results in the supplementary material.

we compare against their best single model which uses a sequence-to-sequence approach and an attention mechanism.

For a fair comparison to (Venugopalan et al., 2015), we follow the same base architecture and only add the SVOs enconder (module B) as a second pathway. As seen in Fig. 3, our model generates longer and more complex paragraphs. So, our two level annotation scheme, brings a clear improvement over the baseline model in all cases where the SVO encoding is used. The fact that a simple SVO generation scheme, with 8 SVOs per video, uniformly sampled in time (PG-SW-Pred-SVO model in Tab. 1), brings a substantial improvement of 6% over the baseline (6% is very large for the video-to-language literature), is strong evidence, that the intermediate level of SVO description is a solid contribution. Then, our full PG-NMS-Pred-SVO model, which produces superior SVO sequences through a more sophisticated confidence network with non-maxima suppression (Sec. 4.4.1), further improves to more than 16% on

Method	B@1	B@2	B@3	B@4	M	R	C
Baseline	49.0	29.2	17.4	10.3	15.2	30.0	16.6
(Duta et al., 2018)	48.2	28.4	16.7	9.9	15.8	30.0	18.3
PG-SW-Pred-SVO	**51.2**	**33.6**	**22.2**	14.4	17.3	33.8	22.9
PG-NMS-Pred-SVO	49.3	32.5	21.7	**14.5**	**19.7**	**37.3**	**26.2**

Table 1: Comparison with state of the art methods, using several evaluation metrics (BLEU@N (B@N), ROUGE (R), METEOR (M) and CIDEr (C)). The compared methods were trained from scratch on the training set of our database. At test time, all methods have access only to the input video. Our approach leads by a significant margin especially w.r.t the most recent metrics. Both PG-SW-Pred-SVO and PG-NMS-Pred-SVO take advantage of the two level scheme annotation, the difference between them is that PG-SW-Pred-SVO uses a more naive way of detecting SVO regions - sliding window, while PG-NMS-Pred-SVO uses a more sophisticated confidence net in conjunction with nonmaxima suppersion - module D.1

Method	Time GT info	Space GT info	**B@1**	**B@2**	**B@3**	**B@4**	**M**	**R**	**C**
Inter-human	-	-	44.0	29.6	19.9	13.1	19.0	34.4	32.6
Baseline	-	-	49.0	29.2	17.4	10.3	15.2	30.0	16.6
Seq-GT-SVO	Y	Y	46.7	35.3	26.6	**19.5**	**27.6**	**44.6**	43.7
PG-GT-SVO	Y	Y	**59.3**	**40.9**	**28.0**	**19.5**	22.6	41.8	**45.9**
Seq-Pred-SVO	Y	Y	39.9	25.3	15.4	9.1	**19.2**	33.5	22.9
PG-Pred-SVO	Y	Y	**57.0**	**38.5**	**25.5**	**17.2**	**19.2**	**38.5**	**31.3**
Seq-Pred-SVO	Y	N	38.9	25.1	15.6	9.3	**18.7**	33.6	23.9
PG-Pred-SVO	Y	N	**55.0**	**36.3**	**24.3**	**16.5**	**18.7**	**38.1**	**35.8**
Seq-SW-Pred-SVO	N	N	36.8	22.4	13.0	7.3	16.2	29.0	10.4
PG-SW-Pred-SVO	N	N	**51.2**	**33.6**	**22.2**	**14.4**	**17.3**	**33.8**	**22.9**
Seq-NMS-Pred-SVO	N	N	32.1	20.3	12.6	7.6	17.7	29.1	3.3
PG-NMS-Pred-SVO	N	N	**49.3**	**32.5**	**21.7**	**14.5**	**19.7**	**37.3**	**26.2**

Table 2: Evaluating the effect of using the second-level paragraph generation (”PG” models) vs the first-level of SVO sequences (”Seq” models), using common evaluation metrics (BLEU@N (B@N), ROUGE (R), METEOR (M) and CIDEr (C)). The ”Baseline” model uses no SVO modules. For the ”Seq” cases we compared the generated SVO sequence (using only module D in Fig. 2) with the ground truth paragraph description. For the ”PG” cases, we compared the generated paragraph (using the full system, with all modules A, B, C and D in Fig. 2) to the ground truth. Note that generating the paragraph at the second level greatly helps vs. using the more simple SVO sequence. The ”SW” case uses a sliding window approach to detect SVO regions, while ”NMS” refers to the use of a confidence net in conjunction with nonmaxima suppression (module D.1) for SVO detection. In the Time/Space GT columns we show whether true temporal or spatial SVO location is given (Y) or not (N). In the ”Pred” cases, the SVOs were generated automatically vs. the GT case when true SVOs were given.

CIDEr score over (Venugopalan et al., 2015) and 10% over the state-of-the-art (Duta et al., 2018). This significant jump in performance is strong indication that an intermediate level annotation (such as we propose, in the form of SVOs) is needed to eventually reach the superior levels of complex, fluent, natural language.

5.3 The importance of paragraph description

One can argue that the sequence of SVOs, ordered in time, is a sufficient full description of video content. So, we compare our two-stage model output against the ”sequence” obtained by concatenating together all the SVOs (generated or ground truth) in Tab. 2. There is a major difference between the SVO sequence paragraph and the description. This is not surprising as natural language is diverse. A good description of a video is not limited to just enumerating, in simplistic sentences, all the atomic events that occur in a video. The results clearly show that paragraph-level descriptions are not trivial concatenations of lower-level SVOs. They also set an upper-bound for the dense captioning methods: even if a dense captioning method perfectly predicts the SVOs and concatenates them into a final description, it still cannot capture the final description in natural language. This once again proves that the whole is greater and actually different than the simple sum of its parts (Sternberg and Sternberg, 2016).

This fact is even better seen when comparing the system using ground truth SVOs (PG-GT-SVO) to the ground truth SVO sequence, directly (Seq-GT-SVO). As shown in Tab. 2, even though the results are close, for the CIDEr metric, which is known to better correlate with the human judgment, PG-GT-SVO, which produces second-level paragraphs, is superior by 2% to the ground truth SVO sequence. This observation shows that by processing the SVOs and using them as features, we can obtain a better description than by simply copying the SVOs. It again stresses out that learning first to second level language transformations improves natural language generation. Additionally, in both cases the improvement compared to the baseline is over 25%, meaning that our proposed scheme is useful.

Method	Feature type	B@1	B@2	B@3	B@4	M	R	C
Baseline	(Hara et al., 2018)	49.8	31.5	20.2	13.2	16.2	30.6	19.2
PG-GT-SVO	(Hara et al., 2018)	60.4	42.0	29.5	20.6	21.8	42.3	47.3
PG-NMS-Pred-SVO	(Hara et al., 2018)	53.4	34.8	23.3	15.6	18.7	35.9	28.7
Baseline	(Xie et al., 2018)	51.3	30.5	18.7	12.4	15.8	31.3	24.8
PG-GT-SVO	(Xie et al., 2018)	58.9	42.5	31.3	22.7	23.5	43.3	48.9
PG-NMS-Pred-SVO	(Xie et al., 2018)	52.9	35.2	23.0	14.6	18.6	35.9	27.5

Table 3: Influence of different additional features. We report the commonly used metrics: B@N stands for BLEU@N while R, M and C represent the ROUGE, METEOR and CIDEr metrics. As it can be seen, even when using more powerful features, there is a clear improvement gain when using our two level annotation scheme.

Method	Camera type	B@1	B@2	B@3	B@4	M	R	C
Baseline	mobile	44.5	26.6	15.8	8.9	15	29.9	15.5
Baseline	fixed	**55.2**	**36.1**	23.1	14.4	16.6	35.7	19.5
(Duta et al., 2018)	mobile	43.2	26.0	15.6	10.1	15.1	27.5	18.1
(Duta et al., 2018)	fixed	47.4	29.8	18.3	11.2	16.9	32.1	21.2
PG-NMS-Pred-SVO	mobile	48.6	29.7	18.6	12	17	34	23.7
PG-NMS-Pred-SVO	fixed	48.3	34.2	**23.4**	**15.1**	**21.3**	**37.7**	**26**

Table 4: Comparison between fixed and mobile camera. We report the commonly used metrics: B@N stands for BLEU@N while R, M and C represent the ROUGE, METEOR and CIDEr metrics. We re-trained all models from scratch to include only videos filmed with mobile or fixed camera. The fixed camera videos are easier to describe in natural language, on average and the results are superior to the mobile camera case.

5.4 The importance of features

Lately, it has been shown that the features are highly important for the task of captioning (e.g. 3D convolutional features (Sun et al., 2019b; Sun et al., 2019a)). We show tests with using additionl features in Tab. 3. We add two different feature sets, commonly used in the field (Hara et al., 2018; Xie et al., 2018). Both methods extract features from the action recognition task using 3D Convolutional Neural Networks. As see in Tab. 3 our method brings a clear gain even though the extra features are much stronger than the initial ones. This is yet another indication that the proposed two level annotation scheme brings a solid additional value.

5.5 Comparing humans to other humans

By studying the inter-human agreement between annotations, we should get a better understanding of the system's performance (Tab. 2). As expected, there is a large variety in how different humans describe the same video. The relatively low results for human-to-human evaluation, reveal once again the innate difficulty of the task and the limitation of the current evaluation metrics. It indicates the strong need for a better knowledge representation at the intersection of vision and language, which could help both in evaluation, as well as in language understanding and generation at a deep semantic level. This is a future direction that is worth pursuing and which could take advantage of our proposed hierarchical structure for language generation, from simple events to complex paragraphs.

6 The effect of camera movement

The introduced Videos-to-Paragraphs dataset contains videos filmed from two viewpoints: a mobile one that follows the action and a fixed one that captures the whole scene. Please note that for the fixed camera, the whole scene is captured and there is no action that occurs outside of the camera viewpoint. As seen in Tab. 4 there is a clear difference between the fixed camera and the mobile one, the fixed-camera scenario being clearly superior. This shows that a larger context, as captured by the fixed camera scheme, helps the model. The power of context is also emphasized in experiments when we restricted the SVOs to the annotated bounding box, obtaining worse performance than when the whole frame is used.

7 Conclusions and Future Work

In this paper we introduced a novel, hierarchical approach to language generation from videos. It starts with generating simple sentences ordered in time, followed by the generation, at a second stage, of a longer and more complex description, which semantically relates the simpler events in space and time using more coherent, natural language. This is to our best knowledge, the first approach of this kind in the literature, with two explainable stages for video to language generation. Additionally, we introduce a novel dataset, Videos-to-Paragraphs, with full annotations at both levels of textual descriptions, again, the first of its kind in the literature. The dataset is instrumental in comparing our work to the standard, direct video-to-language published approaches, which are not able to generate longer coherent paragraphs. Our extensive experiments strongly suggest that the idea of generating explicit language in several phases, going from simpler sentences, to longer paragraphs and then to complex stories, could offer an interesting direction towards strong common vision and language representations.

Acknowledgements This work was supported through UEFISCDI, in part by EEA Grant EEA-RO-2018-0496 and project PN-III-P1-1.1-TE-2016-2182. The authors express their deepest appreciation and gratitude towards the graduate students volunteers, from Univerisity Politehnica of Bucharest, University of Bucharest and the Institute of Mathematics of the Romanian Academy, who helped with the full acquisition and annotation of the Videos-to-Paragraphs Dataset.

References

Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 521–530.

Noam Chomsky and David W Lightfoot. 2002. *Syntactic structures*. Walter de Gruyter.

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.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

Iulia Duta, Andrei Liviu Nicolicioiu, Simion-Vlad Bogolin, and Marius Leordeanu. 2018. Mining for meaning: from vision to language through multiple networks consensus. *British Machine Vision Conference 2018*, page 275.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 706–715.

Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7492–7500.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 873–881.

Kishore Papineni, 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*, pages 311–318. Association for Computational Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2017a. Multi-task video captioning with video and entailment generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1273–1283, Vancouver, Canada, July. Association for Computational Linguistics.

Ramakanth Pasunuru and Mohit Bansal. 2017b. Reinforced video captioning with entailment rewards. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 979–985, Copenhagen, Denmark, September. Association for Computational Linguistics.

Adam Paszke, 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*, pages 8024–8035.

Jeffrey Pennington, Richard Socher, and Christopher 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.

Steven Pinker. 2003. *The language instinct: How the mind creates language*. Penguin UK.

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

Anting Shen. 2016. Beaverdam: Video annotation tool for computer vision training labels. *EECS Department, University of California, Berkeley, Master Thesis*.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer.

Robert J Sternberg and Karin Sternberg. 2016. *Cognitive psychology*. Nelson Education.

Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. 2019a. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019b. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7464–7473.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4534–4542.

Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. 2018a. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7190–7198.

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

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321.

Ran Xu, Caiming Xiong, Wei Chen, and Jason J Corso. 2015. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4507–4515.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.

Mihai Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. 2016. Spatio-temporal attention models for grounded video captioning. In *asian conference on computer vision*, pages 104–119. Springer.