
**ROMANIAN ACADEMY**

**School of Advanced Studies of the Romanian Academy**

**"Simion Stoilow" Institute of Mathematics**

# PhD Thesis Summary

Describing complex activities in natural language through multiple network consensus and hierarchical representations

**PhD advisor:**

Prof. Dr. Marius Leordeanu

**PhD Student:**

Simion-Vlad Bogolin

**2022**

# Contents

# Abstract

In this work, we aim to describe complex activities in natural language. We argue that this is the most appropriate way, since natural language is so versatile and, in this way, you are not limited to a fixed amount of labels. Moreover, automatically describing actions depicted in videos in natural language is an ambitious problem, which could bridge our understanding of vision and language. We start by introducing a novel, consensus based approach that leverages the power of multiple models and through consensus generates powerful textual descriptions. Then, we discuss the relationships between visual content and language and highlight some of the challenges behind this task. Finally, 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 paragraph 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. In this way, we obtain a system that is able to leverage the two level annotation scheme and describe complex activities in natural language.

# Chapter 1

# Introduction

In this work, our goal is to describe complex activities in natural language. In this way, we aim to merge two different, but very popular Computer Vision fields: activity recognition and video to text translation.

In our approach, we begin by first describing videos in sentences by finding the consensus among multiple video to text models. Each model uses an encoder-decoder paradigm, where the encoder receives as input the video sequence and the decoder outputs the description in natural language. These model usually predict valid, well-formed sentences but may lack in semantic meaning, not always accurately depicting what actually happens in the video. The consensus among many models helps in overcoming this problem and in improving the semantic correlation between the generated sentence and the video.

We continue by making an extensive study of this approach in order to find out its limitations. Even though the generated sentences are grammatically correct and describe the visual content, they are not very complex and use a simple vocabulary. As a second major contribution, we make an extensive study on the relationship between text and visual data. In order to eliminate the unknowns related to textual generation part, we conduct most of the experiments on the text-video retrieval task. In this case, given a query in natural language, the goal is to find a video (from a given collection of videos) that is best described by the query. Moreover, based on this analysis we are able to develop a system that achieves state of the art results on the retrieval task.

Lastly, in order to overcome some of the limitations, we further propose a novel hierarchical

approach to video to text translation where we first generate some simpler sentences and then a larger more complex paragraph that describes the whole video content [4]. In order to make such an approach possible and to achieve our goal of describing complex activities through natural language, we propose a novel Video-To-Paragraph dataset that uses a two level annotation scheme. It this way, we are able to design a system that allows fine-grained paragraph description of human activities.

So, our final 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 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, which contains indoor videos from a well-contained ”universe”.

Then, we conduct extensive experiments, which show that our two-stage video to language generation system 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).

# Chapter 2

# Video to text translation

The task of describing videos in natural language is one of the most exciting and still unsolved problems in artificial intelligence today. Solving this task would help decode many important questions about how the mind works, how we perceive the world, how we think and then communicate to one another. Efficient methods for vision to language translation would also have an immense practical value, with applications in many areas ranging from technology to medicine and entertainment.

In this work, we aim to describe human activities in natural language. We start by introducing a general video to text translation model which is described next.

## 2.1 Network architectures

In this section, we will describe all the used architectures by our text to video translation system. We start by describing one of the most used paradigms in video to text translation, sequence-to-sequence (Seq2Seq) and then introduce our proposed architectures.

Early works [1] for video to text translation tackle this problem as a machine translation one. However, things are not that simple. Instead of translating words from one language to another, we now need to understand and translate visual features into language. So, some changes in the pipeline are definitely needed. Videos, being a sequence of frames, are naturally suited to be processed by recurrent neural networks. In this way, most sequence-to-

*This chapter is based on the method published in* Duta, Nicolicioiu, Bogolin and Leordeanu [9]

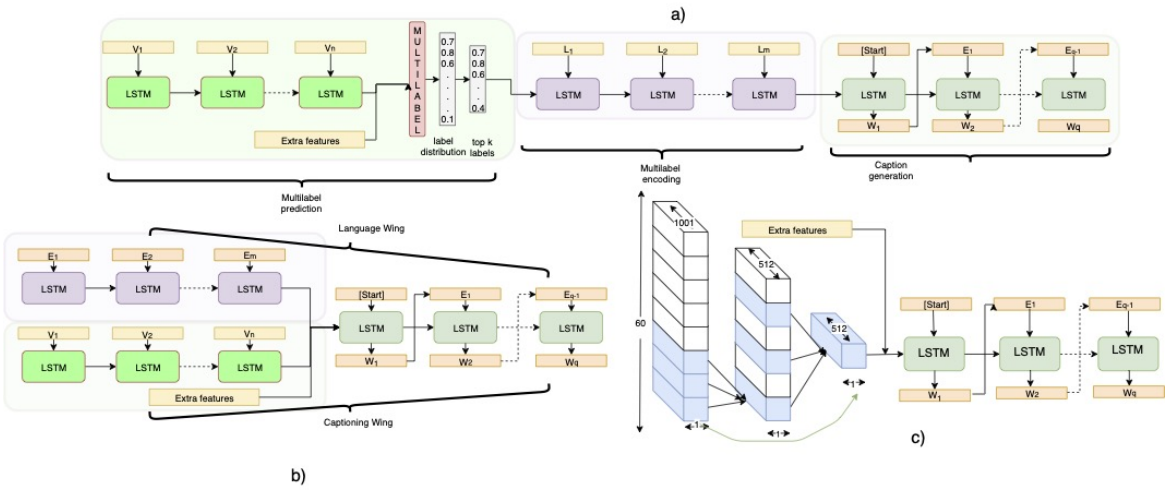



Figure 1: **Main architectures.** a) Two-Stage: vision to words to sentences, b) Two-Wings network, c) Temporal Convolutional Network (TCN). The architectures differ in structure significantly and generally output different sentences, but have a similar overall performance. Picture reused from [9].

sequence models, use LSTMs to produce a video level embedding. This is usually called encoder. Once the video is processed (encoded), then it is fed to a decoder that generates the text, hence the sequence-to-sequence naming scheme.

Now it is time to described our proposed architectures.

**Two-Wings network with sentence reconstruction.** The seq2seq model tends to produce, in experiments, simple sentences with very limited vocabulary. Ideally, we would want a stronger decoder, able to capture more realistic, complex sentences. We aim to accomplish this by a model which we term the **Two-Wings network** due to its dual language and vision encoder (as shown in Figure 1b). The two branches (wings) are trained alternatively, with the decoder having shared parameters. Note that the language reconstruction model is used only during training in order to learn a more powerful decoder. During testing, only the second video to sentence network is used.

**Two-Stage Network, from video to words to sentences.** The second model we propose, the Two-Stage net, puts two encoder-decoder nets one after the other (Figure 1a). The first stage net learns to output words from videos. The second stage net learns to produce sentences from the sets of words given by the first. Thus, word labels provide an intermediate semantic interpretation standing between video data and the final sentence.

**Temporal convolutional network.** Inspired by [2], we adapt their idea of a temporal convo-

| | Group A | | Group A+B | | Group A+B+C | |
|---|---|---|---|---|---|---|
| **Model** | **Cider** | **Meteor** | **Cider** | **Meteor** | **Cider** | **Meteor** |
| Seq2Seq | 36.0 | 25.5 | 44.0 | 27.4 | 46.1 | 28.3 |
| Two-Wings | 32.2 | 25.2 | 42.2 | 27.3 | 46.2 | 28.8 |
| Two-Stage | 34.9 | 25.2 | 43.3 | 27.4 | 45.7 | 28.4 |
| TCN | 36.80 | 25.5 | 43.9 | 27.4 | 46.1 | 28.4 |
| Attention | 41.0 | 26.6 | 44.2 | 27.5 | 46.4 | 28.5 |
| MEAN | 36.0 ±2.5 | 25.6 ±0.5 | 43.9 ±0.6 | 27.4 ±0.2 | 46.0 ±0.9 | 28.4 ±0.3 |
| Best individual model | | | | | 46.2 | 28.8 |

Table 1: **Performance vs features.** Performance of our models using different image, video and audio features added during three experimental phases: Group A - Inception features; Group B - C3D + MFCC audio features; Group C - VGG audio + Y8M word labels features. In each phase we report the average results of each type of models and the average of all the models. Note how additional features pretrained on different tasks significantly improve performance.

lution network architecture (TCN) to replace the recurrent neural network encoder. The idea behind TCN (Figure 1c) is to capture how features change over time by using one dimensional temporal filters. By employing a hierarchy of convolutions with increasing dilation rate, the amount of information combined increases exponentially, over different time scales, until it reduces the temporal dimension to one, to capture global content.

## 2.2 Experimental Analysis

From our experiments, as shown in Table 1 you can see the additional how each used architectures affects the results. Moreover, it can be seen that complementary high level information is brought in by features pre-trained on different tasks. This fact strongly suggests that the intermediate level of semantics captured by these features is important for better bridging the gap between vision and language.

In this chapter, we have introduced three new network architectures for the task of video to text translation. Next we will describe our proposed consensus approach.

# Chapter 3

# Video to text translation through multiple network consensus

In this chapter, we present an approach to address the intrinsic challenges that arise when talking about vision and language. Traditional models suffer from a loss in diversity as compared to humanly generated sentences. Moreover, each sentence is unique and the same message can be said in an unlimited number of ways. In order to solve these challenges, we propose a new approach based on finding the consensual linguistic description among multiple vision to language translation models. While each model individually is able to generate well formed sentences that generally obey grammatical rules, it is the consensus among many models that best captures the hidden meaningful content and significantly outperforms the individual models on the tested evaluation metrics.

## 3.1 Multiple networks consensus

While our models reach a level of accuracy that stands well against published literature, there is a relatively high degree of variation in their output sentences due to the different ways we encode the video content. Some models tend to have complex, descriptive results with a richer vocabulary, while others generate simple, concise sentences.

We noticed that the group of sentences very often contains correct sentences. Moreover, we

*This chapter is based on the method published in* Duta, Nicolicioiu, Bogolin and Leordeanu [9]

observed that models generally produce sentences that gravitate around the correct meaning. Thus, noisy sentence variations could be eliminated if the ensemble of networks could work jointly, as a whole.

Here we propose an efficient **consensus algorithm** for selecting the best sentence in the group, composed of two stages - a first consensus stage using simple agreements between sentences and a second stage that involves training an Oracle network which picks the better sentence from the ones with the top score as follows:

- Step 1: For each sentence in the group, compute its CIDEr score against the others.
- Step 2: Keep the top-C scored sentences.
- Step 3: Re-rank the top C using the Oracle Net and output the top scored sentence.

## 3.2 Experimental analysis

| | **Cider** | **Meteor** | **Rouge** | **Bleu 4** |
|---|---|---|---|---|
| **HRL [26]** | 48.0 | 28.7 | 61.7 | 41.3 |
| **dense [22]** | 48.9 | 28.3 | 61.1 | 41.4 |
| **CIDEnt-RL [17]** | 51.7 | 28.4 | 61.4 | 40.5 |
| **TGM [11]** | 52.9 | **29.7** | - | **45.4** |
| **Ours (Single model)** | 46.2 | 28.4 | - | - |
| **Ours (Consensus)** | **53.8** | **29.7** | **63.0** | 44.2 |

Table 2: **Comparison with the top models on MSR-VTT 2016 test dataset.** We obtain state of the art results on three evaluation metrics.

In Table 2 we compare our method against the top submissions from the MSR-VTT 2016 competition, but also against top models published after the competition on that dataset. The consensus between all models significantly improves the performance achieving state of the art results on several metrics.

In this chapter, we introduced a novel consensus based algorithm that is able to generate better textual descriptions. This means that we are a step closer from our goal of describing actions through natural language. Moreover, we have obtained a general captioning system that may be applied to all kind of videos that achieves state of the art results on a challenging dataset, namely MSR-VTT.

# Chapter 4

# A discussion about the video and text relationship

On a qualitative inspection, we have found out that while the generated sentences usually obey grammatical rules, they are often quite generic quite general. So, we wanted to better understand the relationship between video and language. While our observations are in line with the literature ([20, 7]) that models usually have a less diverse vocabulary and generate shorter sentences, we have also wanted to study the problem from a different angle. Because of this, we considered that the best option in order to better understand the relationship between video and text is to study it for a different task, namely text-video retrieval. In this way, we can study the relationship between video and text without having to account for the unknowns that arise in the language generation part. In this chapter, we will present the key findings that stem from an error analysis step of multiple state of the art models for the text-video retrieval task. So, we investigate cross-modal embeddings learned by modern text-video retrieval systems where the setup is very similar to the video to text translation task.

Text video retrieval is the task where, given a query in natural language the goal is to retrieve a video from a predefined collection of videos that is best described by the query. Cross-modal embeddings form an attractive computational model for the task of retrieving content from large-scale collections of videos with text queries. However, while performance has improved considerably on popular retrieval benchmarks such as LSMDC [21], the models themselves are inscrutable: it remains unclear which factors influence their performance

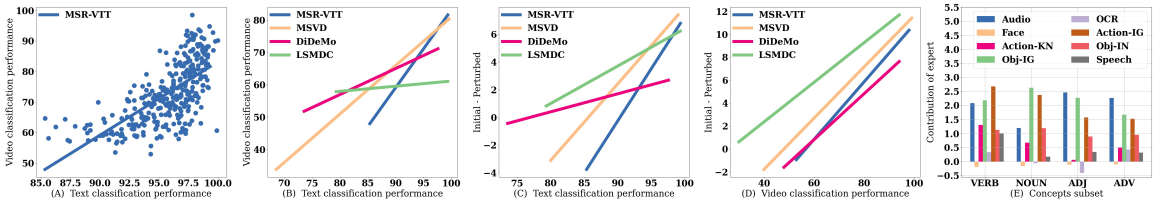



Figure 2: **Analysis on text and video classifiers.** In figure (A) the performance of the text and video classifiers is pictured along with the fitted line for the points (it can easily be observed that the trend is linear). In figure (B) the trends for all datasets are presented for CE. The next two figures show perturbation results (the original retrieval performance in geometric mean minus the performance computed using the perturbed sentences) w.r.t text classification - (C) and video classification - (D) performance. In figure (E) the contribution of each expert for MSR-VTT per part of speech is presented. The contribution of each expert is calculated based on the performance of video classifiers as compared to the baseline - the baseline is the scene expert as presented in [13].

and which factors limit future progress. One interesting aspect arises from the perturbation analysis. The purpose of the perturbation analysis is to find out which words have a higher weight towards a particular retrieval and which words do not count that much. In order to asses this, we have compared the performance obtained by each model before and after removing a particular word from the whole testing set. If the performance changes, then that word is important towards making decisions, while if the performance remains the same, the word does not count towards making a decision.

**Concept Sensitivity.** Identifying words that have a high weight towards a particular retrieval is important, but we wanted to study the relationship between words and the video and text embeddings in order to identify if there is a gap in terms of what concepts can be derived from which embeddings. In order to achieve this, we have trained several linear binary SVM classifiers for each concept, one having as input the video embedding and the other the text embedding for each concept. We have seen that the classifiers usually work well given the text embedding, but not that good for videos. In Fig. 2 (A-D) you can find the performance of the text classifiers compared to the video classifiers for all words. As it can be seen, the performance of the classifiers is correlated, however there is a gap in performance between the textual classifiers and the video classifiers.

**Concepts vs visual input features.** Moreover, we wanted to study if and how the modalities used for training the retrieval model count towards a good classifier performance for words that were present in the pre-trained data used for each modality and which do not count.

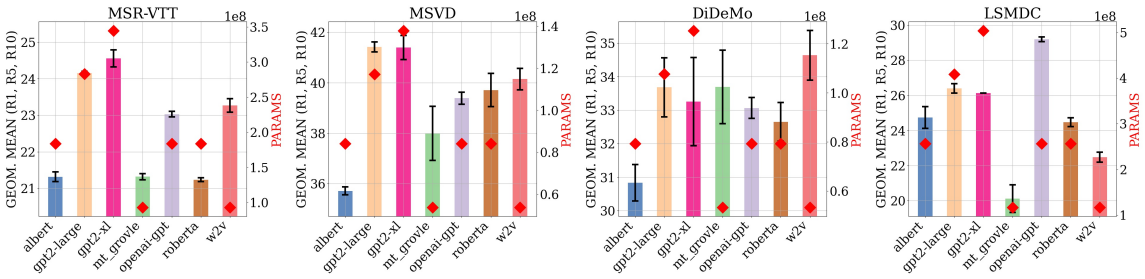



Figure 3: **The influence of various text encoders**. The bars and left y-axes indicate the `t2v` geometric mean of each model at R@1, R@5 and R@10. The red markers and right y-axes indicate the total number of parameters used by each model. As it can be seen, changing the text embedding has a major impact on performance.

As it can be seen in Fig. 2 (E), there is a clear connection between the type of pre-trained modalities and the category of words that can be classified from the video embeddings. Not surprisingly, if we add action recognition features then the category that improves the most are verbs, while adding classification features improves nouns.

Next we wanted to see what happens if you change the used text embedding. So, for our **text embedding study** you can see the influence of changing the text embeddings in Fig. 3, where the y-axis indicates the geometric mean (higher is better) of each model at R@1, R@5 and R@10. We can see that this change has a major impact on the retrieval performance. So, even though these text embeddings are usually pre-trained on vasts amount of textual data, this fluctuation in performance both inter and intra dataset suggests the each text embedding may have an additional information that may be useful for the retrieval task. This is a key insight that stands at the basis of our state of the art text-video retrieval model that we will introduce in the next chapter. In the same time, it also has strong implications for the video to text translation task. It is a strong indication that the correspondence between video and text cannot come from textual data alone. This means, that we need large scale text-video datasets. Moreover, generating a longer text may be problematic. Because of this, we introduce a novel two level annotation scheme that we will describe in Chapter 7.

# Chapter 5

# Text-video retrieval through generalized distillation

As stated before, the goal of this work is the description of complex activities through natural language. However, as discussed in the previous chapter, based on the text-video retrieval task, we were able to better understand the relationship between visual data and natural language. In this chapter, we will briefly present our method for the text-video retrieval task that stems from this analysis. The goal is to develop a system that given a textual query, finds the video that is best described by the query given a collection of videos.

## 5.1 The TEACHTEXT algorithm

By using generalized distillation, we are able to leverage the knowledge from multiple text embeddings and train a student model that is able to achieve state of the art results. The process begins by training multiple teacher models, having the same architecture, but using a different pre-trained text embedding. We then train a student model that is able to learn from the teacher aggregated similarity matrix through generalized distillation. So, given a dataset of text-video pairs, $(x_i, t_j^k)$, the goal is to train a retrieval model $M = (F, Q)$ that learns a joint embedding between video and text. $F$ and $Q$ represent the learnt pre-trained video and text embedding respectively. So, we want to assign a high similarity

*This chapter is based on the method published in* Croitoru, Bogolin, Leordeanu, Jin, Zisserman, Albanie and Liu [6]

| Model | Task | $R@1\uparrow$ | $R@5\uparrow$ | $R@10\uparrow$ | $MdR\downarrow$ | Task | $R@1\uparrow$ | $R@5\uparrow$ | $R@10\uparrow$ | $MdR\downarrow$ | Params |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dual[8] | t2v | 7.7 | 22.0 | 31.8 | 32.0 | v2t | 13.0 | 30.8 | 43.3 | 15.0 | - |
| HGR[5] | t2v | 9.2 | 26.2 | 36.5 | 24.0 | v2t | 15.0 | 36.7 | 48.8 | 11.0 | - |
| MoEE[14] | t2v | $11.1_{\pm 0.1}$ | $30.7_{\pm 0.1}$ | $42.9_{\pm 0.1}$ | $15.0_{\pm 0.0}$ | v2t | $16.5_{\pm 0.1}$ | $43.1_{\pm 0.5}$ | $57.3_{\pm 0.6}$ | $7.7_{\pm 0.5}$ | $400.41M$ |
| CE[13] | t2v | $11.0_{\pm 0.0}$ | $30.8_{\pm 0.1}$ | $43.3_{\pm 0.3}$ | $15.0_{\pm 0.0}$ | v2t | $17.0_{\pm 0.5}$ | $43.5_{\pm 0.4}$ | $57.8_{\pm 0.5}$ | $7.2_{\pm 0.2}$ | $183.45M$ |
| TT-CE | t2v | $11.8_{\pm 0.1}$ | $32.7_{\pm 0.1}$ | $45.3_{\pm 0.1}$ | $13.0_{\pm 0.0}$ | v2t | $19.3_{\pm 0.4}$ | $47.0_{\pm 0.7}$ | $60.0_{\pm 0.4}$ | $6.7_{\pm 0.5}$ | $183.45M$ |
| TT-CE-L | t2v | $13.0_{\pm 0.0}$ | $34.6_{\pm 0.1}$ | $47.3_{\pm 0.2}$ | $12.0_{\pm 0.0}$ | v2t | $22.4_{\pm 0.3}$ | $50.4_{\pm 0.6}$ | $63.8_{\pm 0.3}$ | $5.3_{\pm 0.5}$ | $66.72M$ |
| TT-CE+ | t2v | $\mathbf{15.0}_{\pm 0.1}$ | $\mathbf{38.5}_{\pm 0.1}$ | $\mathbf{51.7}_{\pm 0.1}$ | $\mathbf{10.0}_{\pm 0.0}$ | v2t | $\mathbf{25.3}_{\pm 0.1}$ | $\mathbf{55.6}_{\pm 0.0}$ | $\mathbf{68.6}_{\pm 0.4}$ | $\mathbf{4.0}_{\pm 0.0}$ | $262.73M$ |

Table 3: **MSR-VTT full split: Comparison to state of the art.**

to the pairs $F(x_i), Q(t_i)$ that are fed as input to our model, and a lower similarity to pairs $F(x_i), Q(t_j), i \neq j$. In order to ensure this, we use the max margin retrieval loss $\mathscr{L}_r = \frac{1}{B}\sum_{i=1}^{B}\sum_{i\neq j}\left[max(0, s_{ij} - s_{ii} + m) + max(0, s_{ji} - s_{ii} + m)\right]$ where $s_{ij} = F(x_i)^T Q(t_j)$ denotes the similarity between the encoded video $F(x_i)$ and the encoded text $Q(t_j)$.

The TEACHTEXT algorithm begins by learning several teacher models, that are trained for the retrieval task $T_k$. All models share the same architecture, but use a different text embedding. Once trained, we compute an aggregated similarity matrix $\Phi(S_1, \ldots, S_K) = \frac{1}{K}\sum_{k=1}^{K} S_k$ where $S_i$ is the similarity matrix produced by teacher $T_i, i = 1..K$. Finally, we train our student model that in addition to the retrieval loss $\mathscr{L}_r$ uses a distillation loss $\mathscr{L}_d = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}[l(\Phi(i,j), S_s(i,j))]$ where $l$ represents the Hubber loss and is defined as:

$$l(x,y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{if } |x-y| \leq 1, \\ |x-y| - \frac{1}{2} & otherwise \end{cases} \tag{1}$$

to mimic the aggregated similarity matrix. In this way, we achieve our goal of creating a retrieval system that uses multiple text embeddings.

When it comes to the used architecture, we test our method against several state of the art methods [14, 13, 10] and propose two new variants of the CE [13]: CE+ that has some technical improvements like using the Adam optimizer and gpt2-xl as the text embedding and CE-L - a lightweight variant that uses w2v as the text embedding in order to minimize the number of parameters.

## 5.2 Experimental analysis

In Tab. 3 we make an extensive comparison of our method with other methods from the literature. Moreover, we present the number of parameters of each method where available.

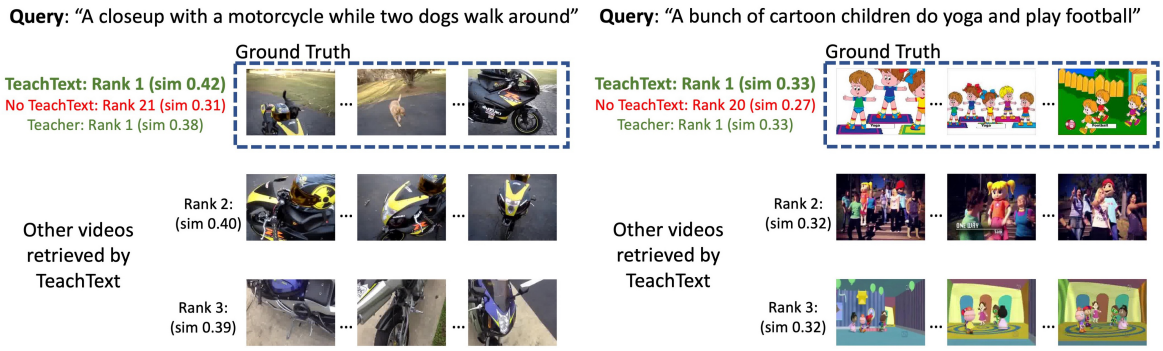



Figure 4: **Qualitative results.** We present the top 3 video retrievals for each query, given by the TEACHTEXT method used on top of a CE+ architecture. Moreover, we show the rank and similarity for the teacher, as well as for the student without using TEACHTEXT for the ground truth video. We mark in green cases where the retrieval is correct in terms of R1 and with red cases where is incorrect. For each of the cases shown, the model learns from the teacher to correct its prediction.

As can be seen, our TEACHTEXT algorithm brings a clear improvement and the total number of parameters remains the same as for the base architecture. Moreover, some qualitative results can be seen in Fig. 4.

In this chapter, we have introduced a novel text-video retrieval method that achieves state of the art results on several datasets. However, maybe the most interesting fact that arises by this usage of multiple text embeddings has direct implications for the video captioning task. In this chapter, we clearly show that despite the huge amount of pre-training, there still is a domain gap between vision and language. By using different pre-trained text embeddings we are able to diminish this gap, however the implications for the captioning task remain. Since for captioning an additional generation step is needed, this is leaned directly from the annotations from the benchmark on which the model is tested. Having this in mind, we argue that a new type of annotations are needed that we will introduce in the next chapter.

# Chapter 6

# A novel Videos-to-Paragraphs Dataset

As we have show in the previous two chapters, by using multiple pre-trained text embeddings, we achieve a big gain in performance for the text-video retrieval task. This aspect is maybe even more interesting for the video captioning task. More exactly, we show that even current text embeddings ([15, 19]) are pre-trained on huge amounts of textual data, there still is a domain gap between vision and language. So, in order to tackle this problem, we propose a new dataset that is annotated at two semantic levels that we will describe next.

So, we introduce our Videos-to-Paragraphs dataset (with appropriate annotations at two linguistic levels), which contains indoor videos from a well-contained ”universe”. The scenes, actors and their activities are centered around humans, in the context of what typically happens in a classroom. In this way, we argue that we can better study how a video captioning system works, but most importantly, we are able to accurately describe human actions through natural language. While it is limited to a simpler setup, a classroom, we consider this to be a first step towards having an automated system that is able to accurately describe human activities.

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

*This chapter is based on the method published in* Bogolin, Croitoru and Leordeanu [4]

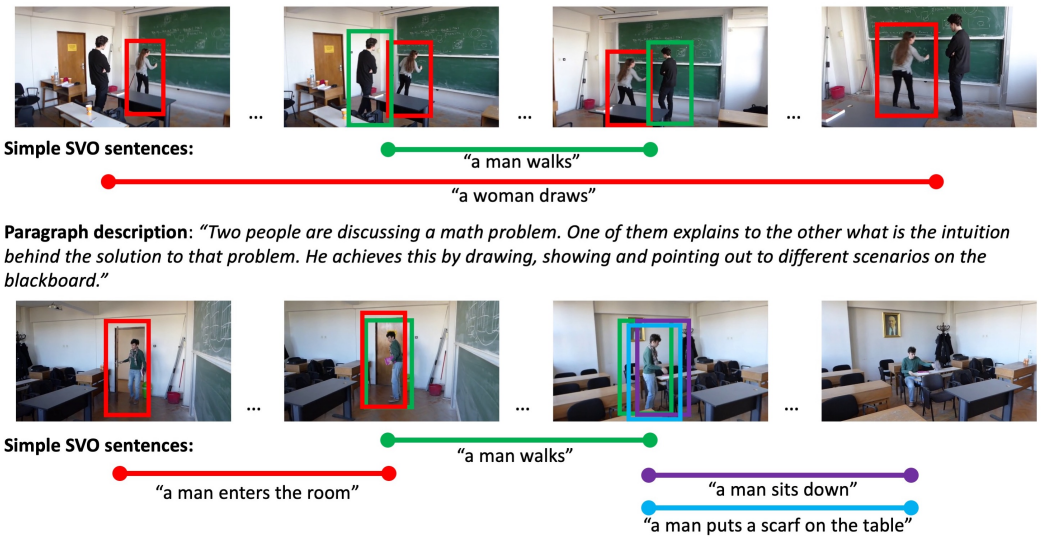



Figure 5: **Examples from Videos-to-Paragraphs Dataset.** We present annotations for two different videos. For each we present a few SVO (subject, verb, object) events along with: their start frame, end frame, corresponding space-time bounding box that contains the event, and its simple sentence description. At the bottom we present the paragraph-level video description. The paragraphs contain longer and more complex sentences than the SVOs. Picture reused from [4].

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 (Figure 5). 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.

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.

In this chapter we described how we collected and created our new Videos-to-Paragraphs dataset. It is the first step needed into achieving our goal of describing human activities through natural language. Moreover, we have introduced a novel two level annotation scheme that we argue stands at the basis of creating a hierarchical captioning model that is able to generate longer, more complex sentences. However, in order to make use of this new annotation scheme, a new hierarchical approach is needed which we will introduce in the next chapter.

# Chapter 7

# A hierarchical representation for video to text generation

As pointed out in the previous chapter, our final 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 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 prob-

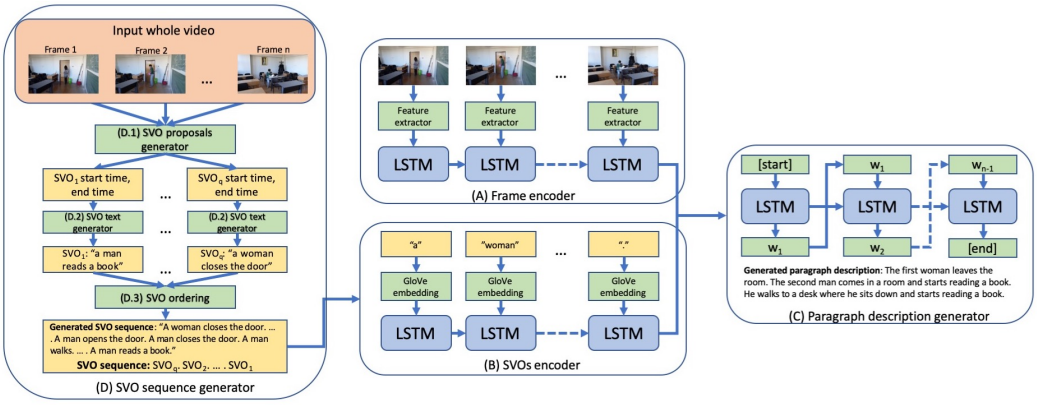



Figure 6: **System overview.** Our system consists of several modules that work together to form two encoding pathways, one that processes the video frames (module *(A) Frame encoder*) and the other that processes the SVOs (module *(B) SVOs encoder* and module *(D) SVOs sequence generator*). At the end, 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. Picture reused from [4].

lem.

It is time to introduce our full system as described in Figure 6. The process begins by pre-extracting the ImageNet [23] features for all the videos in the dataset. These features will be the input to the *Frame encoder*. Similarly, for the text side, we extract the GloVe [18] embeddings. These will be fed into the *SVOs encoder*. Now it is time to train the *Paragraph description generator*. Now, for inference we do not have the SVO sequence. So, we need to predict it. We next train the *SVO text generator*. It receives as input the video features corresponding to each clip that has a SVO associated. Next, we generate the SVO proposals and then use the trained text generator to generate a sentence for each proposal. Finally, we order the predicted SVOs and then we can use them as input for the *Paragraph description generator* at inference time. In this way, we achieve our goal of generating longer, paragraph descriptions.

| 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 |
| [9] | 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 4: **Comparison with state of the art methods.** We use 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

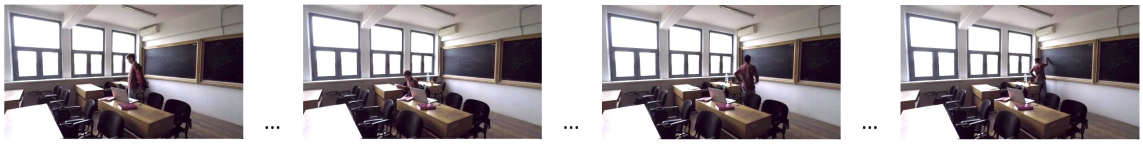
## 7.1 Experimental analysis

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 [16], ROUGE [12], CIDEr [24] and METEOR [3].

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 [25]) and a recent state-of-the-art method [9] (see Table 4). Please note that for a fair comparison, we compare against their best single model which uses a sequence-to-sequence approach and an attention mechanism. Moreover, in Figure 7 you can find some qualitative examples. As it can be seen, the generated paragraph accurately describes what happens in the scene.

In this way, we introduce a novel hierarchical model that is able to leverage the advantage given by the two level annotation scheme introduced with our new Videos-to-Paragraph dataset. We start by 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. In this way, we are able to describe human activities through natural language. Moreover, one of the major things that we think



**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 7: **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.

can be improved in future work is the usage of a graph neural network in order to better capture the semantic links between SVOs. In this work, we have made a first attempt into integrating link information which turned out to be successful, but we feel like this aspect can be further improved.

# Chapter 8

# Conclusions

In this work, we aim to describe complex activities through natural languages. We argue that video to text translation is the best way to describe such activities since the natural language is so versatile that you can closely capture and describe what is happening without being limited to a fixed set of labels. So, we first present several different architectural approaches to encode content of a video for the purpose of learning to generate captions.

On close inspection, we found that very small changes in the sentence structure or at the level of words, without changing the overall meaning, may strongly change the metric scores. Moreover, in order to better understand the relationship between video and text, we make a thorough analysis on a different, but closely related task: text-video retrieval. From this analysis we may draw the following conclusion: there still is a domain gap between visual data and text. Lastly, we introduced a novel, hierarchical approach to language generation from videos that makes use of the two level annotation scheme. 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. 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.

# Bibliography


[1] Bahdanau, D., Cho, K. and Bengio, Y. [2014], 'Neural machine translation by jointly learning to align and translate', *arXiv preprint arXiv:1409.0473* .

[2] Bai, S., Kolter, J. Z. and Koltun, V. [2018], 'An empirical evaluation of generic convolutional and recurrent networks for sequence modeling', *arXiv preprint arXiv:1803.01271* .

[3] Banerjee, S. and Lavie, A. [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', Vol. 29.

[4] Bogolin, S.-V., Croitoru, I. and Leordeanu, M. [2020], A hierarchical approach to vision-based language generation: from simple sentences to complex natural language, *in* 'Proceedings of the International Conference on Computational Linguistics (COLING)'.

[5] Chen, S., Zhao, Y., Jin, Q. and Wu, Q. [2020], Fine-grained video-text retrieval with hierarchical graph reasoning, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.

[6] Croitoru, I., Bogolin, S.-V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S. and Liu, Y. [2021], 'Teachtext: Crossmodal generalized distillation for text-video retrieval', *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* .

[7] Cui, Y., Yang, G., Veit, A., Huang, X. and Belongie, S. [2018], Learning to evaluate image captioning, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.

[8] Dong, J., Li, X., Xu, C., Ji, S. and Wang, X. [2019], Dual dense encoding for zero-example video retrieval, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.

[9] Duta, I., Nicolicioiu, A. L., Bogolin, S.-V. and Leordeanu, M. [2018], 'Mining for meaning: from vision to language through multiple networks consensus', *The British Machine Vision Conference (BMVC)* p. 275.

[10] Gabeur, V., Sun, C., Alahari, K. and Schmid, C. [2020], 'Multi-modal transformer for video retrieval', *Proceedings of the European Conference on Computer Vision (ECCV)* .

[11] Jin, Q., Chen, S., Chen, J. and Hauptmann, A. [Proceedings of the International Conference on Multimedia], 'Knowing yourself: Improving video caption via in-depth recap'.

[12] Lin, C.-Y. [2004], Rouge: A package for automatic evaluation of summaries, *in* 'Proceedings of ACL Workshop on Text Summarization Branches Out', p. 10.

[13] Liu, Y., Albanie, S., Nagrani, A. and Zisserman, A. [2019], 'Use what you have: Video retrieval using representations from collaborative experts', *The British Machine Vision Conference (BMVC)* .

[14] Miech, A., Laptev, I. and Sivic, J. [2018], 'Learning a text-video embedding from incomplete and heterogeneous data', *arXiv preprint arXiv:1804.02516* .

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

[16] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. [2002], Bleu: a method for automatic evaluation of machine translation, *in* 'Proceedings of the Association for Computational Linguistics (ACL)'.

[17] Pasunuru, R. and Bansal, M. [2017], Reinforced video captioning with entailment rewards, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, pp. 979–985.

[18] Pennington, J., Socher, R. and Manning, C. [2014], Glove: Global vectors for word representation, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)'.

[19] Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. [2018], 'Improving language understanding by generative pre-training', *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf* .

[20] Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T. and Saenko, K. [2018], 'Object hallucination in image captioning', *arXiv preprint arXiv:1809.02156* .

[21] Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A. and Schiele, B. [2017], 'Movie description', *International Journal of Computer Vision (IJCV)* **123**(1), 94–120.

[22] Shen, Z., Li, J., Su, Z., Li, M., Chen, Y., Jiang, Y.-G. and Xue, X. [2017], Weakly supervised dense video captioning, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.

[23] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. [2016], Rethinking the inception architecture for computer vision, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.

[24] Vedantam, R., Lawrence Zitnick, C. and Parikh, D. [2015], Cider: Consensus-based image description evaluation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.

[25] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K. [2015], Sequence to sequence-video to text, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)'.

[26] Wang, X., Chen, W., Wu, J., Wang, Y.-F. and Wang, W. Y. [2018], 'Video captioning via hierarchical reinforcement learning', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* .