VEERMATA JIJABAI TECHNOLOGICAL INSTITUTE

BTECH PROJECT

DEPARTMENT OF COMPUTER ENGINEERING & INFORMATION TECHNOLOGY

Dense Video Captioning

Ganadhish Acharekar
Akshat Shah
Arnav Shah
Saharsh Jain

Project Supervisor
Prof. Sandip T. Shingade

Contents

1 Introduction

1.1 Problem Background

There has been a significant progress in the fields of *Computer Vision* (CV) and *Natural Language Processing* (NLP) and other artificial intelligence problems in recent decades. In computer vision, tasks like object detection and classification are well-explored, while in natural language processing, sequence to sequence tasks like machine translation have been well worked upon. The most promising solutions to these tasks come from deep learning methodologies, outperforming techniques like rule-based systems and other traditional machine learning algorithms. Owing to the rich feature representation capabilities of different kinds of neural networks, tasks that are more natural to humans are becoming solvable by machines, especially perceptive and understanding tasks.

The task of image captioning combines the fields of CV and NLP. Video action recognition, which involves classifying the action in a short video clip, is a task that can be considered as the representative task for video understanding. Taking one step further from action recognition and image captioning, the task of video captioning involves labelling a video clip with a single natural language sentence. However, for long videos, a single sentence is not enough to describe all events that may occur in a video. Thus, the task of **Dense Video Captioning (DVC)** was introduced [1]. It involves generating natural language descriptions for each of the multiple events that may occur in a video.

Describing a short video in natural language is trivial for humans, but a challenging task for machines. Automatic video description involves understanding of many entities and the detection of their occurrences in a video employing computer vision techniques. These entities include background scene, humans, objects, human actions, human-object interactions, human-human interactions, other events, and the order in which events occur. All this information should then be articulated using comprehensible and grammatically correct text, employing NLP techniques[2]. Indeed, the task of dense video captioning can be thought of as the culmination of all the perceptive learning tasks of computer vision and the natural language generation task.

As we move towards a digital age, more and more content is generated in the form of multimedia, e.g., videos. The applications of dense video captioning are in tasks such as video summarization, video retrieval (search and indexing), video object detection, video segment localization via queries, increased accessibility to the visually impaired (with combined use of DVC and text-to-speech).

1.2 Problem Formulation

The task of dense video captioning involves two sub-tasks (1) temporal localization of events in a video and (2) describing the localized events in natural language. Given an input video $v = \{v_1, v_2, ..., v_T\}$ where v_i represents i^{th} video frame in temporal order, the target of dense video captioning task is to output a set of sentences $S = \{s_1, s_2, ..., s_{N_s}\}$ where N_S is the number of sentences and $s_i = \{t_i^{start}, t_i^{end}, \{w_j\}\}$ consists of start and end timestamps for each sentence described with set of words from a vocabulary set $w_j \in V$.

Most of the architectures for dense video captioning comprises of two components (1) Proposal module and (2) Captioning module. The modules can be trained in different ways like separate training, alternative training or in a end-to-end manner. The task of Proposal module is to input video frames v and output event proposals $P = \{t_i^{start}, t_i^{end}, confidence_i\}$. Depending on the architecture, the proposal module can also output additional parameters like in [1, 3, 4] which can be utilized by captioning module. The proposals can also be in the form of offsets (center and length) as in [3, 5, 6, 7, 8]. The captioning module outputs descriptions for each of the proposed event in usually a single sentence.

1.3 Datasets

1.3.1 Textually Annotated Cooking Scenes (TACoS)

It is a subset of MP-II Composites. TACoS was further processed to provide coherent textual descriptions for high quality videos. The TACoS dataset was constructed by filtering through MP-II Composites, while restricting to only those activities that involve manipulation of cooking ingredients, and has at least 4 videos for the same activity. As a result, TACoS contains 26 fine grained cooking activities in 127 videos. For each video, 20 different textual descriptions were collected. The dataset comprises 11,796 sentences containing 17,334 actions descriptions. The dataset also provides the alignment of sentences describing activities by obtaining approximate time stamps where each activity starts and ends.

1.3.2 TACoS-MultiLevel

It was collected on the TACoS corpus. For each video in the TACoS corpus, three levels of descriptions were collected that include: (1) detailed description of video with no more than 15 sentences per video; (2) a short description that comprises 3-5 sentences per video; and finally (3) a single sentence description of the video. Annotation of the data is provided in the form of tuples such as object, activity, tool, source and target with a person always being the subject.

1.3.3 Microsoft Video Description (MSVD)

It comprises 1,970 YouTube clips with human annotated sentences. The audio is muted in all clips to avoid bias from lexical choices in the descriptions. Furthermore, videos containing subtitles or overlaid text were removed during the quality control process of the dataset formulation. The duration of each video in this dataset is typically between 10 to 25 seconds mainly showing one activity. The dataset comprises multilingual (such as Chinese, English, German etc) human generated descriptions.

1.3.4 Montreal Video Annotation Dataset (M-VAD)

It is based on the Descriptive Video Service (DVS) and contains 48,986 video clips from 92 different movies. Each clip is spanned over 6.2 seconds on average and the entire time for the complete dataset is 84.6 hours. The total number of sentences is 55,904, with few clips associated with more than one sentence. The vocabulary of the dataset spans about 17,609 words (Nouns-9,512: Verbs-2,571: Adjectives-3,560: Adverbs-857). The dataset split consists of 38,949, 4,888 and 5,149 video clips for training, validation and testing respectively.

1.3.5 MPII Movie Description Corpus (MPII-MD)

It contains transcribed audio descriptions extracted from 94 Hollywood movies. These movies are subdivided into 68,337 clips with an average length of 3.9 seconds paired with 68,375 sentences amounting to almost one sentence per clip. Every clip is paired with one sentence that is extracted from the script of the movie and the audio description data. The Audio Descriptions (ADs) were collected first by retrieving the audio streams from the movie using online services MakeMkV 1 and Subtitle Edit 2. Then the transcribed texts were aligned with associated spoken sentences using their time stamps. The total time span of the dataset videos is almost 73.6 hours and the vocabulary size is 653,467.

1.3.6 MSR Video-to-Text (MSR-VTT)

It contains a wide variety of open domain videos for video captioning tasks. It comprises 7180 videos subdivided into 10,000 clips. The clips are grouped into 20 different categories.. The dataset is divided into 6513 training, 497 validation and 2990 test videos. Each video comprises 20 reference captions annotated by AMT workers. In terms of the number of clips with multiple associated sentences, this is one of the largest video captioning datasets. In addition to video content, this dataset also contains audio information that can potentially be used for multimodal research.

1.3.7 ActivityNet Captions

It contains 100k dense natural language descriptions of about 20k videos from ActivityNet that correspond to approximately 849 hours. On average, each description is composed of 13.48 words and covers about 36 seconds of video. There are multiple descriptions for every video and when combined together, these descriptions cover 94.6% content present in the entire video. In addition, 10% temporal overlap makes the dataset especially interesting and challenging for studying multiple events occurring at the same time.

1.3.8 ActivityNet Entities

It is the first video dataset with entities grounding and annotations. This dataset is built on the training and validation splits of the ActivityNet Captions dataset, but with different captions. In this dataset, noun phrases (NPs) of video descriptions have been grounded to bounding boxes in the video frames. The dataset comprises 14281 annotated videos, 52k video segments with at least one noun phrase annotated per segment and 158k bounding boxes with annotations. The dataset employs a training set (10k) similar to ActivityNet

Captions. However, the validation set of ActivityNet Captions is randomly and evenly split into ANet-Entities validation (2.5k) and testing (2.5k) sets.

1.3.9 YouCook

It consists of 88 YouTube cooking videos of different people cooking various recipes. The background (kitchen / scene) is different in most of the videos. The dataset is divided into six different cooking styles, for example grilling, baking etc. For machine learning, the training set contains 49 videos and the test set contains 39 videos. The object categories for the dataset include "utensils", "bowls" and "food" etc.

1.3.10 YouCook2

YouCook-II Dataset consists of 2000 videos uniformly distributed over 89 recipes. The cooking videos are sourced from YouTube and offer all challenges of open domain videos such as variations in camera position, camera motion and changing backgrounds. The complete dataset spans a total play time of 175.6 hrs and has a vocabulary of 2600 words. The videos are further divided into 3-16 segments per video with an average of 7.7 segments per video elaborating procedural steps. Individual segment length varies from 1 to 264 seconds. The average length of each video is 316 seconds reaching up to a maximum of 600 seconds. The dataset is randomly split into train, validation and test sets with the ratio of 66

1.3.11 MP-II Cooking

Max Planck Institute for Informatics (MP-II) Cooking dataset comprises 65 fine grained cooking activities, performed by 12 participants preparing 14 dishes such as fruit salad and cake etc. The 65 cooking activities include "wash hands", "put in bowl", "cut apart", "take out from drawer" etc. When the person is not in the scene for 30 frames (one second) or is performing an activity that is not annotated, a "background activity" is generated. In total, the dataset comprises 44 videos (888,775 frames), with an average length per clip of approximately 600 seconds. The dataset spans a total of 8 hours play length for all videos, and 5,609 annotations.

1.3.12 VideoStory

VideoStory is a multi sentence description dataset comprising 20k social media videos. This dataset is aimed to address the story narration or description generation of long videos that may not sufficiently be illustrated with a single sentence. Each video is paired with at least one paragraph. The average number of temporally localized sentences per paragraph are 4.67. There are a total of 26245 paragraphs in the dataset comprising 123k sentences with an average of 13.32 words per sentence. On average, each paragraph covers 96.7% of video content. The dataset contains about 22% temporal overlap between co-occurring events. The dataset has training, validation and test split of 17908, 999, and 1011 videos respectively and also proposes a blind test set comprising 1039 videos.

1.3.13 Charades

It contains 9848 videos of daily indoor household activities. Videos are recorded in 15 different indoor scenes and restricted to use 46 objects and 157 action classes only. The dataset comprises 66500 annotations describing 157 actions. It also provides 41104 labels to its 46 object classes. Moreover, it contains 27847 descriptions covering all the videos. The videos in the dataset depict daily life activities with an average duration of 30 seconds. The dataset is split into 7985 and 1863 videos for training and test purposes respectively.

1.3.14 Video Titles in the Wild (VTW)

It contains 18100 video clips with an average of 1.5 minutes duration per clip. Each clip is described with one sentence only. However, it incorporates a diverse vocabulary, where on average one word appears in not more than two sentences across the whole dataset. Besides the single sentence per video, the dataset also provides accompanying descriptions (known as augmented sentences) that describe information not present in the visual content of the clip. The dataset is proposed for video title generation as opposed to video content description but can also be used for language-level understanding tasks including video question answering.

1.3.15 Kinetics

A collection of large-scale, high-quality datasets of URL links of up to 650,000 video clips that cover 400/600/700 human action classes, depending on the dataset version. The videos include human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands and hugging. Each action class has at least 400/600/700 video clips. Each clip is human annotated with a single action class and lasts around 10 seconds.

Dataset	Domain	#classes	#videos	Avg len	#clips	#sent	#words	vocab	len (hrs)
TACoS	cooking	26	127	360 sec	7,206	18,227	146,771	28,292	15.9
TACos Multilevel	cooking	1	185	360 sec	14,105	52,593	2,000	-	27.1
MSVD	open	218	1970	10 sec	1,970	70,028	607,339	13,010	5.3
M-VAD	movie	-	92	6.2 sec	48,986	55,904	519,933	17,609	84.6
MPII-MD	movie	-	94	3.9 sec	68,337	68,375	653,467	24,549	73.6
MSR-VTT	open	20	7,180	20 sec	10,000	200,000	1,856,523	29,316	41.2
ActivityNet Captions	open	-	20,000	180 sec	-	100,000	1,348,000	-	849.0
ActivityNet Entities	social media	-	14,281	180 sec	52k	-	-	-	-
YouCook	cooking	6	88	-	Nil	2,688	42,457	2,711	2.3
YouCook II	cooking	89	2,000	316 sec	15.4k	15.4k	-	2,600	176.0
MP-II Cooking	cooking	65	44	600 sec	-	5,609	-	-	8.0
VideoStory	social media	-	20k	-	123k	123k	-	-	396.0
Charades	human	157	9,848	30 sec	-	27,847	-	-	82.01
VTW	open	-	18,100	90 sec	-	44,613	-	-	213.2
Kinetics 700	human	700	6,50,000	10 sec	700	-	-	-	1806

Table 1: Datasets and their characteristics

1.4 Evaluation Metrics

Text Generation is a tricky domain. Academics as well as the industry still struggle for relevant metrics for evaluation of the generative models' qualities. Every generative task is different, having its own subtleties and peculiarities.These metrics can be applied to the numerous tasks such as:

- short or long-form text generation
- Machine Translation
- Summarisation
- Chatbots and dialogue systems
- Multimedia systems like speech2text, image/video captioning

Following are certain metrics that are used to evaluate the above natural langauge generation tasks:

1.4.1 BLEU: Bilingual Evaluation Understudy)

BLEU [9] is a precision focused metric that calculates n-gram overlap of the target and generated texts. This n-gram overlap means that this evaluation scheme is word-position independent apart from n-grams' term associations. BLEU also consists of a brevity penalty i.e. a penalty applied when the generated text is too small compared to the target text.

1.4.2 ROUGE: Recall Oriented Understudy for Gisting Evaluation

ROUGE [10] is a set of metrics for evaluating automatic summarization of texts as well as machine translations. There are 3 types of ROUGE: ROUGE-N, the most common ROUGE type which means n-gram overlap. Second is ROUGE-L which checks for Longest Common Subsequence instead of n-gram overlap. The third is ROUGE-S which focuses on skip grams.

1.4.3 METEOR: Metric for Evaluation of Translation with Explicit Ordering

METEOR [11] is an metric that works on word alignments. It computes one to one mapping of words in generated and reference texts. Traditionally, it uses WordNet or porter stemmer. Finally, it computes an F-score based on these mappings. The metric was designed to fix some of the problems found in the more popular BLEU metric, and also produce good correlation with human judgement at the sentence or segment level. This differs from the BLEU metric in that BLEU seeks correlation at the corpus level.

1.4.4 CIDEr: Consensus based Image Description Evaluation

CIDEr [12] is a metric used to evaluate image descriptions that uses human consensus. It uses a triplet-based method of collecting human annotations to measure consensus followed by stemming and grouping the words into n-grams. This metric is especially useful in image and video annoated tasks.

1.4.5 WMD: Word Mover's Distance

WMD [13] is a fundamental technique for measuring the similarity of two documents. As the crux of WMD, it can take advantage of the underlying geometry of the word space by employing an optimal transport formulation. WMD leverages the results of advanced embedding techniques like word2vec and Glove. It suggests that distances between embedded word vectors are to some degree, semantically meaningful. It utilizes this property of word vector embeddings and treats text documents as a weighted point cloud of embedded words.

1.4.6 SPICE: Semantic Propositional Image Captioning Evaluation

SPICE [14] is an automated caption evaluation metric defined over scene graphs. It first transforms both generated and target captions into an intermediate representation that encodes semantic propositional content. It then creates a scene graph based on certain object classes, attribute types and relations which in turn is used to calculate the F-score b/w the generated and target captions.

2 Literature Review

The problem of dense video captioning was introduced by Krishna *et al* in [1] by proposing ActivityNet Captions dataset. Their architecture involved a proposal module and captioning module. The proposal module was inspired from DAPs [15], while the captioning module incorporated LSTM with contextual features along with event feature as inputs. The architecture was able to detect events and generate captions of the video in single pass without the need of time consuming sliding window approach. However, since the features were dependent on the end location of the event, the model generated same captions for events ending at same timestamp.

Following the release of ActivityNet Captions dataset in 2017, many researchers were able to surpass the results of baseline model and achieve state-of-the-art. Zhou *et al* [8] addressed the problem of little influence of language desciptions on event proposals if the two modules are trained separately. They introduced an end-to-end masked transformer for propogating captioning error to proposal module for better performance. Furthermore, they proposed a self-attention mechanism for learning long-range dependencies in video. The proposal module was based on ProcNets [16]. The caption decoder employed a differentiable proposal mask to account for features in the respective event. Wang *et al* [4] employed Bi-SST as their proposal generator to account for past and future context information. The captioning module consisted of attentive fusion of context features, the weights of which were decided by a context gating mechanism. The architecture selected final captions based on joint ranking method which accounted for both proposal and caption confidence. Li *et al* [3], inspired by object localization networks like [17, 18], presented an end-to-end model with descriptiveness regression. An attribute-augmented LSTM network optimized using reinforcement learning was used for captioning module. Xiong *et al* [19] strived to generate relevant, coherent and concise descriptions using SSN [20] for event localization and LSTM for event selection and caption generation. Reinforcement learning with sentence-level reward is used to train the captioning network. Xu *et al* [6] proposed JEDDi-Net, an end-to-end architecture incorporating visual and language contexts. Segment Proposal Network inspired from R-C3D [21] is used for proposal generation. Hierarchical LSTM with caption-level controller network and word-level sentence decoder is used for caption generation. The proposal features are represented using 3D

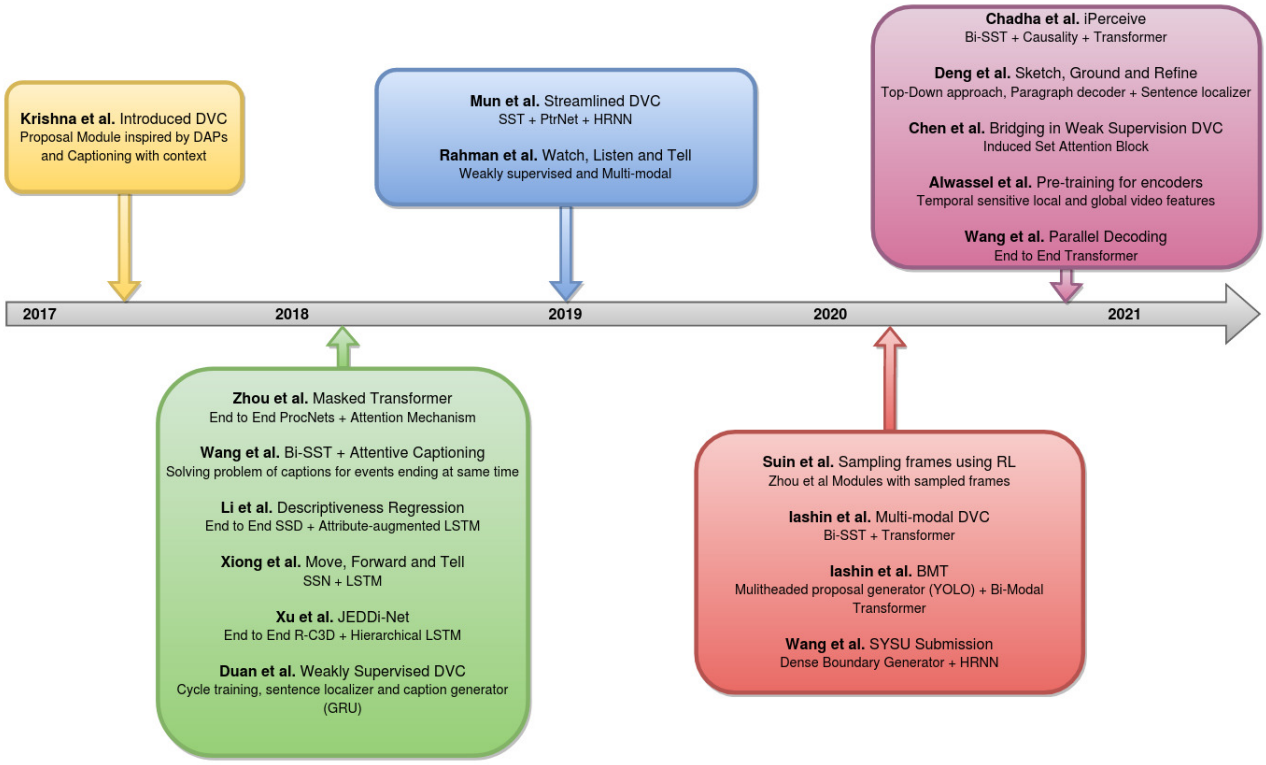

Figure 1: Evolution of dense video captioning methods over time.

Segment-of-Interest Pooling. Duan *et al* [22] introduced weak supervision training with no need of temporal annotations of events. They trained sentence localizer and caption generator in cyclic manner minimizing the reconstruction loss. However, the method struggles to detect beginning of events properly.

Mun *et al* [23] tackles the challenge of coherent captioning by considering temporal dependency between events. They used SST for event proposals, PtrNet for event sequence generation and HRNN for captioning. Rahman *et al* [24] utilized weak supervision method from [22] and were the first to try multi-modal approach for dense video captioning. They show how audio alone can be competitive to the previous visual based results. The paper also discusses various methods to encode audio (MFCC, CQT, SoundNet) and context fusion techniques for multiple modalities (Multiplicative Mixture, Multi-modal context fusion, MUTAN). The architecture suffered from proposal localization accuracy due to weak supervision. Furthermore, the results are also affected due to unavailability of part of the dataset as some videos are not available on their respective urls.

Suin *et al* [25] aimed to reduce computational cost by processing fewer frames. They used deep reinforcement-based approach to describe multiple events in a video by watching a portion of the frames. The event proposal and captioning modules were inspired from [8]. Iashin *et al* [26] shows the importance of audio and speech modalities alongside visual features for dense video captioning task. They employ a Bi-SST for proposal and the captioning module consists of a Transformer architecture with three blocks: an encoder, decoder and generator. The model was able to achieve better performance than the then existing methods despite of unavailability of full dataset for multiple modalities. Iashin *et al* [5] utilized audio and video with Bi-modal Transformer for captioning. The proposal generator consisted of multiheaded method, inspired from YOLO object detector [27]. Their ablation analysis depicted stronger contribution of visual cues alone than audio cues alone. However, both modalities combined gave better results. Wang *et al* [28] adapt DBG [29] for temporal event proposals alongwith ESGN [23] for candidate subset selection. The captioning module consists of an encoder-decoder architecture with CMG (cross-modal gating) block to adaptively balance the visual and linguistic information.

Chadha *et al* [30] proposed to handle cognitive error (causality between events) and incorrect attention (attending to objects in the frame). Their end-to-end model consisted of Bi-SST for proposals, Common-Sense Reasoning for causality learning and Transformer based architecture [26] for captioning. The common-sense

Paper	Set	Ground Truth Proposals				Learnt Proposals			
		M	C	B@3	B@4	M	C	B@3	B@4
Krishna *et al* [1] DCE	Validation	8.88	25.12	4.09	1.60	5.69	12.43	1.90	0.71
	Test	9.46	24.56	7.12	3.98	4.82	17.29	3.86	2.20
Zhou *et al* [8] Masked Transformer	Validation	11.16	47.71	5.76	2.71	4.98	9.25	2.42	1.15
	Test	10.12				10.12			
Wang *et al* [4] Bi-SST	Validation	10.89				5.86	7.99	2.55	1.31
	Test					9.65			
Li *et al* [3] Descriptivness Regr.	Validation	10.33	26.26	4.55	1.71	6.93	13.21	2.27	0.74
	Test					12.96			
Xu *et al* [6] JEDDi-Net	Test			4.06	1.63	8.58	19.88	4.06	1.63
Mun *et al* [23] Streamlined DVC	Validation	13.07	43.48	4.41	1.28	8.82	30.68	2.94	0.93
	Test					8.19			
Duan *et al* [22] Weakly Supervised DCE	Test			2.62	1.27	6.3	18.77	2.62	1.27
Rahman *et al* [24] Watch, Listen, Tell	Validation *	7.23	25.36	3.04	1.46	4.93	13.79	1.85	0.9
Suin *et al* [25] Efficient framework for DVC	Validation			2.87	1.35	6.21	13.82	2.87	1.35
Iashin *et al* [26] Multi-modal DVC	Validation *	11.72		5.83	2.86	7.31		2.6	1.07
Iashin *et al* [5] BMT	Validation *	10.90		4.63	1.99	8.44		3.84	1.88
Xiong *et al* [19] Move Forward and Tell	Validation			2.84	1.24	7.08		2.84	1.24
Wang *et al* [28] SYSU	Validation	14.85				10.31			
	Test					9.28			
Chadha *et al* [30] iPerceive	Validation *	12.27		6.13	2.98	7.87		2.93	1.29
Deng *et al* [7] Sketch, Ground and Refine	Validation				1.67	9.37	22.12		1.67
Chen *et al* [31] Towards Bridging EC-SL	Validation			2.78	1.33	7.49	21.21	2.78	1.33
Alwassel *et al* [32] TSP with BMT	Validation			4.16	2.02	8.75		4.16	2.02
Wang *et al* [33] Parallel decoding	Validation *	11.26	53.65		3.12	8.08	28.59		1.96

Table 2: Performance comparison of previous methods on the ActivityNet Captions dataset (* some videos unavailable)

reasonining module employed a borrow-put experiment for determining dependency between events and generate context-aware features. Deng *et al* [7] introduced a top-down approach, reversing the usual detect-then-describe method. The architecture first generates a multi-sentence paragraph to describe the whole video and then localize each sentence for events. The captions are then refined using dual-path cross attention module. The top-down approach increased coherency in captions. Chen *et al* [31] worked on closely bridging event localization and captioning modules for weakly supervised learning. The Induced Set Attention Block helps the captioner to learn highly abstracted global structure of the video. The method suffers from detecting visually small concepts/objects. Alwassel *et al* [32] introduce a supervised pre-training paradigm for temporal action localization. The paradigm also considers background clips and global video information, to improve temporal sensitivity. Wang *et al* [33] formulated dense video captioning as a set prediction task and introduced an end-to-end dense video captioning framework with parallel decoding (PDVC). PDVC adopts the vision transformer to learn attentive interaction of different frames. Two prediction heads run in parallel over query features, leveraging the mutual benefits between two tasks of localization and captioning.

2.1 Identified Themes

Dense video captioning can be decomposed into two parts: event localization and event description. Existing research methodologies can be grouped into different categories based on training methods, modalities used and ordering of tasks.

Based on Training schemes:

1. **Independent training**
2. **Alternate training**: alternate between i) training the proposal module only and ii) training the captioning module on the positive event proposals while fine-tuning the proposal module.
3. **End-to-End**

4. **Weakly Supervised**

Based on Modalities:

1. **Uni-modal**: Only visual features
2. **Multi-modal**: Combinations of visual, audio and speech features.

Based on Task ordering:

1. **Top-Down**: localize-then-describe
2. **Bottom-Up**: describe-localize-refine

3 Methodology

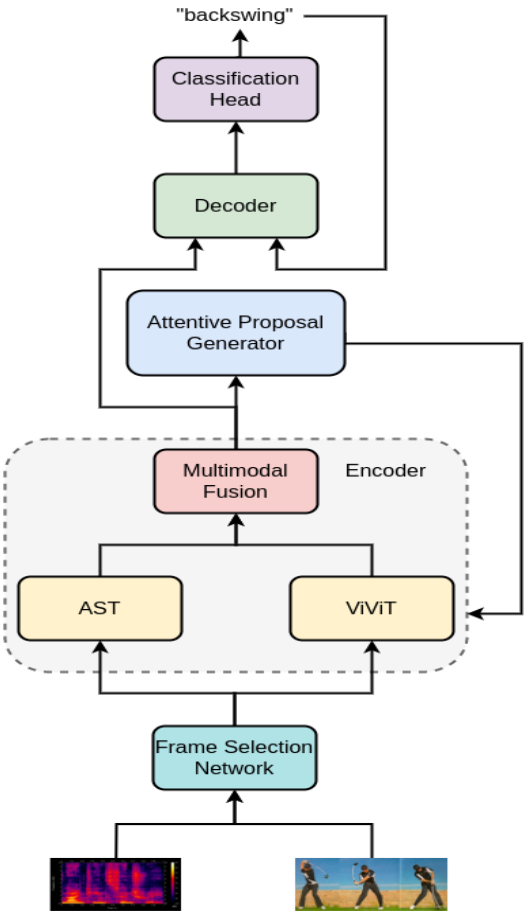

Figure 2: Proposed model for Dense Video Captioning

3.1 Multimodal Feature Extraction

Feature extraction is the backbone of our solution to tackle the task of Dense Video Captioning [1]. A rich feature space would enhance the representational power of the model, thereby leading to more meaningful and accurate event proposals. Most previous works have utilized one modality only (i.e. video) to generate feature vectors as input to the proposal generator. However, audio cues, in conjuction with video is a strong event indicator. These events are accompanied with a sharp change in their corresponding audio spectogram which can be learnt by the model to better determine the precise boundary of events. Thus, we aim to combine features generated using both video and audio. Currently, we are exploring two methods of fusion which include the cross-attention mechanism [5] and common space proejction using contrastive learning [34].

3.1.1 Video Features

Video features are the most important part of the encoder. Without a robust and rich feature space for videos, the proposal generator would never be able to learn accurate event boundaries. Previous state-of-the-art video encoders [35], [36], [37] use CNN-based architectures. Although they have strong inductive bias and translational invariance, they fail to model long-range temporal dependencies which are of paramount importance when encoding videos. Moreover, CNNs require different architectures to model different modalities which can become complex when combining multiple modalities such as video and audio. Transformers [38] can overcome these barriers using its attention mechanism without comprimising on its statistical and computational efficiency. Moreover, transformers can use the same building blocks across different modalities without many modifications. We aim to use the recently proposed ViViT model [39], a purely attention-based video encoder which has outperformed previous approaches across several datasets such as Kinetics 400 and 600, Epic Kitchens 100, Something-Something v2 and Moments in Time. Even though ViViT requires several orders of magnitude more training data as compared to its CNN counterparts, the authors of ViViT propose several methods to limit its training by using pretrained weights in conjunction with strong regularisation and specific fine-tuning.

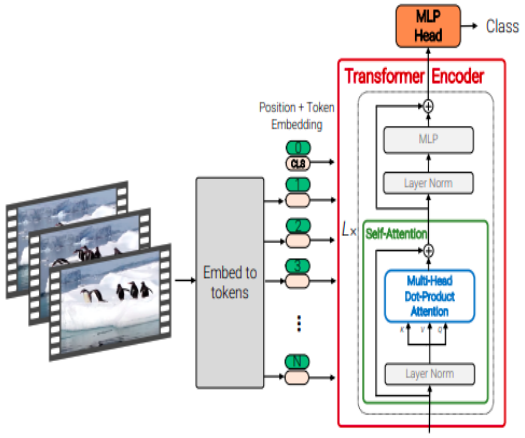

Figure 3: ViViT encoder by Dosovitskiy *et al*. (Image courtesy [39])

3.1.2 Audio Features

Audio is an important aspect of any video. Not only does audio suggest the duration of an event, it can also signify the magnitude or significance of that event. Thus, audio becomes a powerful accessory to images when representing a video. We aim to use the recently proposed Audio Spectogram Transformer (AST) [40] which has outperformed current state-of-the-art models in audio classification. Moreover, AST uses pretrained weights from ViT [41], just like ViViT. We believe that this would thus, work cohesively with ViViT and lead to richer multimodal features.

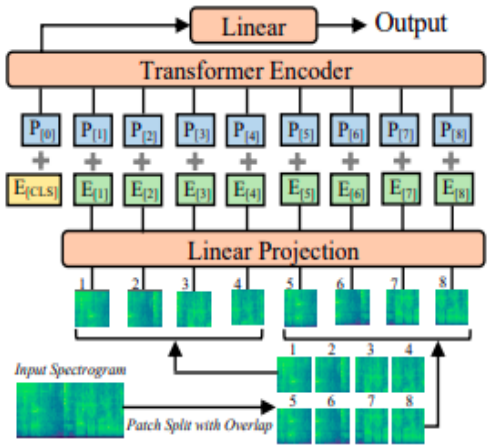

Figure 4: AST encoder by Gong *et al*. (Image courtesy [40])

3.2 Training

3.2.1 Knowledge Distillation

We aim to use the Knowledge Distillation framework (student-teacher paradigm) to train the model, either in individual modules or as a whole. For the video encoder (ViViT), we would use a strong CNN-based video classifier as the teacher model to introduce inductive bias within the transformer and reduce training.Touvron *et al* introduced a method for knowledge distillation [42] using a distillation token to compute the loss based on the softmax generated by the teacher model.

We also aim to train the proposal generator using a student-teacher paradigm with the current SOTA DVC models.

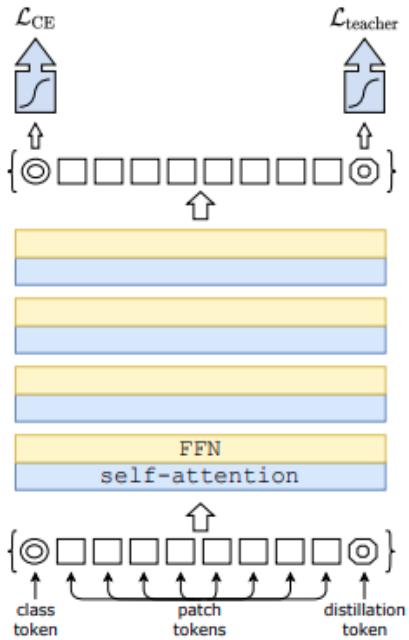

Figure 5: Distillation through attention by Touvran *et al*. (Image courtesy [42])

References

[1] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

[2] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description. *ACM Computing Surveys*, 52(6):1–37, Jan 2020.

[3] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning, 2018.

[4] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning, 2018.

[5] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer, 2020.

[6] Huijuan Xu, Boyang Li, Vasili Ramanishka, Leonid Sigal, and Kate Saenko. Joint event detection and description in continuous video streams, 2018.

[7] Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 234–243, 2021.

[8] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. *arXiv preprint arXiv:1804.00819*, 2018.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation, 2002.

[10] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries, 2004.

[11] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, 2005.

[12] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015.

[13] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances, 2015.

[14] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation, 2016.

[15] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *ECCV*, 2016.

[16] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos, 2017.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science*, page 21–37, 2016.

[19] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions, 2018.

[20] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks, 2017.

[21] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection, 2017.

[22] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. Weakly supervised dense event captioning in videos, 2018.

[23] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning, 2019.

[24] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning, 2019.

[25] Maitreya Suin and A. N. Rajagopalan. An efficient framework for dense video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12039–12046, Apr. 2020.

[26] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning, 2020.

[27] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger, 2016.

[28] Teng Wang, Huicheng Zheng, and Mingjing Yu. Dense-captioning events in videos: Sysu submission to activitynet challenge 2020, 2020.

[29] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator, 2019.

[30] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering, 2020.

[31] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8421–8431, 2021.

[32] Humam Alwassel, Silvio Giancola, and Bernard Ghanem. Tsp: Temporally-sensitive pretraining of video encoders for localization tasks, 2021.

[33] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding, 2021.

[34] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021.

[35] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.

[36] Lorenzo Torresani Du Tran, Heng Wang and Matt Feiszli. Video classification with channel-separated convolutional networks, 2019.

[37] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition, 2017.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[39] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer, 2021.

[40] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[41] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.