Benchmarking Deep Learning based Automatic Scenes Description with User Study.

A thesis submitted in part fulfilment of the degree of
Master of Machine Learning and Computer Vision

by
Siddharth Sachan
U7072580

Supervisor: Prof. Tom Gedeon
Examiner: Dr. Sabrina Caldwell

College of Engineering and Computer Science
the Australian National University

November 2021

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university. To the best of the author's knowledge, it contains no material previously published or written by another person, except where due reference is made in the text.

Siddharth Sachan
9 November 2021

Acknowledgements

Thanks to Prof. Tom Gedeon and Dr. Sabrina Caldwell for being my guides for this project. Thanks to Mr. Yue Yao for discussions, feedback and help in general. Thanks to Mr. Zi Jin for designing the platform used for experiments. Thanks to the participants who took part in the human evaluation of the system. Most of all, thanks to my parents and sister for believing in me and supporting in anyway possible. Also, thanks to Ms. Ada Prova Omera for all her emotional support.

Abstract

Audio descriptions (AD) explain the key events in the videos, making them more accessible for the Blind and Visually Impaired (BVI) people. Automated generation of AD is needed to make the vast amount of videos online accessible to BVI people. In this thesis, we employed a transformer-based Dense Video Captioning (DVC) model to generate the descriptions. To evaluate the descriptions, we have also conducted a user study with 70 participants to evaluate the potential of the generated descriptions to serve as AD. 57.5% of the participants considered the added descriptions to be helpful for the visualisation of the video content when compared to the raw audio. However, due to the limitation of the current state of the art DVC models, the generated sentences are often confusing and grammatically erroneous. Analysing the critical limitations of the current system, we also proposed further research and experimentation opportunities.

Contents

List of Figures

List of Tables

Introduction

According to the white paper[1] by Cisco published in 2020, 81 % of all the IP traffic would be video IP traffic by 2022. Cha (2013) and Lee and Lee (2015) have found that the increasing online video services reduces time spent on cable TV or playing games. Instructional and educational videos have created a whole new industry focused on delivering effective learning at customer demand. Though the impacts of this style of learning are still being researched (Meehan and McCallig (2019)), it is clear that online learning has better reachability and accessibility. Besides the educational content, Lagger et al. (2017) found that the participants used online streaming services multiple times a week and, in some cases, several times a day. All this points to the importance of audio-visual media for sharing information and communication, yet people who are blind or have low vision can not consume video content directly. While the instructional and educational videos are mostly in a commentary style, often accompanied by a transcript, the content on online streaming sites (Netflix, Amazon Prime Videos, YouTube *etc.*) is not. This is to say that there are visual cues that are not accompanied by any corresponding audio cue, which makes it difficult for people with impaired vision to understand the full context.

Audio Descriptions (AD) explain the key audiovisual events in the video, making them more accessible for visually impaired people. Recently all the big streaming services have pushed for adding audio descriptions in their existing titles. According to the AD project by American Council for Blind[2], out of about 26000 titles, Amazon Prime Videos have AD for 3000. In contrast, Netflix has 1700 AD enabled titles out of about 6000 in its portfolio. They also mention the current rates at which AD is being added by these streaming services, which is 50 and 30 titles per month for Amazon Prime Videos and Netflix, respectively. At current rates, it would take several years to add AD to the existing content. This points to the challenge of the sheer amount of video content already present on the web. The amount of video content being added becomes more apparent when we start looking at video sharing sites like YouTube. About 500 hours of video content is uploaded to YouTube every hour[3]. Courtesy of deep learning and automatic speech recognition, YouTube can automatically generate closed captions in English and some European languages, making all of its content more accessible to people with hearing disabilities (Parton (2016)). Hence, the percentage of videos with AD is much less than those with closed captions across all the video sharing and streaming platforms (Ellis (2015)). Moreover, even if the streaming platforms were to add AD to all their content and AD are included in all new productions (TV shows, movies) like closed captions; the vast majority of user-generated content shared online would still be left without AD, making them challenging for blind users to consume. Since there is no automatic system for generating AD, they are usually added after the production and cost about 15-50 USD per minute of video content (Plaza (2017)). Hence,

[1] Cisco White Paper
[2] https://adp.acb.org/
[3] https://expandedramblings.com/index.php/youtube-statistics/

manual generation of AD is not enough to handle the sheer amount of video content being generated and shared on a daily basis.

Schmeidler and Kirchner (2001), in a study with 111 legally blind participants, found that the participants had positive cognitive (correctly interpreting what was shown), psychological (interest in the content shown) and social (level of comfort while talking about the content in their social groups) impacts when watching an animated cartoon with AD as compared to the raw version. Caro (2016) and Walczak and Fryer (2017) have studied the emotions that various degrees of AD can entice in visually impaired (BVI) people when watching award-winning cinema as compared to the standard baseline. AD makes it easier for BVI people to understand the emotions in a scene as they describe the body language and the tension between characters. Hence, BVI people have more confidence in the video content. Snyder (2005), while working on children's picture books, received appreciation from parents of blind children who were able to access the picture books using the more narrative versions of stories. López (2008) points out the role of AD in aiding literacy development and language learning among visually challenged. Moreno and Vermeulen (2015) studied the potential benefits of using AD to learn foreign languages quickly, an application that is not limited to visually impaired people. Perego (2016) concluded in their work that films with AD have no negative impact on the experience of sighted people and can be used as an opportunity for them to mingle with BVI people. All these works show the positive influence that AD has in people's lives, from helping blind children learn the language in a better way to improving adult BVI people's interaction with video content (Greening and Rolph (2007)). Adding AD also addresses the basic need of making video content more accessible for BVI people. The fundamental idea is that everyone should have a similar opportunity to experience videos like movies, TV shows, YouTube videos and education/instructional videos.

The positive impact of AD on BVI people's lives depends on accurate narrations, which in all the studies are manually written and inserted into the video content. Mazur and Chmiel (2012) researched to ascertain the commonalities between AD among different European languages and come up with general guidelines for developing AD. Finbow (2010) studied the various features of AD and their relevance to the story-line of famous films. Hence there exist language-specific styles and guidelines for writing AD for movies and shows. However, there are no well-researched requirements for the automatically generated AD. Wang et al. (2021b) have proposed the use of the three aspects of standard speech recognition v.i.z. When what and how to generate the descriptions. Although the three aspects are pretty much predetermined for automatic speech recognition system, they are not well-defined for an AD generating system:

- *When* should an AD be inserted? The basic idea behind AD is to describe the inaudible but meaningful events in a video. However, it is difficult to generalize in meaningful audible events. For example, a camera zooming in or panning out might not be noteworthy, depending on the context. Scene transitions/cuts might or might not be inaudible (due to the presence of different background audio in two scenes) but affect the flow of the story by quite a bit. Also, AD may be required for the cases when background audio is present but not consistent with the video. The generalization of this consistency is not straightforward as this is more complex than the vanilla audio-visual inconsistency problem (Arandjelovic and Zisserman (2017) where audio from the same and different videos are used to train networks to predict consistency) and requires specifically annotated data.

- *What* should be the content of the AD? In any video frame, there are many audio-visual events happening at once. How does a system determine what is essential and what is not? What

is the most salient way to describe a scene? This is probably the most complex challenge for automatic AD as there is more than one probable ground truth as different people expect/focus on different video details. Stangl et al. (2020) worked with 28 BVI participants to study the variety of content and details users expect when encountering images with AD from different sources and contexts.

- *How* does the generated AD fit with other information? The generated AD is not just to describe what happens in a small portion of a video. It is expected to be consistent with the audio-visual content in the past and future with respect to the clip. Hence, the automatic system needs to generate context-aware audio descriptions, which tell a cohesive story when organized with the dialogue (spoken words), background sounds, and other descriptions. Even when a human is generating AD, they might iterate through the video multiple times to optimize the descriptions to tell the story in a better way. As with other context-based text generation, this task becomes more and more difficult as the length of the whole video or the video clip being described is increased.

Splitting the idea automatic audio descriptions into the above three aspects gives a more precise idea of the requirements and challenges such an automated system must address. They also give us a way to make design decisions when faced with choices, as a better design will be better in one or more of the three aspects. We used the Dense Video Captioning (DVC) task employing deep learning to address the above challenges. DVC model tries to describe a video using multiple sentences (*what* and *how*) with the time-stamps of each sentence (*when*). A more detailed discussion about this is presented in chapter 2. We start with the first work in the field of video captioning and build up to the newer models. Using the state of the art model available for DVC, we generate the descriptions of 19 selected videos. The generated descriptions are inserted into the original videos for their evaluations with users. Chapter 3 summarises the overall pipeline used to go from original videos to user responses. Chapter 4 discusses the quality of the generated descriptions in light of the responses from the user study with 70 participants. The main contribution of our work is the data collection and analysis to evaluate the automatic AD system. We find that although the descriptions might be helpful in the visualisation of the video content, they are often confusing and grammatically erroneous. Finally, Chapter 5 concludes the project by discussing our approach's limitations and potential future research directions.

Literature Review

In this chapter, we provide the relevant background to place our work among the research already done. First, we look at the accessibility of visual media and a recent push to make online images, GIFs and memes more accessible to BVI social media users. Then we examine the existing systems which solve the core task of automatic audio description generation. Then we glance at the different datasets used by the researchers in the field of video captioning. Next, automated evaluation metrics, their importance, meaning and pitfalls are discussed. Finally, we focus on the recent works in the area of automatic audio description generation.

2.1 Accessibility of visual media

While designing web pages or documents, developers will usually put the image descriptions in HTML code's "alt" variable. This alt text (ulterior text) has been used to make visual content on web pages more accessible to BVI people. Hanley et al. (2021) report a set of guidelines used by different institutions for alt text. Gleason et al. (2019a) studied the alt text in the user uploaded twitter content and found that for over 1.09 million tweets with images, only 0.1 % has some description. Simply providing users with the feature to add alt text (which Twitter has as of 2016) does not ensure that descriptions will be added. Guinness et al. (2018) had similar observations when researching consumer and government sites where either the descriptions are largely missing or replaced with the filename of the image.

To address these issues Guinness et al. (2018) used reverse image search to find better descriptions for a given image. In essence, given a query image, the system searches the web/database for similar images with descriptions. Appropriate descriptions are found, added using VoiceOver and rated according to blind and sighted users satisfaction. Wu et al. (2017) developed and tested an automatic alt text generator using object detection to tag the images in question. These tags are fed into a syntax-based sentence generator to produce a set format of descriptions. However, this architecture fails when the additional context is added to the images. Overlay text is has become very common in recent times, with memes and GIFs becoming a part of communication. Gleason et al. (2019b) used classification (for meme template) and text extraction from overlay images to generate a template-based description of memes. However, they only experimented with 11 meme templates, which makes it far from a general solution. Gleason et al. (2020) worked on improving the accessibility of GIFs on Twitter. They combined the alt text and source audio from the GIF to produce audio descriptions which do not overlap with the source audio. Combining the meme and GIFs applications, Low et al. (2019) made a web-based extension to make Twitter more accessible for BLV people. They report a ten times increase in user satisfaction in their study with 50 self-identified blind participants.

Hence a lot of work has been done to make images, memes and GIFs more accessible to BLV people. And recently, researchers have moved away from extracting alt text from HTML code towards

automated systems using reverse image search, captioning based alt text generators and crowdsourcing. The next type of data to consider is user-generated/uploaded videos. Videos are usually longer than GIFs which implies the presence of more complex interactions between people and objects. Then there are scene transitions and cuts which changes the nature of the background sound. Moreover, scenes are interdependent, and what has been shown in the first few seconds may impact what happens much later in the video. This makes the task of making AD for videos much more difficult as compared to more straightforward and shorter memes and GIFs. Next, we explore the research done in making AD for videos.

2.2 Audio Description for Video Content

2.2.1 What is Audio Descriptions? Why is it important?

According to the Audio Description Project (launched in 2009) of American Council of the Blind[1], audio descriptions are "Narration added to the soundtrack to describe important visual details that cannot be understood from the main soundtrack alone. Audio description is a means to inform individuals who are blind or who have low vision about visual content essential for comprehension. Audio description of video provides information about actions, characters, scene changes, on-screen text, and other visual content. Audio description supplements the regular audio track of a program. Audio description is usually added during existing pauses in dialogue." Access is improved through such precise and concise descriptions of people-object interaction, facial expressions, body language and colours. Many researchers have studied the implication of videos with added AD. Ely et al. (2006) found that the information retention improved among children after watching content with AD as compared to the one without AD. Other works have shown positive impacts of AD on increasing understanding (Schmeidler and Kirchner (2001); Simpson and Australia. (1999)), interest and enjoyment (Caro (2016); Schmeidler and Kirchner (2001); Walczak and Fryer (2017)), better social connection (Kim et al. (2014); Schmeidler and Kirchner (2001)) and improved connection between language and visuals (López (2008); Snyder (2005)).

2.2.2 Guidelines for Scripting ADs and their pitfalls

Audio descriptions can be viewed as intermodal translations, i.e. the visual information in the scenes is translated into text information. However, it is not simply a matter of replacing video frames with text; the descriptions have to be cohesive and add to the original audio. To address the complexities of AD, several guidelines are available on the web. Vercauteren (2012) studied the AD generation process from a narration point of view in six European languages. Different countries have their own guidelines advocated by their respective social organizations for blind and visually impaired people (A.1). While these guidelines ensure the quality of the descriptions generated, they are complex, time-intensive to implement and require the manual generation of the scripts. Hence, the guidelines may work for the online streaming sites like Netflix and Amazon Prime Video to generate AD for the shows and movies; they can not be enforced for user-uploaded content. Even if they were enforceable, the variety in the user-uploaded content might make the guidelines too complex to be practical. Moreover, the sheer amount of new user-generated video content being posted online makes any human-based method impractical to scale. We aim to solve this practically using the well-defined computer vision task of the automatic video description. The task is to describe a video clip using one or more sentences.

[1] https://adp.acb.org/ad.html

As with other computer vision tasks, the application of deep learning models has made the pre-deep learning era models obsolete. We first summarize the primary datasets to facilitate the discussion of the models later.

2.3 Datasets

Availability of large annotated datasets is the key to success for most of the deep learning tasks. The size of dataset becomes more important as the complexity of the task is increased. Here we summarize the key characteristics of the datasets used for video descriptions. These can be broadly categorized into three classes, namely *Cooking*, *Movies* and *Videos in the wild*. Some of these have a single sentence description of the whole video, while others have paragraphs describing the videos with timestamps (starting point and end point) for each sentence of the paragraph.

2.3.1 Cooking

1. MP-II Cooking: Rohrbach et al. (2012a) at Max Plank Institute for Informatics released a set of 44 finely annotated cooking videos (o length varying from 3 to 41 minutes) of 12 participants preparing 14 dishes such as fruit salad or cake. The data is recorded from a fixed camera and is annotated with 65 cooking-related activities. The videos were annotated by 6 people with start and end frames as well as the activity categories. In total, the 8 hours of video (at 30 FPS) has 5609 annotations, including the background activity.

2. You-Cook: Das et al. (2013) created a new dataset consisting of 88 cooking videos downloaded from YouTube, which are divided into six cooking styles like baking and grilling. Since the videos are from YouTube, the environment is less constrained as compared to MP-II Cooking (Rohrbach et al. (2012a)) which uses a fixed camera in an inert background environment (same kitchen for all videos). Apart from annotating the videos with natural language, the researchers also annotated the object in the videos, like ten different utensils and bowls. This is to facilitate the natural language generation using object detection as a side task. Amazon Mechanical Turk (AMT) was employed to annotate the videos with 67 words or 8 sentence descriptions on average.

3. TACoS: Textually annotated cooking scenes is a subset of MP-II Composites dataset (Rohrbach et al. (2012b)). TACoS was created by filtering MP-II Composites to include only the videos which include interaction with ingredients and have at least four videos per activity. Hence, TACoS contains 127 videos annotated with 26 activities. For each video, multiple descriptions were collected to improve the size of training data. The annotations are quite action-centric, with 28,292 verb tokens among the total 146,771 annotated tokens. The dataset also provides approximate time stamps for the start and end of each activity. In the same manner, as with other cooking datasets, the descriptions are sequential and non-overlapping.

4. TACoS Multilevel: Rohrbach et al. (2014) released a dataset with AMT workers employed to work directly on the TACoS annotations. They were asked to describe the videos at three levels: a detailed description, a three to five-sentence description, and a single sentence. In the detailed version, all the activities and objects (utensils, tools) are mentioned, requiring the models to learn to recognize both the actions and objects. The dataset also provides the timestamps for start and end times for each sentence of the description.

5. You-Cook-II: Zhou et al. (2018a) released an upgrade to the You-Cook dataset, containing over 2000 videos downloaded from YouTube, making 89 recipes. Although they offer more challenges similar to 'open domain' videos, they are still restricted to cooking activities, so they are not fully generalized. The whole dataset has 175.6 hours of video with 7.7 segments on average. The vocabulary is 2600 words. The length of the videos goes up to 10 minutes, with the average size being around 5 minutes. The researchers have provided official splits 67%:23%:10% for training validation and testing, taking the recipes into account.

2.3.2 Movies

1. MP-II MD: Rohrbach et al. (2015) from Max Planck Institute for Informatics released a movie description dataset containing audio descriptions extracted from 94 movies. The dataset includes 68,375 clips (hence sentences) with an average length of 3.9 seconds, amounting to a total length of 73.6 hours. The AD was extracted by processing the audio stream with added AD with the original audio of the movies. Since the audio descriptions might refer to contextual information, the descriptions were cleaned, keeping this in mind to remove references to objects/people not present in the clip.

2. M-VAD: Based on Descriptive Video Service (DVS), Torabi et al. (2015) released the Montreal Video annotation dataset containing 46,589 annotated clips from 92 movies. DVS tracks are put carefully in the DVDs released much after the movie's original release. The average length of the clips in the dataset is 6.2 seconds, amounting to a total of 84.6 hours of running time. The vocabulary of the dataset contains 9,512 nouns out of the 17,609 unique words, which is common with the datasets based on movies where characters have names and pronouns are used less (to avoid confusion for BVI people).

Rohrbach et al. (2017) combined both the datasets to introduce the Large Scale Movie Description Challenge[2] (LSMDC) which contains a corpus of 118,114 sentences and video clips from 202 movies.

2.3.3 Videos in the Wild

1. MSVD: Microsoft Video Description dataset contains 1970 video snippets downloaded from YouTube (Chen and Dolan (2011)). Video clips are filtered to have a single activity or only one main focus with length varying from 10 seconds to 25 seconds. Moreover, they are muted, and any overlay text is removed to reduce any bias in the descriptions. Each clip has 41 single sentence descriptions on average, with some in different languages than the usual English (like Mandarin and German). The standard split for the dataset is 1200 for training, 100 for validation and 670 for testing.

2. MSR-VTT: MSR Video to Text contains about 10,000 video clips, each annotated with approximately 20 sentences (i.e. different descriptions, similar to MSVD, Chen and Dolan (2011)) each, giving 200k sentence-clip pairs in the whole dataset. The videos are collected by using 257 popular queries on a commercial video sharing platform and 118 videos for each query. Different to the MSVD (Chen and Dolan (2011)), the researchers left the audio channel as it is, and hence models using multimodal information can also be employed here.

[2] https://sites.google.com/site/describingmovies/

3. ActivityNet Captions: Krishna et al. (2017) released a large dataset containing approximately 100k annotated sentences for 20k videos from the Activity Net (Heilbron et al. (2015)) dataset. Since the Activity Net dataset contains videos from YouTube, this dataset has various backgrounds, quality, types of content etc. Moreover, since the Activity Net dataset was made for activity classification, most of the videos correspond to some activity. Each video contains 3.65 clips on average, and about 10% of them have overlapping time-stamps. The start and end time-stamps are provided for each annotation. The Activity Net Challenge[3] page also provides pre-extracted features for the videos. The standard split provided by the researchers and used to report results is 50%:25%:25%.

2.4 Automatic Video Captioning and Description

In computer vision, the task of video captioning refers to describing the video clip using a single natural language sentence, based on the premise that a short clip contains one main idea (Aafaq et al. (2019)). This can be thought of as an extension of an image captioning problem, where the features from all the frames in the video are combined to give a single feature that is used to generate the caption. Hence a lot of initial work on single sentence captioning is inspired by the work in image captioning, with models being generalized to handle the time component as well (Amirian et al. (2020)). The multiple sentence description can be thought of as generating a description of each event in the video and then combining the descriptions to make sense together. Before the deep learning models took over the world of computer vision, the classical models employed a two-step strategy, i.e. first, detect the entities (objects/persons, action/verbs, scene/background) and then fit them into fixed sentence templates. Many deep learning models follow a similar idea, with the detection of entities becoming a feature extraction and encoding and the fixed sentence templates replaced by a sequence generation model. Before jumping into deep learning models, which are more closely related to our work, we briefly mention the classical models.

2.4.1 Classical Models:- Before Deep Leaning

Research to use natural language to describe visual information goes back three decades. Koller et al. (1991) designed a system to track vehicles across multiple frames (from traffic camera footage) and characterise the motion using natural language tags. Later, Brand (1997) developed heuristic and probabilistic systems to process an instructional video (with action verbs like grab, put, press, touch etc.) into a storyboard. Specifically for video description, SVO (Subject, Verb, Object) methods are the most famous classical models. SVO tuple tagging methods have two stages, namely content identification and sentence generation. In content identification, the model identifies the actor/person (noun, subject) doing the activity (verb) using or interacting with object/tools (object). Then grammar rules are used to generate template based sentences. This is also common for the systems which already have tags generated from somewhere else. Wu et al. (2017) developed a system that works in this fashion by using the automatically generated tags for Facebook images to create alt text.

Several feature extraction and matching methods have been developed over the years in classical computer vision research. These methods employ the so-called handcrafted features for the first step. One of the very successful human detection methods is the Histogram of oriented gradients (HOG) and used SVM to classify the histograms (Dalal and Triggs (2005)). For activity recognition, features

[3]http://activity-net.org/challenges/2021/

needed to be in space (somewhere on a particular frame) and time (across multiple continuous frames). Some of the approaches used are Bayesian Networks (Hongeng et al. (2000)), Past, now future (PNF) Network (Pinhanez and Bobick (1998)), and Histogram of Oriented Optical Flow (HOOF, Ustundag and Unel (2014)). Similarly for object detection several well know feature extraction methods are established like Scale Invariant Feature Transform (SIFT, Lowe (1999)), HAAR features (Viola and Jones (2001)) and Deformable parts model (DPM, Felzenszwalb et al. (2008)). Once the entities are detected in the first stage, the sentences are generated using fixed templates. Many systematic grammar frameworks including Head-driven phrase structure grammar (HPSG, Pollard and Sag (1994)), planner and surface realizer (Reiter and Dale (1997)) give the rules to be followed to make sentences from given SVO tuples (Bateman (1997)). These templates contains placeholders which are filled by the proper entities based on their class.

We have described the general framework used in classical computer vision to solve video captioning. We do not survey the specific methods as they are unrelated to the work done here. More keen readers are referred to Fig. 5 of Aafaq et al. (2019) and Table 1 of Li et al. (2019b) for a detailed list of models using the SVO framework. As the datasets available at this time were quite limited in terms of variety and size, these models worked well. However, as more general datasets like YouCook (Das et al. (2013)) were introduced, the limitation of these models became clear. They were not equipped to handle large variety of backgrounds, people, objects and activities. Around the same time Alex Net (Krizhevsky et al. (2012)) showed the power of deep leaning models. Within a few years, image captioning models using deep learning (Devlin et al. (2015); Mao et al. (2014)) showed much better performance than the classical methods. A similar trend was observed for video captioning and description task as well. We discuss the models in the next part.

2.4.2 Deep Learning Models

A general approach while solving the problem using deep learning can be broadly divided into the following steps:

1. Visual Content extraction: Extract and encode the video features
2. Text Generation: Decode to sequentially generate words

Most of the models can be dissected in terms of theses two steps.

2.4.2.1 Single sentence descriptions (Video captioning)

The early research in video description focused on generating a single sentence to describe a short clip where one activity is focused upon. CNN (LeCun et al. (1989)) and a variety of RNN architectures (Hochreiter and Schmidhuber (1997)) are used for feature extraction and encoding to learn useful representation for the second stage. Text generation is usually done using some flavour of RNN architecture like LSTM, GRU, Bi-directional RNN or multilayered RNN (deep RNN). Hence, the overall architecture is similar to the encoder-decoder framework used for other sequences to sequence modelling tasks like machine translation (Dabre et al. (2020)). Let us look at some of the models in a little bit more detail.

Donahue et al. (2015) were one of the first ones to attempt to solve video captioning using deep neural networks. They extract the frame-wise features using CNNs pretrained for the ILSVRC-2012 challenge (Russakovsky et al. (2015)). A conditional Random Field (CRF) is used to predict the sequence of subject, object, actions, and background using the features. This sequence is fed into

a vanilla LSTM to generate a proper sentence. Although the model outperforms the baseline work using CRF without LSTM (Rohrbach et al. (2013)), it is not end to end trainable because of the supervised CRF modelling in between the encoder and decoder stages. Venugopalan et al. (2015b) addressed this by mean pooling the CNN features from the video. The CNN architecture is similar to Alex-Net (Krizhevsky et al. (2012)) pretrained on the 1.2 million images from the ILSVRC-2012 challenge. They also used a two-layered LSTM model. Mean pooling the features reduces the problem to image captioning, where the image features are replaced by the mean feature of the video clip. However, simple averaging losses the temporal information, i.e. even if the video is presented in reversed (or some random order for the matter) order of frames to the model, it will generate the same representation. This limits the application of the method to short clips where one significant activity is happening (single subject/object/action).

Open-domain videos have a variety of complexities where multiple subject-object interactions take place. Mean pooling the frame features may lead to a lot of clutter and make the generated text inadequate as proper descriptions. Working upon the success of C3D model in activity classification from videos (Tran et al. (2015)), Yao et al. (2015) used 3D CNN to encode the Spatio-temporal features from the video clips. They first extracted the classical HOG, HOOF features (Konečnỳ and Hagara (2014), to reduce the dimensionality. These features carry the flow information and are put in a fixed size 3D tensor which is processed by 3D convolutions. The extracted feature from 3D CNN is concatenated with frame-wise features extracted using a pretrained (on activity classification) GoogLeNet (Szegedy et al. (2015)). Along with the usual LSTM to generate the text, they also used an attention mechanism to perform a weighted sum of the videos features at each time step, depending on the hidden state of the LSTM at the previous time step. Although they reported a good performance on less restricted datasets like M-VAD (Torabi et al. (2015)), the architecture is only equipped to handle videos of a fixed size of 240 frames. Due to this fixed upper limit, the model ignores any information beyond a certain point.

Venugopalan et al. (2015a) solved this issue by using a two layered LSTM to both encode frame wise features and decode text. They trained two separate networks, one to predict text using the frame wise features and another using the flow features. The predictions from the two networks are combined through various process to get final prediction of the word at a time step. In this way, this is a first model which take into account both the visual and temporal information, is end to end trainable and puts no limits on the length of the videos. Several modifications like use of attention while decoding, embedding sentence and video features into same space (Pan et al. (2016)) and likelihood of generated sentence for a pretrained language model (Venugopalan et al. (2016)) were done on this basic idea. This has lead to works like Gao et al. (2017), who have used multiple such fine-turning techniques to achieve very good performance on the open datasets like MSVD (Chen and Dolan (2011)). However, videos in the wild often contain more than one even, and describing multiple events with one sentence is infeasible because it leads to complex sentence structures, which are difficult to learn. This has lead to work in describing sentences using multiple sentences and using datasets like TACoS Multilevel (Rohrbach et al. (2014)), You-Cook-II (Zhou et al. (2018a)) and ActivityNet Captions (Krishna et al. (2017)) for evaluation.

2.4.2.2 Multi Sentence descriptions (Video description) and Dense Video Captioning

Yu et al. (2016) proposed a hierarchical RNN architecture to generate paragraph descriptions of the videos. Video features extracted by CNN networks like VGG (Simonyan and Zisserman (2015)) and

C3D (Yao et al. (2015)) are used to generate sentences using a GRU. The hidden states during a sentence are averaged and concatenated with the last hidden state to form the sentence encoding. This works as the input for the sentence level LSTM, which produces the initial hidden state of the sentence generating GRU for the following sentence. Xiong et al. (2018) proposed the use of reinforcement learning to generate sentences through their Move forward and Tell model. Recently, Park et al. (2019) used a hybrid discriminator for generating sentences using an adversarial approach. The sentences are generated using a LSTM whose first hidden state is simply the last hidden state of the previous sentence. This generator is trained using three discriminators each for the language characteristic of the sentence, relevance to video clip and coherence w.r.t. the previous generated sentence. Although these works can generate multiple sentence descriptions of a given, most of them either need the ground truth time-stamps of the different clips or model the temporal order of the descriptions, which is not evaluated. This is not directly applicable to our task of automatic audio description, as we need the model to predict the time-stamps of the sentences to be inserted.

Krishna et al. (2017) introduce the new task of dense video captioning (DVC), which involves both the prediction of start-end timestamps of localized clips in the video and the respective sentences. 3D convolution (C3D architecture, Yao et al. (2015)) is used to extract video features. These features are sampled at different intervals and fed into different LSTMs, each of which tries to predict start and end probabilities for all of their respective time-steps. This event proposal network is connected to a vanilla LSTM sentence generator that considers the attention-based encoding of the past and future frames in its input. Wang et al. (2018) extended the work by adding a backward pass LSTM, in addition to a forward pass to encode video features in the backward direction as well. These backward encodings serve a similar purpose as the representation of future frames that the baseline paper (Krishna et al. (2017)) used. Probabilities of events for a set of fixed time windows is predicted by the event proposal network using encodings from both the forward and backward LSTM. The sentence decoder is a LSTM that uses the encoding from forward (past context) and backward (future context) LSTM, along with the visual features in the predicted time window as inputs. One big issue with these methods is their event proposal networks. They produce thousands of proposals per video clip (average number of annotations in 3-10 across different datasets for DVC) and require a manual threshold to select a few to work in an inference setting. This makes them unsuitable for direct application in the automatic audio description where the model should predict a reasonable number of captions without external thresholding.

Duan et al. (2018) proposed a model which can produce captions without requiring the time-stamp annotations. They achieved this by training a sentence localizer, which is a function to map a given sentence to a time window in the video. They first pretrained a sentence generator that fed some random time windows from the video and generated captions. These captions are fed into the sentence localizer to predict the time window the caption came from. This time window can be used to generate the captions for the next iteration. This sort of cyclic training alleviates the need for time-stamp annotations. Although their model performs worse than the Bi-SST model (Wang et al. (2018)), they remove the need for external thresholding. During inference, 10-20 random time windows are initialized, and the model converges to the required number of windows on its own within two iterations. Mun et al. (2019) added an additional event sequence generator between the event proposal and caption generator to streamline the process of DVC. The event sequence generator employs attention-based RNN to dynamically select a few events from the thousands proposed by the event proposal network. The sequence generator selects 2.85 on average when tested with ActivityNet Captions (Krishna et al. (2017)), which has 3.65 sentences per video on average. They also used

hierarchical RNNs as caption generator, like Yu et al. (2016), replacing the simple attention-based RNN common as sentence decoder.

Recently with the success of transformer (Vaswani et al. (2017)) based model in language translation (another sequence to sequence modelling task, Karita et al. (2019)), researchers have tried to use transformers for DVC. Zhou et al. (2018b) used multi-head transformer architecture to both encode video features and generate the sentences. The context-aware video features given by the encoder are put side by side and processed through temporal convolutions with different kernel sizes. The kernel in temporal convolution is of the same size as features in one dimension, and the other dimension is varied to cover large or small number of frame features (i.e. time). Result from each convolution is used to propose events. The encoded features are masked using a differentiable function for a given proposal and put through the decoding transformer. They were the first researchers to show the application of transformer architecture for DVC. Iashin and Rahtu (2020) proposed the use of audio information already present in the video to enhance the descriptions in their Bi-modal transformer model. Both their encoder and decoder transformers have separate heads for audio features and also cross attention between the audio and the video features.

Although these models perform well on the challenging ActivityNet Captions dataset, they also propose too many events and use external thresholding to select the top thousand proposals used for evaluation. This is a similar problem as we have seen in the RNN based architectures, making these not applicable to our goal of automatic audio description. Also, the time complexity of self-attention in a naive transformer structure is $O(n^2d + d^2n)$, where n is the number of terms in the sequence and d is the dimension of the feature vector (Vaswani et al. (2017)). With videos, the number of frames in the sequence is usually very large and slows the training. Wang et al. (2021a) used deformable transformers which predict the range of keys that a query needs to attend to. This reduces the number of dot products required during the attention process. They also use an additional head to predict the number of events in the video besides the temporal localization and caption generation head. The decoder performs these three tasks in parallel for some number (N) of learnt queries. The query vector indirectly tells the decoder which part of the video is being focused upon. Each of the N temporal segments and the captions gets a confidence score used to sort these proposals. The top-k among these are selected during inference, where k is also predicted by one of the decoder heads. Hence, they have shown that the transformer model can be used to get state of the art performance on the ActivityNet Captions dataset. Moreover, the internal prediction of the number of events makes this model directly applicable to our automatically generating a reasonable number of video descriptions.

2.5 Evaluation of Video description and DVC

Evaluating models is an essential part of development and research. Evaluation of video descriptions, machine-generated or annotated by humans, is not a straightforward task as there is no specific ground truth. As the length of videos increases, multiple things might be going on. One human annotator might focus on a different activity than the others. This makes it more difficult to evaluate than other sequences to sequence transformation tasks like machine translation, sign language translation or lip-reading where single ground truth is applicable. Researchers have tried to include this property in the datasets, like videos in the validation set of ActivityNet Captions (Krishna et al. (2017)) have two sets of annotations. Evaluation of machine-generated captions/descriptions can be divided into Automatic and Human. We discuss these metrics and methods in brief detail here.

2.5.1 Automatic Evaluation

Even though there are additional difficulties in evaluating video descriptions, most of the metrics used to evaluate the descriptions are derived from the evaluations metrics for machine translation and text generation tasks. We only mention the core ideas and behaviour of the metrics, leaving the exact empirical formulation to the original works.

1. BiLingual Evaluation Understudy (BLEU@N): BLEU scores depend upon the number of matching *n-gram* in the candidate sentence w.r.t. a reference sentence (Papineni et al. (2002)). Hence, this is a precision metric. The precision is normalized by the total number of *n-gram* (essentially length) in the candidate sentence to prevent longer sentences from scoring high by just repeating some phrase. There is also a brevity penalty for sentences smaller than the reference sentence. BLEU score for a perfect match is 1. It was designed and tested with a corpus of reference sentences, i.e. a candidate text is matched with a bunch of references, and the average scores show correlations with human judgements. Hence, it might not be suitable to use BLEU when only one reference description is available (most of the datasets in video description have one annotation).

2. Recall-Oriented Understudy for Gisting Evaluation (ROUGE): Lin and Och (2004) presented several recall-based statistical methods to evaluate summarised text. ROUGE_L, one of the versions, correlated (Pearson's ρ correlation) the best with human judgements w.r.t. adequacy and fluency. Hence, when DVC works mention ROUGE, they usually refer to ROUGE_L. It works by finding the longest common subsequence (LCS) between the candidate and reference sentences. The recall and precision is calculated w.r.t. this subsequence and the lengths of the candidate and reference sentence, respectively. Then the F-score calculated using these recall and precision is the ROUGE_L score. ROUGE_L score is 1 if the two sequences are exactly the same and 0 if they have no matches.

3. Metric for Evaluation of Translation with Explicit ORdering (METEOR): Banerjee and Lavie (2005) introduced a framework for evaluation that uses WordNet (Miller (1998)) database to account for synonyms and stemming. Before score calculation, the two sequences are aligned first so that the adjacent words in the reference lie in the same order in the candidate. When the reference and candidate are not exact matches, there will be some crossovers during the alignment process. More such crossover implies a smaller METEOR score.Elliott and Keller (2014) reported that METEOR corresponds better with human judgements as compared to BLEU and ROUGE using Spearman's correlation.

4. Consensus-based Image Description Evaluation (CIDEr): Vedantam et al. (2015) developed an evaluation system based on the cosine-similarity between the candidate sentence and corpus of reference sentences for image captioning. Any given sentence is converted into a vector representation using the TF-IDF (term frequency-inverse document frequency) score. The final CIDEr score is the weighted sum of all the *n-gram* TF-IDF scores. The method was designed for image captioning and tested for datasets that have several ground truth annotations per image. Their experiments show the best correlation between human judgements and the CIDEr score when 50 annotations were used per image. Hence, this correlation might not hold well when only one annotation is present (as with most datasets for DVC).

All these metrics were designed for comparing single sentence annotations with single sentences generated by machines. In DVC, each video is described by a set of sentences that are highly interdependent. Moreover, Zhou et al. (2018b) noticed many repeated phrases and redundancies between different captions generated on the same video. We do not want a caption generator that produces five captions (different time windows), with each of them using similar phrases again and again. The metrics designed for the sequence to sequence translation fail to capture this idea, as evaluation is done on a sentence level. A possible way to circumvent this issue is to use these metrics for paragraph-level evaluations. However, that would have to be tested with experiments to see if there is a correlation with human judgements at the paragraph level. In any case, some recent works have introduced ideas to address these issues:

1. Self-BLEU: Zhu et al. (2018) presented the idea of using BLEU scores for sentences generated by the same system to gauge the diversity. Since, BLEU score between two sentences tells the similarity between them, once can use the BLEU score of a sentence with rest of the descriptions generated for the same video to judge the diversity of the generated descriptions. Lower Self-BLEU score shows more diversity.

2. Story Oriented Dense video cAptioning evaluation framework (SODA): Fujita et al. (2020) pointed out several deficiencies of the existing evaluation frameworks like BLEU, METEOR and ROUGE$_L$ for dense video captioning tasks. The core shortcoming of the existing frameworks is how they match the generated captions to the ground truth based on the IoU (Intersection over Union) score for the respective time windows of the two sentences. They address these issues by assigning correspondences using dynamic programming to maximize the overall sum of IoU. The average F_1 score over the whole dataset is the SODA score. However, as this was proposed very recently, only a few works (like Wang et al. (2021a)) have reported SODA scores.

Automatic evaluation of text is still a very much open research topic. Video descriptions pose additional challenges as it does not have specific ground truth. DVC task adds one more layer of complexity by having multiple sentence descriptions. As DVC was introduced recently (in 2017), we expect new and better automatic evaluation frameworks, like SODA, to be constructed aiming at this problem specifically. For this work, we use METEOR scores as the best representation of automatic evaluations for DVC.

2.5.2 Human Evaluation

Several shortcomings of the automatic evaluation metrics have been reported in reference to the machine translation task, for which they were originally designed (Li et al. (2019a)). Moreover, the application of these metrics to multi-sentence descriptions (with time windows also playing a role) has not been extensively researched. Post (2018) also found the discrepancies in the BLEU scores reported by the researchers due to failure to mention the hyperparameters (like smoothing method for zero matches) used during evaluation. Finally, most of the datasets for video description are annotated by visually corrected (corrected visual acuity) AMT workers. M-VAD (Torabi et al. (2015)) and MP-II MD (Rohrbach et al. (2015)) are exceptions to this as they were constructed by extracting the audio descriptions from the movies. However, these were further cleaned by people to remove instances of things not happening in the clips. Overall, the datasets currently are annotated according to visually corrected people and hence do not represent the requirements of BVI people. Considering this, the recent works focusing on alt-text generation for images (Gleason et al. (2020, 2019b); Low et al. (2019);

Wu et al. (2017)) include user studies with both visually corrected and BVI participants. These studies usually try to determine

- whether the BVI participants found the automatically generated descriptions helpful
- whether both types of participants found the descriptions to be grammatically correct
- whether the visually corrected found the descriptions appropriate w.r.t. the visual information

Due to the implicit nature of a user study, human evaluation is subjective. Different people might want the descriptions to focus on different aspects of the video. If a description satisfies the needs of one person, it might sound good enough to them. The same description might be insufficient for others. Understanding this underlying nature, we look for trends in these studies and judge the implication knowing the subjectivity of the evaluation method.

2.6 Automatic Audio Description

Automatically generating audio description which are directly deployable, is still far from reality. It is a complex and nuanced task. Even the most recent methods depend upon either the metadata present in the video (Kurihara et al. (2019)) or extracting non-dialogue lines from an available script (Campos et al. (2020)). These script based methods are clearly only applicable cinema based videos where a pre-made script exists and was followed to make the content of the movie. Meta data present in the video (could even include automatically generated tags like sport, health, home etc.) will often be insufficient to describe the activity happening in the video. Wang et al. (2021b) used video description task as the backbone for their Automatic AD generation. Their system has following three parts:

1. For detecting where the captions should be inserted: They first extract all the parts of the video which have no background audio (like people saying something). The extracted clips are passed through a pretrained audio-visual consistency network (Arandjelovic and Zisserman (2017)) to decide if the clip audio and video information match or not. If they do not match the clip is send for captioning to the second part.

2. For generating captions for each selected clip: They used a pretrained dense video captioning system (Duan et al. (2018)) to generate descriptions of each clip.

3. For selecting a set of captions: Given all the clips and their captions, this module selects the best possible set based on a hybrid loss function (perplexity, diversity and irrelevance costs). Dynamic programming is used to select the set with minimum loss.

The system was employed to generate captions for a set of videos in the ActivityNet Captions dataset (Krishna et al. (2017)). They conducted a user study with 32 participant out of which 12 were partially of fully blind. They reported positive results, with decrease in the number of times BVI people asked for additional information. They also reported that the users feel that the information provided by caption is not accurate. However, we identify the following potential issues:

1. The audio-visual consistency network used to identify clips which need captions is trained by putting false audio in original videos to create negative examples. So, for example if there is lawn mower in the scene, and it makes sound of engine, the audio-visual consistency will say that this does not require description. Hence there is a potential to miss some clips which actually need description.

2. The DVC system used has been trained to produce multiple sentence descriptions of the given video. Using it to generate single sentence might not optimal.

3. Moreover, networks trained for DVC usually generate sentences which are inter-related i.e. they know the context. Due to breaking of clips and pushing them separately through the model, the context information is lost. We observe it's impact in the AD added videos released by the researchers[4]. For longer videos which have multiple AD added, the AD is very redundant and repetitive (for example not using pronouns at all).

4. After looking at the video IDs from the released set, we observed that all these videos corresponded to the training set of the ActivityNet Captions dataset. Although, evaluating video description is not the goal, using videos from the validation or test set might be more representative of the achievable performance.

As a whole, this an interesting work which inspired us to make our own system and experiments.

2.7 Implications of using Deep Learning

Failure modes of any automatic system have to be studied and analyzed to understand the reasons behind the errors. Deep learning methodologies, although produce great results, are much more obscure than a rule based or template based system. Finding the reasons behind the network not producing correct results is not straight forward at all. Moreover, being a data dependent methods, neural networks simply learn the relationships they are presented with. For example - consider all the gym videos present in a dataset. If 90 % of the gym videos show men doing exercise and are annotated with the pronoun *he* somewhere in the description, the model might learn to associate gym to *he*. This may lead to gender mis-identification during inference. Similar issues were raised in a panel discussion in CVPR 2020[5], where blind panelist noted that all the errors generated by an automatic system are not equal. Errors made w.r.t. sensitive characteristics for minoritized community (like mis-identifying genders) can be very harmful. Wu et al. (2017) also noted that there has not been much research into understanding the impacts of these errors for the BVI people.

We should keep in mind that the biases present in the data are guaranteed to show up during inference. Sometimes we can use weighted presentation during training, where rare examples are shown to the model more times than the abundant one. For DVC datasets, it is not straightforward to remove these biases. Addressing the explainability and biases of deep leaning models and datasets respectively is an active research area in the community. All in all, we have to mindful of the results we get and understand that the model can fail.

[4] https://bitwangyujia.github.io/research/project/ad_demo.html
[5] CVPR 2020 VizWiz Grand Challenge Workshop – Panel Discussion with Blind Technology Experts

Methodology

The overall design of our approach is heavily inspired from the recent work of Wang et al. (2021b). The main difference is in the way we generate the audio descriptions. Then the generated descriptions are inserted in the original video to get the modified videos. These modified videos and their audio tracks are used in a user study where participants are asked to judge the descriptions. This data is collected for a variety of videos in the AcitivityNet Captions (Krishna et al. (2017)) dataset. The whole pipeline to get the experiment data is discussed in detail in this chapter. Finally, we use simple statistics to analyse the survey data and present our results in the next chapter.

3.1 Video Description Generation

While Wang et al. (2021b) used three sub modules to generated captions, we employ the dense captioning module directly. In the DVC datasets the annotations are added by AMT workers at the places where they see some activity going on. If the DVC model can predict these time-stamps correctly, then it is indirectly capable of recognizing and localizing activities in the videos. We think this localization is better than the use audio-visual consistency network trained on cross-sampled negative examples. For example the audio-visual consistency network may find the audio cues consistent with the visual information if a parade is going on in a video, but we would still want to generate a description. These activities are more likely to be covered by directly using the dense captioning module to predict the time-stamps as well.

In the second stage, Wang et al. (2021b) use a DVC model trained on the AcitivityNet Captions (Krishna et al. (2017)) dataset. However, the original model is trained to produce multiple sentence descriptions for a given video. Here the researchers used it to produce one sentence caption for the clips suggested by their first stage. Hence the DVC model is used in an untested way. A single sentence captioning model might be more appropriate if we already have the clip time-stamps. Moreover, captioning all the segments separately removes the context. For example, if the first stage recommends two time windows and we have generated the captions for the first one, the captions generated using the second time window are completely ignorant of what was said in the first caption. This might lead to repeated phrases (like describing the background again and again for each caption) and redundancies in the captions. Although they have the third stage tries to reduce these repetitions, we still see the impacts in the data released by the researchers[1]. Hence, the captions generated by external division of the video into clips produces captions with recurring phrases, which reduces the clarity of the whole description. By employing a DVC model that produces context aware descriptions, we remove this problem. We use the Parallel Decoding Video Captioning (PDVC, Wang et al. (2021a)) model for generating the descriptions. PDVC was chosen because of it's state of the art METEOR score on the validation set of AcitivityNet Captions (Krishna et al. (2017)) and You-Cook-II (Zhou et al. (2018a))

[1] https://bitwangyujia.github.io/research/project/addemo.html

datasets. Moreover, we choose to work with AcitivityNet Captions dataset as it has more variety of scenes and activities as compared to You-Cook-II, which is a cooking dataset.

We summarize the key steps for using the PDVC model:

1. *Feature extraction method*: In DVC research, the main focus of the models has been time-window prediction and caption generation. Models trained for activity recognition task (Heilbron et al. (2015)) are employed to extract the video features and used across different models. 3 D Convolution (Tran et al. (2015)) and two stream convolution networks (TSN) (Simonyan and Zisserman (2014)) are the most famous architectures used in Action recognition. In almost all of the DVC literature (Mun et al. (2019); Wang et al. (2021a); Zhou et al. (2018b)), the METEOR scores reported with TSN features are higher than those with C3D features. Hence, we also choose the TSN features extracted following the modifications presented by Wang et al. (2017). These are made available by the publishers of the Masked Transformer model (Zhou et al. (2018b)) in their official implementation[2].

2. *Training and Inference*: We followed the instructions given in the official implementation[3] of PDVC model to train the model and generate descriptions for the videos in the validation set. It takes about 20 hours to train for 30 epochs on single GPU.

Running inference on the validation set gives a dictionary where video IDs can be used to get the video duration, predicted time-stamps and the corresponding sentences.

3.2 Inserting the Video Descriptions as AD

Since the video IDs for ActivityNet Captions are the YouTube watch IDs of the videos, these are readily available to watch and download. We use pytube[4], a python library to down the videos from YouTube at their highest resolution. These are referred to as the original videos for the rest of the text. For a given video, the captions are:

1. converted to audio files using gTTS[5] (Google Text-to-Speech), a python library to interface with Google Translate's text-to-speech API. These are stored as mp3 files for further processing.

2. Each caption also has the predicted start and end time. We choose to insert the generated audio at the start time of the respective caption.

3. To insert multiple audio file into a single video, we used FFmpeg[6]. FFmpeg is a powerful open source command line tool to decode, encode, transcode, mux, demux, stream, filter and play multimedia files across different formats. We use the adelay filter to add the audio files at given delays (start time predicated by the PDVC model).

Since FFmpeg requires command-line instructions, we made a python script to automatically create a bash file containing all the commands for a given set of videos. Running this bash file gives us the modified videos. A modified video has the predicted captions inserted as audio beginning at the predicted start timestamp of the caption. Subsequently, we use FFmpeg to extract the audio in the

[2]https://github.com/LuoweiZhou/densecap
[3]https://github.com/ttengwang/PDVC
[4]https://pytube.io/en/latest/
[5]https://pypi.org/project/gTTS/
[6]https://www.ffmpeg.org/

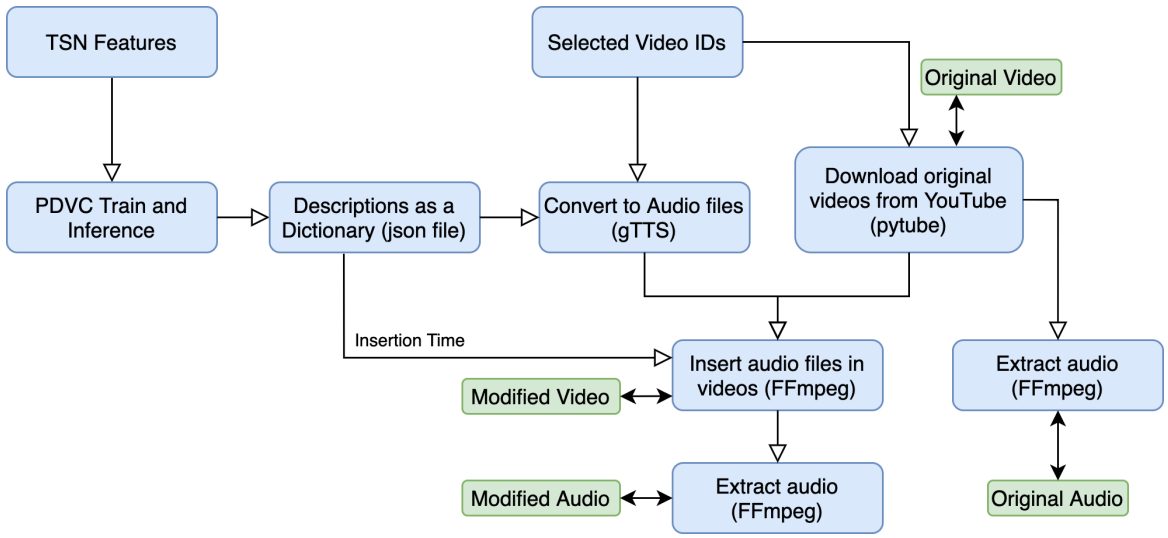

Figure 3.1: The pipeline to get the four variants of the media used for human evaluation. PDVC (Wang et al. (2021a)) is used to generate the descriptions which are converted to audio using gTTS. The audio files are overlayed at predicted positions in the video using FFmpeg. The selected Video IDs are given in Table B.1.

original videos and the modified video to get the original audio and modified audio. Hence, each video ID now has four variants in which the information is presented to the participants. The pipeline to get the four variants is summarized in Fig. 3.1.

3.3 Design of the Experiment with participants

As noted earlier we chose to work with ActivityNet Captions (Krishna et al. (2017)) owing to the diversity of videos the dataset has. Since the dataset contains about 20,000 videos (50:25:25 split), a certain number of representative videos have to be selected for the user study. Unlike Wang et al. (2021b), who chose videos from the training split, we decided to go with the videos from the validation split. Since the dataset is essentially open, there is a wide variety of videos and it is difficult to categorise the selected videos. Anyway we provide the list of the 19 videos used in Table B.1, which include sports clips (gymnasium, basketball and weightlifting), instructional videos, advertisements, music and other miscellaneous activities. This demonstrates the diversity among the selected videos.

Due to Covid-19 restrictions, the access to physical spaces was very limited and the experiments were conducted completely online. We used an in house platform[7] to conduct the survey. The participants were recruited through the Psychology Research Participation Scheme[8] of the Research school of Psychology at the Australian National University. They are awarded credits (necessary for other courses) as compensation for their time and effort towards the experiment. Before signing up, a prospective participant is shown the general information about the experiment. This contains the study's aims, along with the web browser specifications and approximate time needed to complete the study. After signing up, they are given a link to the experiment, where they are given instruction about the flow of the experiment. Before beginning the actual study, we collect demographic information and the user's previous exposure to audio descriptions. More details are given in the appendix B.2.

Each participant is shown six video sets and six audio set. A set is composed of the original media

[7] AD-Survey, developed by Zi Jin
[8] https://anupsych.sona-systems.com

(audio or video) content and the modified media. Random sampling is used to first select a 6 video set from the total 19 and then select 6 audio set from the remaining 13. After each piece of media is shown the participants asked a set of question depending upon the variant of the media. These question are listed in appendix B.2. After presenting the original audio or video the participants are asked about their familiarity and understanding of the content. They are also asked about the extra information they think the AD should include. After showing the modified video, the participants asked whether the descriptions are grammatically and logically appropriate given the information already present. This will help us estimate the correctness of our model. After showing the modified audio, the participants are asked if the descriptions help them picture the content of the video. Grammatical correctness is also judged for the modified audio. This will help us ascertain the potential help the AD system can provide to BVI people. As we do not have any BVI people in our experiments, we can not be certain that the ADs will actually be helpful to BVI people as their requirements might be different from vision corrected people. All these questions are answered in a 5 point Likert scale (Nemoto and Beglar (2014)). For both the modified audio and video, we also ask the participants if they would like additional information from the AD.

All in all, the questionnaire is designed to help us gauge the appropriateness, grammatical correctness and potential usefulness of the descriptions generated by automatic AD system. Moreover, the descriptive answers may be a useful guide to future research in the area.

Results and Analysis

As there are no pre-trained models available for PDVC (Wang et al. (2021a)), we trained the video description network as per the instruction given in the official implementation of PDVC[1]. We get a 7.83 METEOR score against the reported 8 for sentence wise evaluations and a 15.9 METEOR score against the reported 15.96 for paragraph-level evaluation. For paragraph-level evaluation, all the sentences generated for a video are treated as a paragraph and arranged w.r.t. their start time. The minor differences in the performance arise from deep learning models' fundamental nature to minimise a complex error hyperplane with several local minima. We also get the best results around 12-15 epochs (out of 30) of training which is similar to the reported number in the original work. Hence, in terms of the automatic evaluation, we get similar results as reported in the original work, which is the state of the art method for dense video captioning. This implies that the automatic description generation system that we have employed in our work is very close to the state of the art methods that deep learning models have to offer. Now we discuss the results from the experimental part of our work.

4.1 Demographics and simple Statistics of the collected data

The main aim of our work is the human evaluation of the descriptions. Out of the 70 participants who could complete the experiment, 50 recognized as female and 20 as male. As each participant is shown six video sets followed by six audio sets, we should get a total of 420 responses in each category (that are original video, modified video, original audio and modified audio). Due to some bugs in the online platform collecting data, the audio set data of two participants were not recorded. So we ended with 420 responses for the video set and 408 for the audio set. This averages out to about 22 responses per set. Due to random sampling, all the 19 video/audio sets do not receive the same number of responses. We get 15 to 29 responses per set for the video sets, and for the audio sets, we get 15 to 28 responses per set. Hence, on average 22 (and at least 15) different people evaluated a pair of original and modified media. These numbers must be kept in mind when drawing general implications and conclusions from the responses.

4.2 Survey data Analysis

In this section, we go analyse the answers we got for the Likert scale questions (appendix B.2) regarding the generated descriptions.

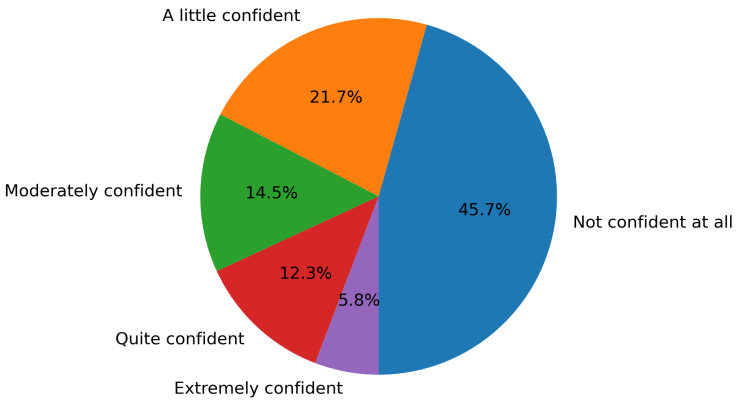

Figure 4.1: Overall participant response to the question regarding confidence in the knowledge of the video content after listening to the original audio. In a large proportion of the cases, the participants were unsure about the content in a video by just listening to the original audio present in the video.

4.2.1 Effectiveness of the system

In the audio set, only the audio is presented to the participant. This is to simulate the manner in which BVI people receive the video information in extreme cases. After listening to the original audio, the participants were asked if they were confident about video content. Their responses, summarized in Fig. 4.1, show that in 45.7% of the 408 cases, the participants did not know what could be happening in the video. This shows the need of the audio descriptions in the content selected by us for the experiments. Also, in 98.1% of the cases the participants reported to not have watched the video before, implying that they have no pre-existing information about the content of the videos. In other words, we are not using famous videos which could be guessed just by listening to the audio.

The modified audio represents the manner in which BVI people would receive the video information if our system was used. After listening to the audio with added descriptions, the participants are asked if they agree with the statement "The descriptions help you picture the video content.". As Fig. 4.2 shows, in about 55% of the cases the participants feel that the descriptions are helpful for visualization of the video content. Moreover, a much smaller proportion (26.3%) of the participants feel that the descriptions are not helpful. This is similar to the finding by Wang et al. (2021b), who report 23.8% of the sighted users found the descriptions 'Not Helpful". To understand the impact of description in more details we also plot the response to to this question w.r.t. to the response to the question about confidence in video content in Fig. 4.3. We see that for different levels of confidence in the video content after listening to original audio, the descriptions are helpful to a similar extent. About 100 (black (30) and green (70)) participants out of 185, who had no confidence about the video content after listening to the original audio, found the descriptions helpful. Even out of the 50 participants who felt quite confident about the video content, about 30 think that the descriptions are helpful. This agreement shows that the automatic AD system might be helpful for visualization, even when the user feels that they know what might be going on in the video.

Despite of this positive result, the automatically generated descriptions using deep learning are

[1]https://github.com/ttengwang/PDVC

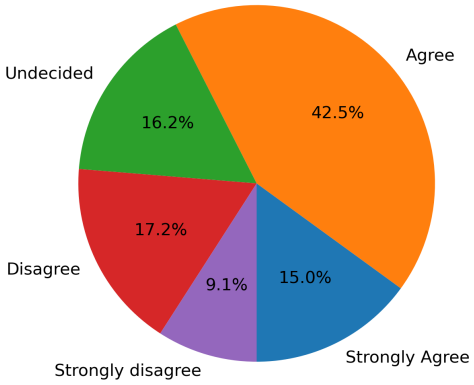

Figure 4.2: Overall participant response to the question regarding agreement with the statement "The descriptions help you picture the video content", after listening to the modified audio. In over 50% of the cases the description seem to be helpful as compared to about 25% cases where they are not.

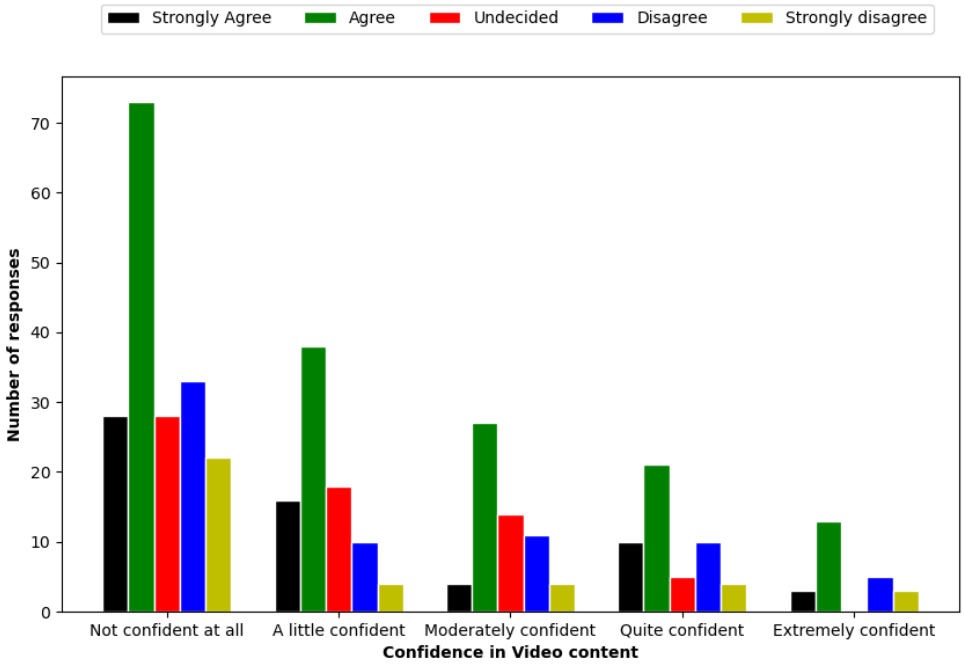

Figure 4.3: The response to the confidence in video content after hearing to the original audio is on the x-axis. The 5 bars in each group show the corresponding response to the question about helpfulness of the descriptions after listening to the modified audio. We see that a majority of people agree (black and green bars) that the descriptions are helpful for visualization of the video content. The trend is similar for all levels of confidence just after listening to the original audio.

know to be not completely accurate. We ask the similar helpfulness question to the participants watching the video along with added descriptions. As shown in Fig. 4.4, the overall responses are quite different from the modified audio. With video also shown along with the added descriptions, the AD were helpful in 28.4% of the cases. In 50.7% cases respondents felt that the descriptions did

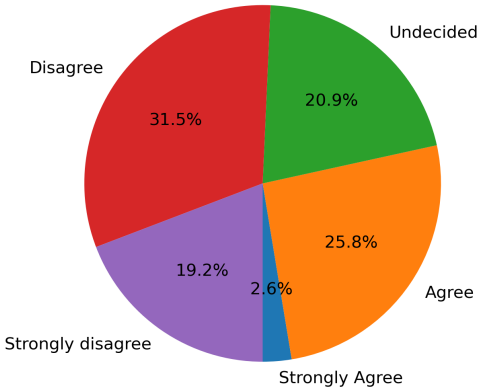

Figure 4.4: Overall participant response to the question regarding agreement with the statement "The descriptions help you understand the video content.", after watching the video with added captions. In 50.7% of the cases the description seem not to be helpful as compared to 28.4% cases where they are.

not help in improving the understanding of the video. This is expected because of the following two reasons:

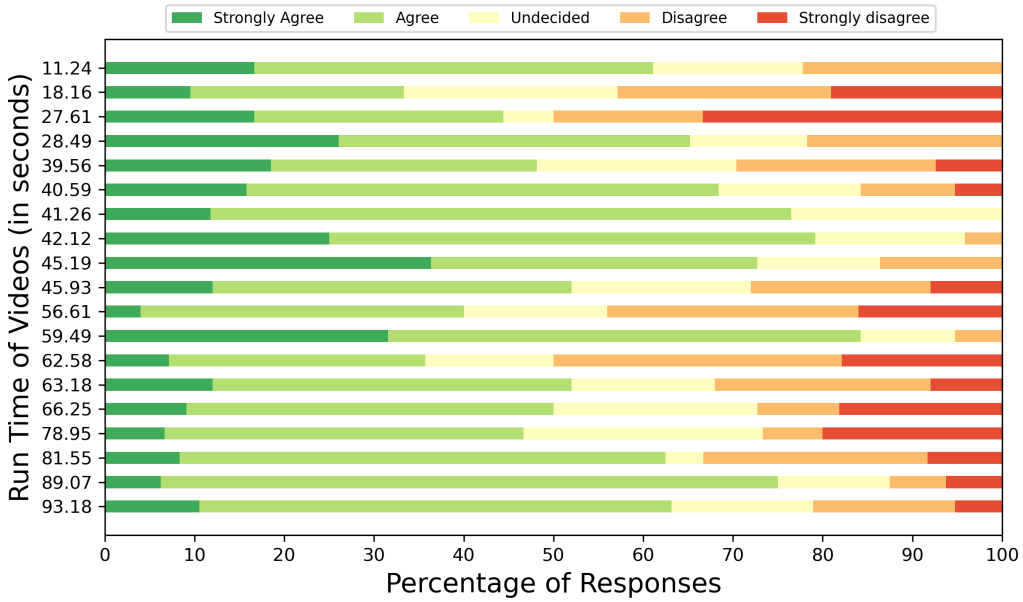

Figure 4.5: Video-wise response to the helpfulness in visualisation of video content question after listening to the modified audio. We see more than 50% agreement for 12 out of the 19 videos. We also observe no apparent affect of the length of the clip.

1. When the participants can see the video, they see and understand all the available information. Any description would only be less helpful as compared to the case when only the audio information is presented.

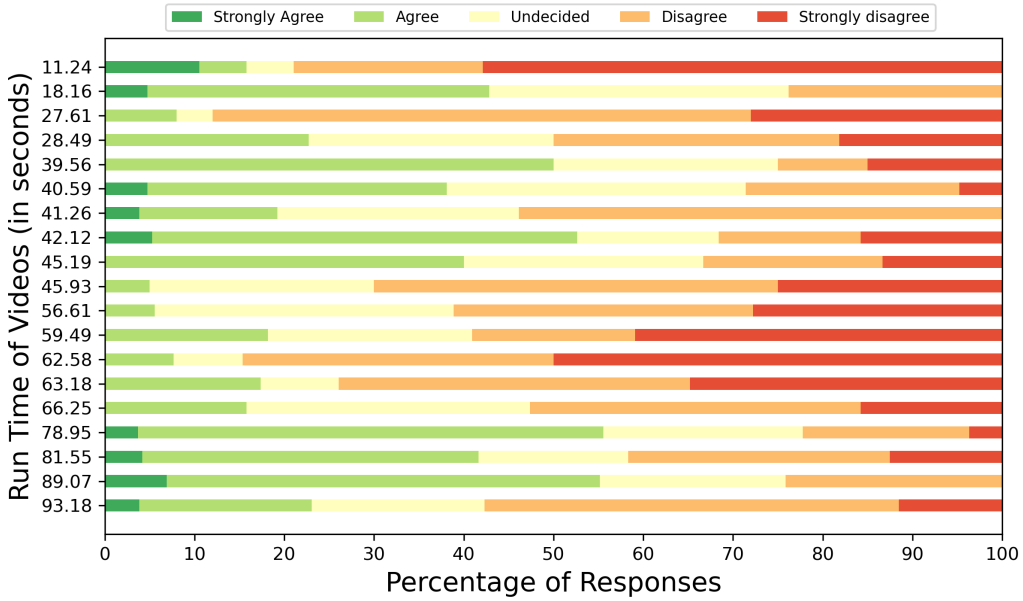

Figure 4.6: Video-wise response to the helpfulness in visualisation of video content question after watching the video with added descriptions. We see more than 50% disagreement for 10 out of the 19 videos. We also observe no apparent affect of the length of the clip. Some of the longer videos have more people agreeing to the helpful nature of the descriptions than the shorter clips.

2. As the participants see exactly what is going on in the video, even the slightest mistakes in the descriptions are much more apparent.

The numbers we get are similar to what Wang et al. (2021b) reported where the sighted people deemed 54.17% of the descriptions to be 'Not Helpful' in their user study. Another interesting observation is that the indecisiveness in judging the description increased for the modified video (20.9%) as compared to the modified audio (16.2%). This is counter intuitive, as the presence of video information should make it easier to decide whether the description is helpful or not. For a deeper understanding, we also plot the video-wise analysis for both the modified audio and video in Fig.4.5 and Fig.4.6 respectively. Comparing the corresponding videos for the two cases, we first observe a reduction in the helpfulness for understanding across all the videos. Also the indecisiveness is not a result of a single video, we see increase yellow portions in Fig.4.6 as compared to Fig.4.5 for several of the 19 videos.

Hence, although the descriptions might be helpful for visualization, they are less useful in understanding the video content. To further investigate the reasons behind the potential shortcoming of the description, we analyze the data from first (confusion/logic), second (information match) and third (grammatical Errors and redundancies) questions from the questionnaire (appendix B.2).

4.2.2 Confusion in the descriptions

After listening to the audio or watching the video with added descriptions, the participants are asked if they found the descriptions to be confusing. The overall response is shown in Fig. 4.7. Comparing the two cases, we observe an increase in confusion when the video is presented with the added descriptions. This is expected considering the potential disparity between the visual information and the descriptions. Another reason could be the lack of fluency in the descriptions which can be due to repeated phrases or grammatical errors. We further analyse both of these possibilities.

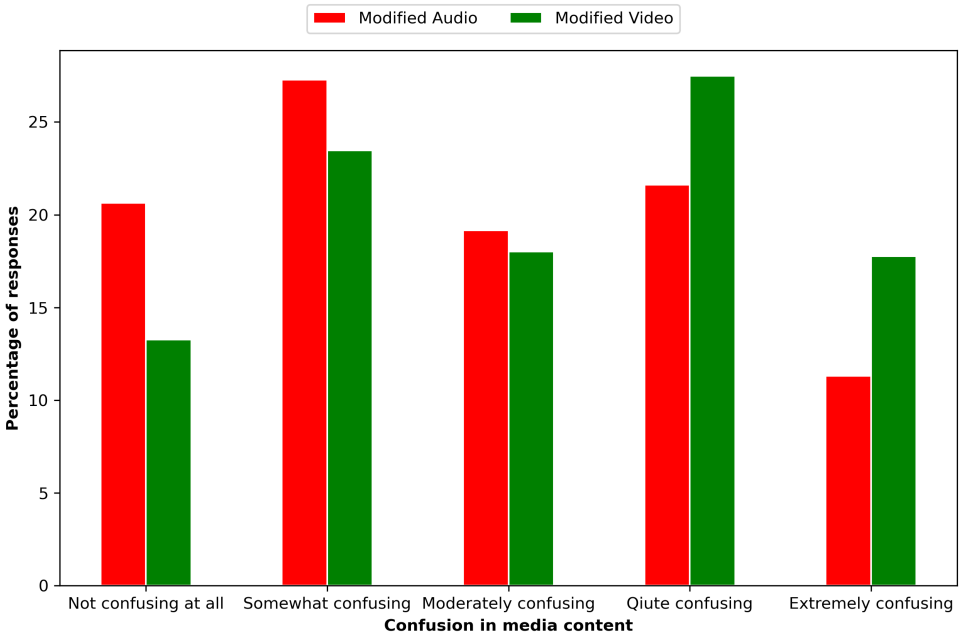

Figure 4.7: Overall response to the question regarding the descriptions being confusing. We see that in general, people found it more confusing when the video information was also presented along with the added descriptions. This is somewhat expected as video shows more information which might contradict the descriptions.

4.2.2.1 Do the description match other information?

If the descriptions do not match the other information (visual for modified video and background noise and dialogues for modified audio) presented, they can be deemed as confusing. In the case of modified video, the participants would exactly know if there is a match between the descriptions and the video content. However, for the modified audio, the participants may be unsure, as the original audio track might not have enough information to corroborate with the descriptions. Fig. 4.8 shows the percentage of responses for the information matching question. We see that for both the 'Moderately matching' and 'Quite well matching', we get a similar proportion for both the modified video and audio. Similarly, for 'No match at all', the percentage for modified video is not much higher than that for modified audio. One would expect less matching for the video case as quite a lot more information is already there.

For a better understanding of the relation between information matching and confusion, we make a grouped bar plot for different level of confusion. Fig. 4.9 shows the response of the participants for two questions, namely the confusion and information match. We observe that the majority of the 'No match at all' responses are coming from the case when the participant found the modified videos 'Extremely confusing'. Similarly, most of the videos deemed 'Quite Confusing' have only 'A little bit matching' information between the descriptions and the visual content. Moreover, the 'Quite well matching' responses co-occur with the videos judged 'Not confusing' or 'Somewhat confusing'. Hence, we see a clear negative co-relation between information matching and the modified video being confusing. This is expected as the descriptions which do not match the visual information will cause confusion. This implies that descriptions which are found confusing do not corroborate with the visual information present in the video. This points to the limitations of the current deep learning methods in

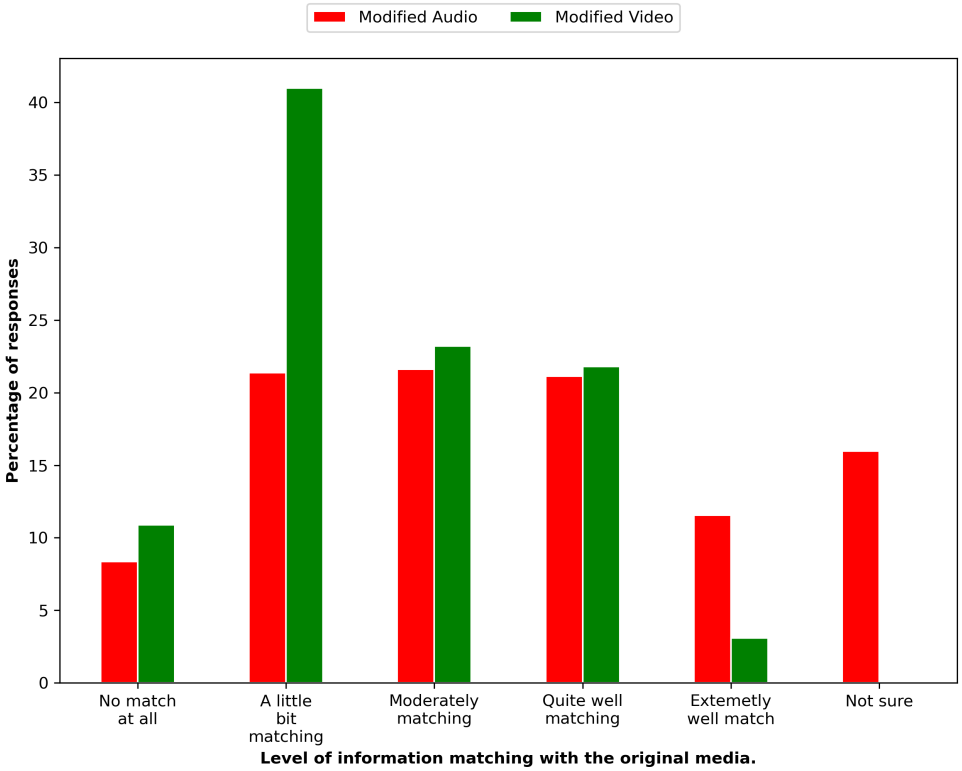

Figure 4.8: Overall response to the question regarding the descriptions matching the other information present in the media. We observe similar proportions of moderate and quite well matching in both the cases. Interestingly, the proportions of no match at all is not very high for the modified videos when compared to the modified audio.

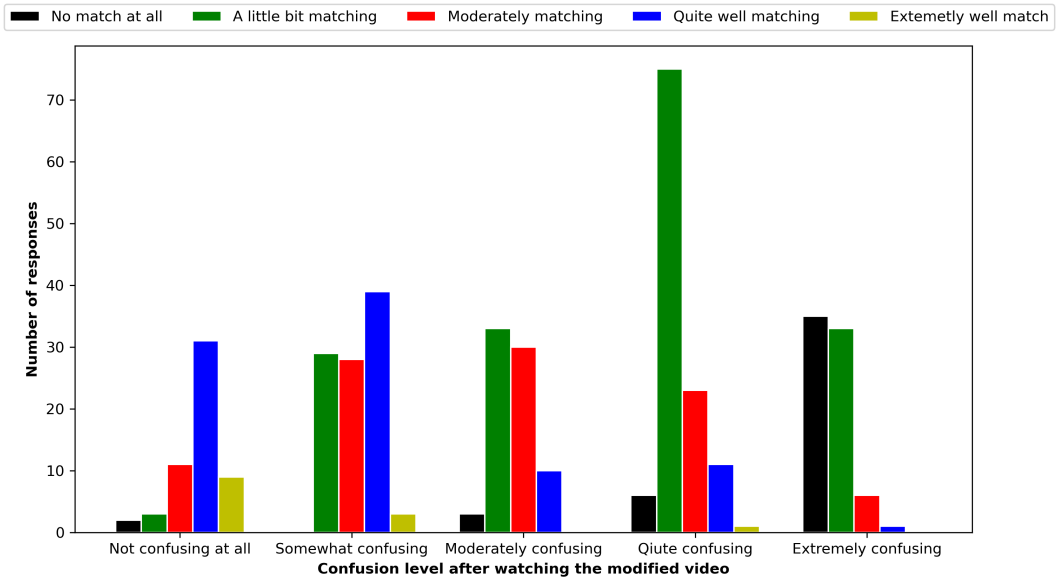

Figure 4.9: Cross-correlation between the modified video being confusing and the level of matching information between the visual information and the descriptions. We see that the majority of the 'No match at all' responses are coming from the case when the participant found the modified videos 'Extremely confusing'.

extraction of good enough features and also the caption generating module. Since the extracted video features have no ground truth, it is difficult to localize the part of the model that is more deficient in performance. It can be the case that the caption generator is able to generate complex sentences but the extracted features are not good enough to capture the relevant information. On the other hand, one can have a good feature extraction method, but the caption generator might be unable to use these properly. The grammatical accuracy and redundancies in the captions will give us more insight in the role played by the fluency of captions in creating confusion.

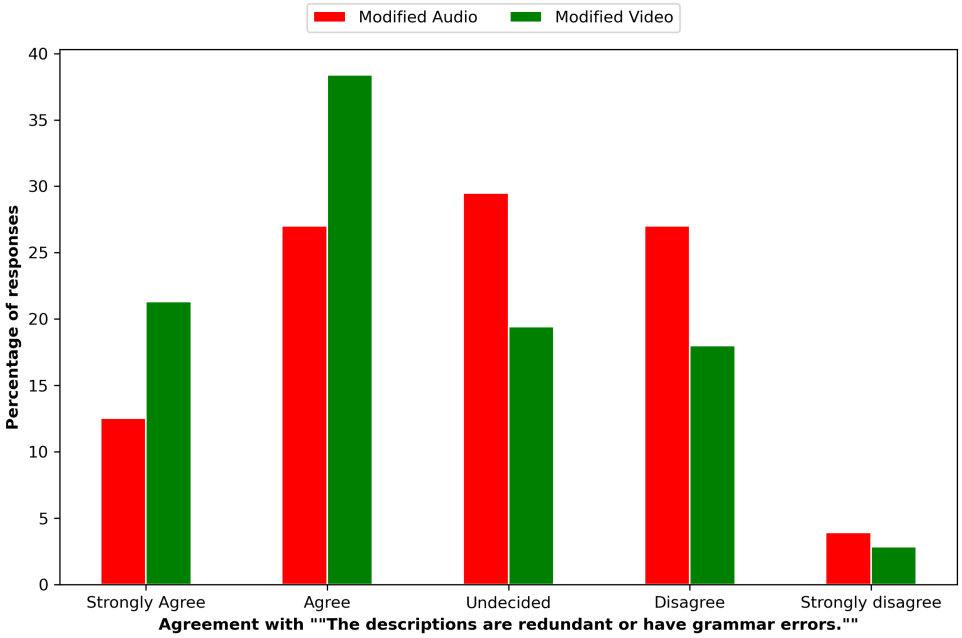

Figure 4.10: The overall response to the level of agreement with the statement 'The descriptions are redundant or have grammar errors'. With visual information, a higher proportion of the descriptions are deemed grammatically inaccurate and redundant.

4.2.2.2 Grammatical errors and Redundancies

The eloquence of the generated descriptions is a requirement for them to be helpful. Fig. 4.10 shows the responses we got when the participants were asked to judge the grammatical suitability and the presence of repeated expressions in the descriptions. We observe that people watching the video with added AD found the descriptions to be more redundant and grammatically incorrect as compared to people listening to the audio version. This is some what expected as some grammar is informed by the visual information like the number of people and gender of the main actor. However the repeated phrases should have a similar impact for both the cases. We also provide a media-wise comparison through Fig. 4.11 and Fig. 4.12. Although the participants are presented with the same descriptions and video content, for most of the videos in both the categories, we have all the five responses selected. This shows the subjectivity of the human evaluation. However, considering that the average number of responses per media about 22, a 10% response means only two people on average. Hence, the 'Strongly disagree' and 'Strongly agree' responses that are less than 10% can be considered outliers as they are the opinions of probably one or two people. In spite of this, there is huge portion of people on both agree and disagree. This is where the subjectivity of the human evaluation really makes it difficult to

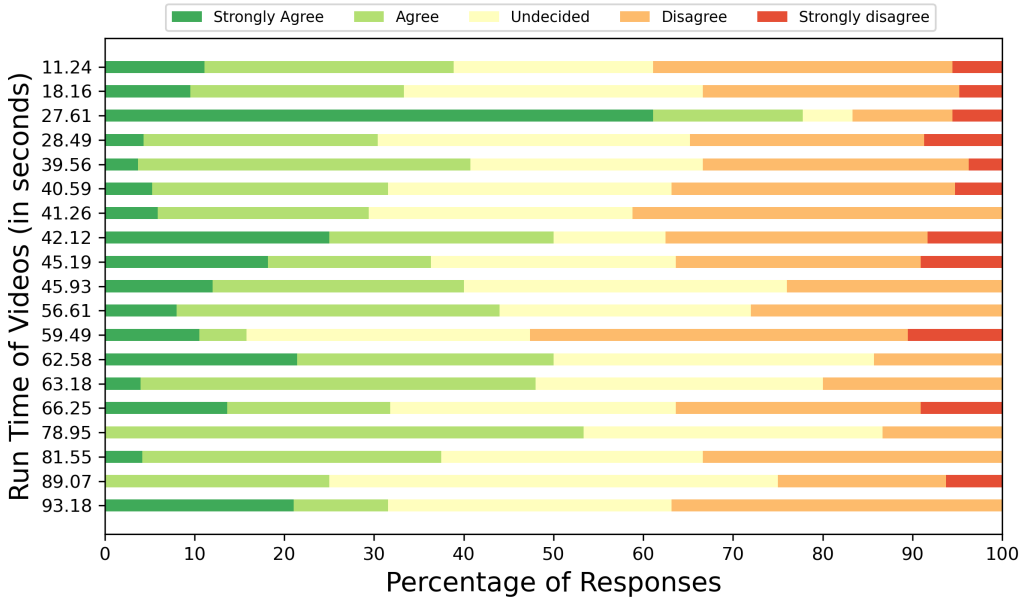

Figure 4.11: Media-wise agreement to the statement 'The descriptions are redundant or have grammar errors' after listening to the modified audio. For almost all the videos we get all the 5 possible responses. This shows the subjectivity of opinion when it comes to grammatical accuracy and redundancies.

draw conclusions. For most of the modified videos (Fig. 4.12), we see a higher level of agreement with w.r.t. the grammatical correctness and redundancies as compared to the corresponding modified audio (Fig. 4.6). Hence the overall responses shown in Fig. 4.10 are a result of all the description behaving in a general trend, where the descriptions with the visual information are deemed more grammatically incorrect.

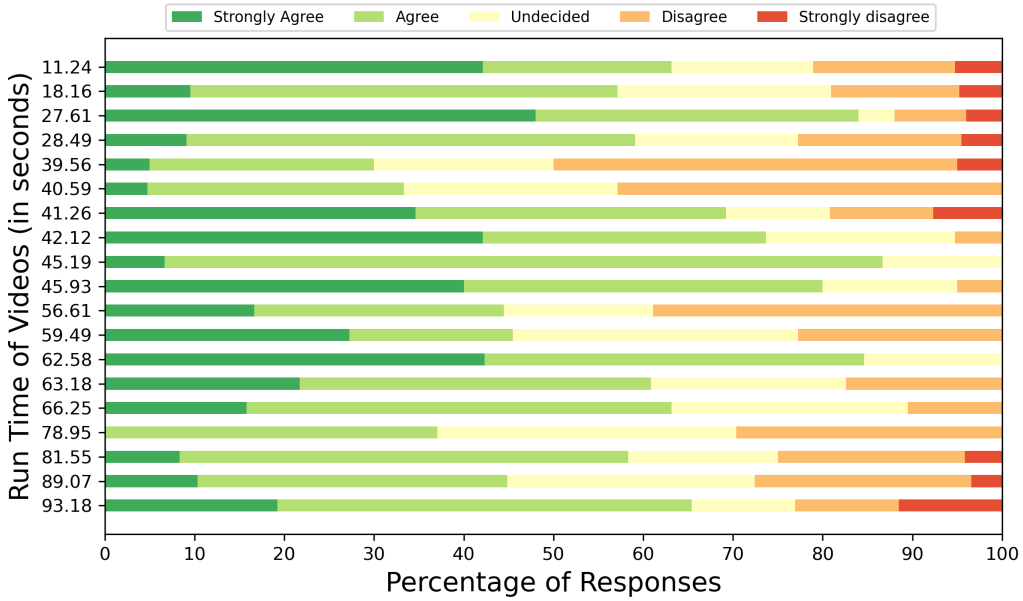

Figure 4.12: Media-wise agreement to the statement 'The descriptions are redundant or have grammar errors' after watching the video with added descriptions. As compared to the audio case, we clearly observe less fraction of undecided people. Also, we see that for 13 out of the 19 videos, the majority of participants think there are grammatical mistakes and redundancies in the descriptions.

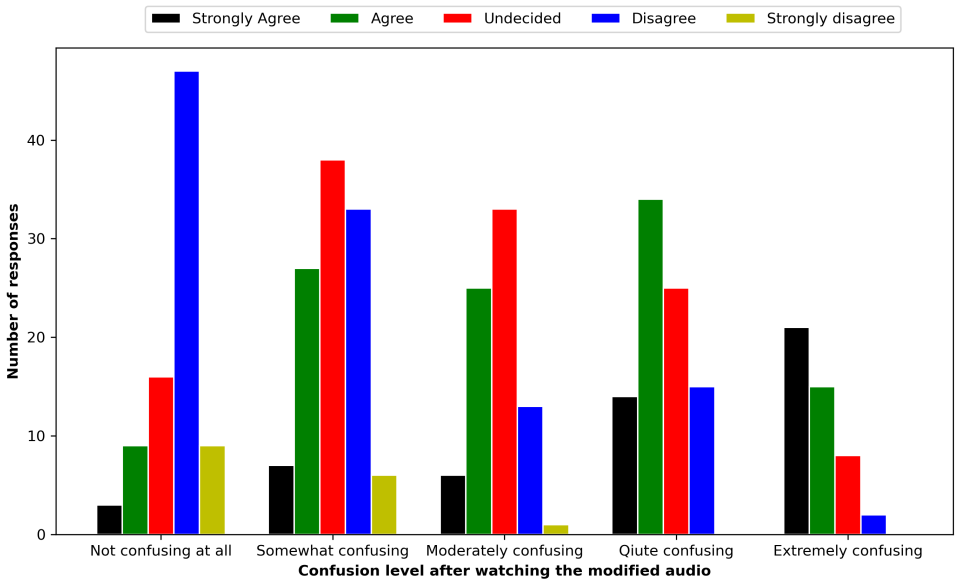

Figure 4.13: Cross-correlation between the modified audio being confusing and the grammatical accuracy of the audio descriptions. We see that the majority of the 'No confusion at all' responses are coming from the case when the participant disagreed with the statement. As the confusion level increases, more people find the descriptions to be grammatically inaccurate and redundant.

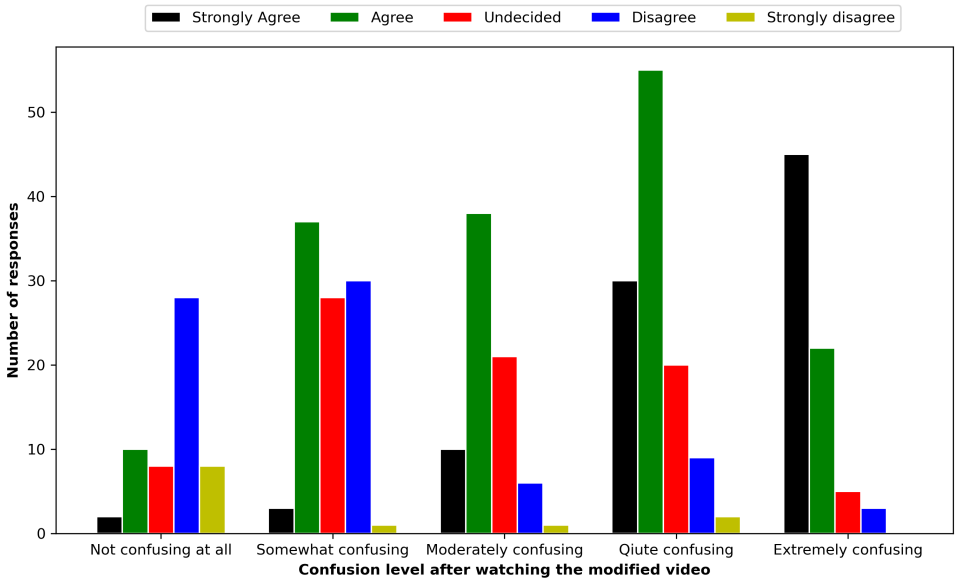

Figure 4.14: Cross-correlation between the modified video being confusing and the grammatical accuracy of the audio descriptions added on to the video. We see that the majority of the 'No confusion at all' responses are coming from the case when the participant disagreed with the statement. As the confusion level increases, more people find the descriptions to be grammatically inaccurate and redundant. This affect is shown by the increasing length of the black bar.

To understand the relationship with confusion we plot the co-occurrence of the responses for the grammatical errors question and the confusing question (appendix B.2). Fig. 4.13 shows the responses for different levels of confusion and agreement with the statement 'The descriptions are redundant

or have grammar errors'. We observe that the people who reported the modified audio to be not confusing, also think that the descriptions are pretty accurate and do not have repeating phrases. The increasing size of the black bar implies that as the AD is judged to be more confusing, it also considered grammatically inaccurate. Hence, we have can see a simple relation between the grammatical quality of the descriptions and the confusion it may create. A similar, but more apparent, trend is observed for the modified video, as shown in Fig. 4.14. The grammatical quality of the descriptions decrease as they become more confusing. Therefore, we conclude that the overall fluency (grammatical quality and absence of repeated phrases) play an important role in the deciding the level of confusion that the audio description creates.

Closing the discussion about the causes of confusion in the automatically generated captions, we found experimental evidence that both the information mismatch and grammatical errors have a crucial role to play. This points to the deficiencies in both the feature extraction and caption generation part of the deep learning module used for generating AD.

4.2.3 Analysis of the Responses to Descriptive Questions

After showing the original audio or video, the participants are asked why they think that audio descriptions might be required. Also, after showing the modified audio or video, they are asked if they expected any additional information from the descriptions. As these are descriptive answers and we received 420*2+408*2 = 1656 responses, we only mention some patterns, rather than analysing every one of them.

1. After watching and listening to tutorials, news or narration heavy original audio, many participants felt there is little need to include any AD. However, after watching the original video for these cases, most participants could point to something or another that should have a description to enhance the user experience. This shows that even the instructional videos have avenues for AD as they contain some (all be it less) elements that need descriptions.

2. After watching the modified video, at least one participant mentioned that the descriptions lacked the background details for almost all the cases. Background information usually means where the video is taking place. For example, the descriptions do not specify where the activity is being held for sports-related videos, i.e. is it outside? Is it a gymnasium? or is it a stadium?. Even with the narration/tutorial style videos, participants thought that the knowledge of background scenes and weather might be helpful for BVI people to understand the content. This is a dataset issue as most of the annotations are activity-specific and do not focus on the background details.

3. Gender misidentification is also an issue which the descriptions suffer from. For multiple cases, the automatically generated description use male nouns and pronouns for females. Many participants recognised this issue. We guess this should be dataset bias where more examples are present with men as compared to women. A deeper analysis of the dataset annotations is required to verify this hypothesis.

4. Participants also suggested that the text appearing on the screen during the videos should be read aloud to make it more accessible for BVI people. There are well established text extraction methods which might have to be integrated with the description generation. We would need a similar detection module to localize the time windows where text appears in the videos.

5. Some videos in the experiment set have a loud background noise, cheering or overlay music. As we have used the same volume level for the descriptions, a few participants felt that either the background noise needs to be reduced or the descriptions need to be more load for a clearer delivery.

6. Several participants found the descriptions to be inadequate when multiple people or people object interactions are present. After watching the video with added descriptions, participants wanted more AD to have more details like "Clarity of other individuals present in the video", "Number of players" and "multiple people playing". As there is no person or actor identification, the descriptions also misidentify people when more than one person is present and might attribute the actions of the additional people to one person. As the original dataset is only annotated with respect to activities, there is less focus to the people in the background. However, there is no part of the model that decides who/what is foreground actor and what is the background information. This causes issues when multiple people are present in the video. A person identification module might be helpful, limiting the input given to the description model to the bounding box only concerning the main actor. However, in that case one would need a separate methods to consider the background events and interactions between different people.

In conclusion, the analysis of the experimental data has shown us that the automatically generated audio descriptions might be helpful for the visualisation of the video content. However, sometimes the descriptions do not match the information presented in the video and lack grammatical accuracy making them confusing. The subjective responses point to some of the critical shortcomings of the descriptions like limited background information, gender misidentification, inadequacy for multi-people interactions etc. These might help design future systems for automatic audio descriptions.

Conclusions and Future Development

This work aimed to develop and evaluate an automatic audio description (AD) generation system to make the video content more accessible to the blind or visually impaired (BVI) people. We acknowledge the social, linguistic benefits of AD and its overall positive impact on the life of BVI people. However, as most AD are scripted by humans and then carefully inserted in the video, the pipeline is time and resource-intensive. The high cost of manually generated automatic descriptions makes them infeasible for the large amount of user-uploaded data on social media and video sharing sites. Following the pipeline suggested by Wang et al. (2021b) for automatic AD generation, we identified the video description as the core task. We provide an extensive literature survey to go through the pre and post deep learning methodologies for video description and dense video captioning (DVC). We also summarise the datasets and automatic and human evaluations metrics used to assess these systems. Referring to the potential shortcomings of the pipeline used by Wang et al. (2021b), we employed a state of the art dense video captioning model to generate both the descriptions and the respective time windows. The detailed procedure to put the descriptions in the original videos and evaluate them using human participants is given in Methodology. Although the automatic evaluation metrics help in choosing the DVC model, they are not designed to cater to the potential needs of BVI people. Hence we decided to perform experiments with human participants to judge the suitability of the descriptions generated by the DVC model as AD.

The responses from the participants suggest that the descriptions are helpful for the visualisation of the video content compared to the raw audio. However, when the visual information is also presented along with the generated descriptions, participants report more confusion, information mismatch, grammatical mistakes, and redundancies than the modified audio. This implies that, although the descriptions give a general idea of what might be happening in the video, the descriptions are deficient when it comes to details and fluency. As the dataset used for the experiment (ActivityNet Captions, Krishna et al. (2017)) is annotated with the activities in focus, it lacks the background details and the information that a blind person may need from the video. Moreover, some participants also reported that the descriptions were incorrect in terms of the nouns and pronouns used to refer to women present in the video. This could be the result of potential gender bias in the dataset, where more examples are present of one gender as compared to the other. Also, the DVC model we used is incapable of using any text information present on the video. A better system would be able to localise the part of the video where text is present, extract the text, decide if it should be included in the description and then incorporate it with the rest of the audio information. A similar case-specific issue is the presence of background and how our system with fixed volume descriptions cannot handle this. Hence, relatively simple subsystems to address these issues for specific videos might be helpful in increasing the overall quality of the automatic AD system. A more general problem made evident by the user responses were the poor descriptions when multiple people were present in the frame. The current model cannot differentiate between different people and might assign activities to the

wrong person or use wrong pronouns when multiple people are present. A person identification and tracking module (like a bounding box across different frames) might help limit the input given to the description generator. Still, it might create other issues like missing interactions between people.

Expecting a much smaller registration for the experiments, we limited the description generated by a single method. Although this lets us evaluate our system, but does not help compare the performance against other systems. For example, in the ground truth implementation of DVC, the time windows of the events are given to the model, and it predicts just the descriptions. Evaluating the descriptions through this pipeline would have let us assess the performance of the captioning module. Moreover, we can just use the ground truth annotations to get a more realistic idea of the shortcomings of the annotations with respect to the requirements of BVI people. This would tell us the best descriptions we can achieve with the current datasets. Another variant might include the implementation of the pipeline given by Wang et al. (2021b). Although they released their results (modified videos), they have used the videos from the training set of the dataset used in the captioning module. As we worked with the validation set videos, the direct use of their videos is not possible. Finally, since the models for single sentence captioning of videos have achieved better performance (in terms of METEOR scores for their datasets like MVAD, Torabi et al. (2015)), a video captioning network can be used to describe each clip of the given video. As this would need the ground truth time windows for the event, it can be compared to the respective DVC representation. However, captioning clips independently (as done by Wang et al. (2021b) in their automatic AD system) can increase the redundancies in the descriptions as the model has no contextual information besides the frames inside the clip time window. Recently, Deng et al. (2021) have used the concept of grounding the descriptions to refine them further. Grounding refers to localising the time window in the video, given the descriptions. They have reported the state of the art METEOR score for the ActivityNet Captions dataset. Even though we have all these pipelines to generate descriptions, we must remember that most of the datasets used in DVC are not annotated for increasing accessibility for blind people. Creating an appropriate dataset is difficult as annotating videos with descriptions suitable for blind people is more expensive (the whole reason behind automatic AD) than simply describing the videos. A potential solution would be to train the description models on the AD extracted from the media on video streaming platforms, like Netflix and Amazon Prime. However, as all these films and shows are professionally shot, the learned model might not generalise the user-uploaded video content. Hence, there are several possibilities for further experimentation and research.

Guidelines for AD

Country	Guidelines
Unite States	`https://dcmp.org/learn/descriptionkey`, `https://adp.acb.org/ad.html`
France	`https://www.audiodescription-france.org/`
Australia	mediaaccess.org.au
UK	`https://www.ofcom.org.uk/about-ofcom/website/regulator-archives`
Canada	`https://crtc.gc.ca/eng/info_sht/b322.htm`
European	`http://www.adlabproject.eu/Docs/adlab%20book/index.html`

Table A.1: Some guidelines for scripting and voicing audio descriptions for videos, TV-broadcasts and Films.

Experiment Details

B.1 Videos selected for user study

Following is the list of 19 videos utilized in the user study. The category gives an overall idea about the core information in the video.

Category	Video watch ID on YouTube
Sport	GGSY1Qvo990, 3hZjxdMcG6o, otWTm1_aAqI, DFJBJkCR0Bk, 0GWJ-VHFlTk
Instructional	wPYr19iFxhw, HpJ2pr0ykqo, WqDep-4l0yc
Pianta hitting	rgAALWYnRrg
Grass cutting	nezTU6Bq5hM, 1hiyhNqakMI
Knit and run	Nqh3RtLRleU
Hopscotch	fdd5ixvEXOE, OvGxDaayPcw
Music	Ti1ZaH0VGfg
Advertisement	6rOmYOU7748
Bumper Cars	yQ2AirKmnTM
Dog tricks	Jp86pFKlsw
Kids and slide	Br1Ty6PCrv8

Table B.1: The categories and the YouTube watch IDs of the selected videos. For accessing these on YouTube, use '*https://www.youtube.com/watch?v=**vid***', where ***vid*** is replace by the watch IDs given here.

B.2 Questionnaire used in the experiments

Following instructions are given to the participants before the experiment:

The main objective is to determine whether the automatically generated descriptions improve the understandability of the media (video/audio). The original videos are from YouTube so you might have seen them. You will be shown some sets of audio/video where each set has the original media and the media with added audio descriptions.

Then we collect some demographic information. After this the experiment starts and the audio/video information is presented to the participant. Depending upon the class of the media a set of questions is asked. Most of them are to be answered in the 5 point Likert Scale, while some descriptive answers are also allowed. Following is the set of questions asked in each case:

B.2.1 Demographics data

1. Age
2. Gender
3. Primary Language
4. Previous used of AD: What is your experience with videos added with audio descriptions?

B.2.2 Question after Original-Audio

1. Based on what you heard, how confident are you that you can describe the video content? (5 point Likert Scale)
2. What information do you want to get from the audio descriptions? (Descriptive)

B.2.3 Question after Original-Video

1. Have you seen this video before? (Yes/No)
2. For what reason did you think the video currently needs to be added with audio descriptions? (Descriptive)

B.2.4 Question after Modified-Audio

1. Did you think the descriptions are confusing? (5 point Likert Scale)
2. Do the descriptions match the information given by existing audio? (5 point Likert Scale, Not sure)
3. How much do you agree with the following statement? The descriptions are redundant or have grammar errors. (5 point Likert Scale)
4. How much do you agree with the following statement? The descriptions help you picture the video content. (5 point Likert Scale)
5. What other information should the description provide? (Descriptive)

B.2.5 Question after Modified-Video

‘

1. Did you think the descriptions are confusing? (5 point Likert Scale)
2. Do the descriptions match the information present in the video? (5 point Likert Scale)
3. How much do you agree with the following statement? The descriptions are redundant or have grammar errors. (5 point Likert Scale)
4. How much do you agree with the following statement? The descriptions help you understand the video content. (5 point Likert Scale)
5. What other information should the description provide? (Descriptive)

Bibliography

Aafaq, N.; Mian, A.; Liu, W.; Gilani, S. Z.; and Shah, M., 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Comput. Surv.*, 52, 6 (Oct. 2019). doi:10.1145/3355390. https://doi.org/10.1145/3355390.

Amirian, S.; Rasheed, K.; Taha, T. R.; and Arabnia, H. R., 2020. Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap. *IEEE Access*, 8 (2020), 218386–218400. doi:10.1109/ACCESS.2020.3042484.

Arandjelovic, R. and Zisserman, A., 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, 609–617.

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*, 65–72. Association for Computational Linguistics, Ann Arbor, Michigan. https://aclanthology.org/W05-0909.

Bateman, J., 1997. Sentence generation and systemic grammar: an introduction. *Iwanami Lecture Series: Language Sciences, Iwanami Shoten Publishers, Tokyo*, (1997).

Brand, M., 1997. The" inverse hollywood problem": From video to scripts and storyboards via causal analysis. In *AAAI/IAAI*, 132–137. Citeseer.

Campos, V. P.; de Araújo, T. M.; de Souza Filho, G. L.; and Gonçalves, L. M., 2020. Cinead: a system for automated audio description script generation for the visually impaired. *Universal Access in the Information Society*, 19, 1 (2020), 99–111.

Caro, M. R., 2016. Testing audio narration: the emotional impact of language in audio description. *Perspectives*, 24, 4 (2016), 606–634. doi:10.1080/0907676X.2015.1120760. https://doi.org/10.1080/0907676X.2015.1120760.

Cha, J., 2013. Do online video platforms cannibalize television?: How viewers are moving from old screens to new ones. *Journal of advertising research*, 53, 1 (2013), 71–82.

Chen, D. and Dolan, W., 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 190–200. Association for Computational Linguistics, Portland, Oregon, USA. https://aclanthology.org/P11-1020.

Dabre, R.; Chu, C.; and Kunchukuttan, A., 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53, 5 (Sep. 2020). doi:10.1145/3406095. https://doi.org/10.1145/3406095.

Dalal, N. and Triggs, B., 2005. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 886–893 vol. 1. doi:10.1109/CVPR.2005.177.

Das, P.; Xu, C.; Doell, R. F.; and Corso, J. J., 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, (2013), 2634–2641.

Deng, C.; Chen, S.; Chen, D.; He, Y.; and Wu, Q., 2021. Sketch, ground, and refine: Top-down dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 234–243.

Devlin, J.; Cheng, H.; Fang, H.; Gupta, S.; Deng, L.; He, X.; Zweig, G.; and Mitchell, M., 2015. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*, (2015).

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2625–2634.

Duan, X.; Huang, W.; Gan, C.; Wang, J.; Zhu, W.; and Huang, J., 2018. Weakly supervised dense event captioning in videos. *arXiv preprint arXiv:1812.03849*, (2018).

Elliott, D. and Keller, F., 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 452–457.

Ellis, K., 2015. Netflix closed captions offer an accessible model for the streaming video industry, but what about audio description? *Communication, Politics Culture*, 47, 3 (2015), 3–20. `https://search-informit-org.virtual.anu.edu.au/doi/10.3316/informit.113665255090751`.

Ely, R.; Emerson, R. W.; Maggiore, T.; Rothberg, M.; O'Connell, T.; and Hudson, L., 2006. Increased content knowledge of students with visual impairments as a result of extended descriptions. *Journal of Special Education Technology*, 21, 3 (2006), 31–43. doi: 10.1177/016264340602100304. `https://doi.org/10.1177/016264340602100304`.

Felzenszwalb, P.; McAllester, D.; and Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, 1–8. Ieee.

Finbow, S., 2010. The state of audio description in the united kingdom – from description to narration. *Perspectives*, 18, 3 (2010), 215–229. doi:10.1080/0907676X.2010.485685. `https://doi.org/10.1080/0907676X.2010.485685`.

Fujita, S.; Hirao, T.; Kamigaito, H.; Okumura, M.; and Nagata, M., 2020. Soda: Story oriented dense video captioning evaluation framework. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 517–531. Springer.

Gao, L.; Guo, Z.; Zhang, H.; Xu, X.; and Shen, H. T., 2017. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19, 9 (2017), 2045–2055.

Gleason, C.; Carrington, P.; Cassidy, C.; Morris, M. R.; Kitani, K. M.; and Bigham, J. P., 2019a. "it's almost like they're trying to hide it": How user-provided image descriptions have failed to make twitter accessible. In *The World Wide Web Conference*, WWW '19 (San Francisco, CA, USA, 2019), 549–559. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3308558.3313605. https://doi.org/10.1145/3308558.3313605.

Gleason, C.; Pavel, A.; Gururaj, H.; Kitani, K.; and Bigham, J., 2020. Making gifs accessible. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '20 (Virtual Event, Greece, 2020). Association for Computing Machinery, New York, NY, USA. doi:10.1145/3373625.3417027. https://doi.org/10.1145/3373625.3417027.

Gleason, C.; Pavel, A.; Liu, X.; Carrington, P.; Chilton, L. B.; and Bigham, J. P., 2019b. Making memes accessible. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19 (Pittsburgh, PA, USA, 2019), 367–376. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3308561.3353792. https://doi.org/10.1145/3308561.3353792.

Greening, J. and Rolph, D., 2007. *Accessibility: raising awareness of audio description in the UK*, 127 – 138. Brill, Leiden, The Netherlands. ISBN 9789401209564. doi:https://doi.org/10.1163/9789401209564_010. https://brill.com/view/book/9789401209564/B9789401209564-s010.xml.

Guinness, D.; Cutrell, E.; and Morris, M. R., 2018. Caption crawler: Enabling reusable alternative text descriptions using reverse image search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18 (Montreal QC, Canada, 2018), 1–11. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3173574.3174092. https://doi.org/10.1145/3173574.3174092.

Hanley, M.; Barocas, S.; Levy, K.; Azenkot, S.; and Nissenbaum, H., 2021. Computer vision and conflicting values: Describing people with automated alt text. *CoRR*, abs/2105.12754 (2021). https://arxiv.org/abs/2105.12754.

Heilbron, F. C.; Escorcia, V.; Ghanem, B.; and Niebles, J. C., 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 961–970. doi:10.1109/CVPR.2015.7298698.

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

Hongeng, S.; Bremond, F.; and Nevatia, R., 2000. Bayesian framework for video surveillance application. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, vol. 1, 164–170 vol.1. doi:10.1109/ICPR.2000.905296.

Iashin, V. and Rahtu, E., 2020. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *British Machine Vision Conference (BMVC)*.

Karita, S.; Chen, N.; Hayashi, T.; Hori, T.; Inaguma, H.; Jiang, Z.; Someki, M.; Soplin, N. E. Y.; Yamamoto, R.; Wang, X.; et al., 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 449–456. IEEE.

Kim, J.; Kim, J.; and Rye, H., 2014. Heart-to-feel': a new audio description coding scheme for the visually impaired on affective cinematography and emotive vibration. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video, TVX*, vol. 14.

Koller, D.; Heinze, N.; and Nagel, H.-H., 1991. Algorithmic characterization of vehicle trajectories from image sequences by motion verbs. *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (1991), 90–95.

Konečnỳ, J. and Hagara, M., 2014. One-shot-learning gesture recognition using hog-hof features. *The Journal of Machine Learning Research*, 15, 1 (2014), 2513–2532.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J., 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, 706–715.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kurihara, K.; Imai, A.; Seiyama, N.; Shimizu, T.; Sato, S.; Yamada, I.; Kumano, T.; Tako, R.; Miyazaki, T.; Ichiki, M.; Takagi, T.; and Sumiyoshi, H., 2019. Automatic generation of audio descriptions for sports programs. *SMPTE Motion Imaging Journal*, 128, 1 (2019), 41–47. doi:10.5594/JMI.2018.2879261.

Lagger, C.; Lux, M.; and Marques, O., 2017. What makes people watch online videos: An exploratory study. *Comput. Entertain.*, 15, 2 (Apr. 2017). doi:10.1145/3034706. https://doi.org/10.1145/3034706.

LeCun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; and Jackel, L. D., 1989. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1, 4 (1989), 541–551. doi:10.1162/neco.1989.1.4.541.

Lee, S.-Y. and Lee, S.-W., 2015. Online video services and other media: Substitutes or complement. *Computers in Human Behavior*, 51 (2015), 293–299. doi:https://doi.org/10.1016/j.chb.2015.03.073. https://www.sciencedirect.com/science/article/pii/S0747563215002745.

Li, P.; Chen, C.; Zheng, W.; Deng, Y.; Ye, F.; and Zheng, Z., 2019a. Std: An automatic evaluation metric for machine translation based on word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27, 10 (2019), 1497–1506. doi:10.1109/TASLP.2019.2922845.

Li, S.; Tao, Z.; Li, K.; and Fu, Y., 2019b. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3, 4 (2019), 297–312. doi:10.1109/TETCI.2019.2892755.

Lin, C.-Y. and Och, F. J., 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 605–612. Barcelona, Spain. doi:10.3115/1218955.1219032. https://aclanthology.org/P04-1077.

López, A. P., 2008. Audio description as language development and language learning for blind and visual impaired children. In *Thinking Translation: Perspectives from Within and Without: Conference Proceedings, Third UEA Postgraduate Translation Symposium*, 113. Universal-Publishers.

Low, C.; McCamey, E.; Gleason, C.; Carrington, P.; Bigham, J. P.; and Pavel, A., 2019. Twitter a11y: A browser extension to describe images. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19 (Pittsburgh, PA, USA, 2019), 551–553. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3308561.3354629. https://doi.org/10.1145/3308561.3354629.

Lowe, D., 1999. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1150–1157 vol.2. doi:10.1109/ICCV.1999.790410.

Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; and Yuille, A., 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, (2014).

Mazur, I. and Chmiel, A., 2012. Towards common european audio description guidelines: results of the pear tree project. *Perspectives*, 20, 1 (2012), 5–23. doi:10.1080/0907676X.2011.632687. https://doi.org/10.1080/0907676X.2011.632687.

Meehan, M. and McCallig, J., 2019. Effects on learning of time spent by university students attending lectures and/or watching online videos. *Journal of Computer Assisted Learning*, 35, 2 (2019), 283–293. doi:https://doi.org/10.1111/jcal.12329. https://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12329.

Miller, G. A., 1998. *WordNet: An electronic lexical database*. MIT press.

Moreno, A. I. and Vermeulen, A., 2015. Using visp (videos for speaking), a mobile app based on audio description, to promote english language learning among spanish students: a case study. *Procedia-Social and Behavioral Sciences*, 178 (2015), 132–138.

Mun, J.; Yang, L.; Ren, Z.; Xu, N.; and Han, B., 2019. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6588–6597.

Nemoto, T. and Beglar, D., 2014. Likert-scale questionnaires. In *JALT 2013 conference proceedings*, 1–8.

Pan, Y.; Mei, T.; Yao, T.; Li, H.; and Rui, Y., 2016. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4594–4602.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J., 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02 (Philadelphia, Pennsylvania, 2002), 311–318. Association for Computational Linguistics, USA. doi:10.3115/1073083.1073135. https://doi.org/10.3115/1073083.1073135.

Park, J. S.; Rohrbach, M.; Darrell, T.; and Rohrbach, A., 2019. Adversarial inference for multi-sentence video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6598–6608.

Parton, B. S., 2016. Video captions for online courses: Do youtube's auto-generated captions meet deaf students' needs? *Journal of Open, Flexible and Distance Learning*, 20 (2016), 8–18.

Perego, E., 2016. Gains and losses of watching audio described films for sighted viewers. *Target. International Journal of Translation Studies*, 28, 3 (2016), 424–444.

Pinhanez, C. and Bobick, A., 1998. Human action detection using pnf propagation of temporal constraints. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, 898–904. doi:10.1109/CVPR.1998.698711.

Plaza, M., 2017. Cost-effectiveness of audio description production process: comparative analysis of outsourcing and 'in-house' methods. *International Journal of Production Research*, 55, 12 (2017), 3480–3496. doi:10.1080/00207543.2017.1282182. `https://doi.org/10.1080/00207543.2017.1282182`.

Pollard, C. and Sag, I. A., 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

Post, M., 2018. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771 (2018). `http://arxiv.org/abs/1804.08771`.

Reiter, E. and Dale, R., 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3, 1 (1997), 57–87.

Rohrbach, A.; Rohrbach, M.; Qiu, W.; Friedrich, A.; Pinkal, M.; and Schiele, B., 2014. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, 184–195. Springer.

Rohrbach, A.; Rohrbach, M.; Tandon, N.; and Schiele, B., 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3202–3212.

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*, 123, 1 (2017), 94–120.

Rohrbach, M.; Amin, S.; Andriluka, M.; and Schiele, B., 2012a. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 1194–1201. doi:10.1109/CVPR.2012.6247801.

Rohrbach, M.; Qiu, W.; Titov, I.; Thater, S.; Pinkal, M.; and Schiele, B., 2013. Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Rohrbach, M.; Regneri, M.; Andriluka, M.; Amin, S.; Pinkal, M.; and Schiele, B., 2012b. Script data for attribute-based recognition of composite activities. In *Computer Vision – ECCV 2012*, 144–157. Springer Berlin Heidelberg, Berlin, Heidelberg.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115, 3 (2015), 211–252. doi:10.1007/s11263-015-0816-y.

Schmeidler, E. and Kirchner, C., 2001. Adding audio description: Does it make a difference? *Journal of Visual Impairment & Blindness*, 95, 4 (2001), 197–212. doi:10.1177/0145482X0109500402. https://doi.org/10.1177/0145482X0109500402.

Simonyan, K. and Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14 (Montreal, Canada, 2014), 568–576. MIT Press, Cambridge, MA, USA.

Simonyan, K. and Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1409.1556.

Simpson, J. A. and Australia., B. C., 1999. *When a word is worth a thousand pictures : improved television access for blind viewers in the digital era / John A. Simpson.* Blind Citizens Australia Prahran, Vic. ISBN 0958706522.

Snyder, J., 2005. Audio description: The visual made verbal. *International Congress Series*, 1282 (2005), 935–939. doi:https://doi.org/10.1016/j.ics.2005.05.215. https://www.sciencedirect.com/science/article/pii/S0531513105009817. Vision 2005.

Stangl, A.; Morris, M. R.; and Gurari, D., 2020. " person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A., 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Torabi, A.; Pal, C.; Larochelle, H.; and Courville, A., 2015. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, (2015).

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Ustundag, B. C. and Unel, M., 2014. Human action recognition using histograms of oriented optical flows from depth. In *Advances in Visual Computing*, 629–638. Springer International Publishing, Cham.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I., 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D., 2015. Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Venugopalan, S.; Hendricks, L. A.; Mooney, R.; and Saenko, K., 2016. Improving lstm-based video description with linguistic knowledge mined from text. *arXiv preprint arXiv:1604.01729*, (2016).

Venugopalan, S.; Rohrbach, M.; Donahue, J.; Mooney, R.; Darrell, T.; and Saenko, K., 2015a. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, 4534–4542.

Venugopalan, S.; Xu, H.; Donahue, J.; Rohrbach, M.; Mooney, R.; and Saenko, K., 2015b. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1494–1504. Association for Computational Linguistics, Denver, Colorado. doi:10.3115/v1/N15-1173. https://aclanthology.org/N15-1173.

Vercauteren, G., 2012. A narratological approach to content selection in audio description. towards a strategy for the description of narratological time. *MonTI. Monografías de Traducción e Interpretación*, , 4 (2012), 207–231.

Viola, P. and Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, I–I. doi:10.1109/CVPR.2001.990517.

Walczak, A. and Fryer, L., 2017. Creative description: The impact of audio description style on presence in visually impaired audiences. *British Journal of Visual Impairment*, 35, 1 (2017), 6–17. doi:10.1177/0264619616661603. https://doi.org/10.1177/0264619616661603.

Wang, J.; Jiang, W.; Ma, L.; Liu, W.; and Xu, Y., 2018. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7190–7198.

Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; and Gool, L. V., 2017. Temporal segment networks for action recognition in videos. *CoRR*, abs/1705.02953 (2017). http://arxiv.org/abs/1705.02953.

Wang, T.; Zhang, R.; Lu, Z.; Zheng, F.; Cheng, R.; and Luo, P., 2021a. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6847–6857.

Wang, Y.; Liang, W.; Huang, H.; Zhang, Y.; Li, D.; and Yu, L.-F., 2021b. Toward automatic audio description generation for accessible videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12.

Wu, S.; Wieland, J.; Farivar, O.; and Schiller, J., 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17 (Portland, Oregon, USA, 2017), 1180–1192. Association for Computing Machinery, New York, NY, USA. doi:10.1145/2998181.2998364. https://doi.org/10.1145/2998181.2998364.

Xiong, Y.; Dai, B.; and Lin, D., 2018. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 468–483.

Yao, L.; Torabi, A.; Cho, K.; Ballas, N.; Pal, C.; Larochelle, H.; and Courville, A., 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, 4507–4515.

Yu, H.; Wang, J.; Huang, Z.; Yang, Y.; and Xu, W., 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4584–4593.

Zhou, L.; Xu, C.; and Corso, J. J., 2018a. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhou, L.; Zhou, Y.; Corso, J. J.; Socher, R.; and Xiong, C., 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8739–8748.

Zhu, Y.; Lu, S.; Zheng, L.; Guo, J.; Zhang, W.; Wang, J.; and Yu, Y., 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1097–1100.