1. Every Picture Tells a Story: Generating Sentences from Images

Summary: This paper investigates methods to generate short descriptive sentences from images. Here they describe a system that can compute a score linking an image to a sentence. This score can be used to attach a descriptive sentence to a given image, or to obtain images that illustrate a given sentence. The score is obtained by comparing an estimate of meaning obtained from the image to one obtained from the sentence. Each estimate of meaning comes from a discriminative procedure that is learned using data.

Methodology: There are two applications for methods used here that link text and images: *Illustration*, where one finds pictures suggested by text (perhaps to suggest illustrations from a collection); and *annotation*, where one finds text annotations for images (perhaps to allow keyword search to find more images.

The scoring procedure is built around an intermediate representation, which they call the meaning of the image. In effect, image and sentence are each mapped to this intermediate space, and the results are compared; similar meanings result in a high score. The advantage of doing so is that each of these maps can be adjusted discriminatively. While the meaning space could be abstract, in our implementation we use a direct representation of simple sentences as a meaning space. This allows us to exploit distributional semantics ideas to deal with out of vocabulary words.

2. Automatic Generation of Descriptive Titles for Video Clips Using Deep Learning:

Summary: The paper proposes an architecture that utilizes image/video captioning methods and NLP systems to generate a title and a concise abstract for a video. The proposed system functions and operates as follows: it reads a video; representative image frames are identified and selected; the image frames are captioned; NLP is applied to all generated captions together with text summarization; and finally, a title and an abstract are generated for the video.

Methodology: The proposed architecture consists of two different, complementary processes: Video Captioning(Decoder-Encoder) and Text Summarization

Video Captioning: The captioning part consists of two phases: Encoder and Decoder. The En-coder part extracts the image information using convolutional neural networks like object detection methods to extract the objects and actions and then put them in a vector. ResNet, DenseNet, RCNN series, Yolo and can be used as object detection methods. Then the vector enters the decoder phase. The Decoder gets the vector and then with RNN methods generates a meaningful caption for the image.

Text Summarization : Authors deployed an unsupervised learning algorithm which ranks sentences according to similarity and builds a similarity matrix which chooses the top N sentences to formulate a title for the video and abstractive text summarization was utilized to generate an abstract for the videos. Abstractive summarization methods interpret and examine the text by using advanced natural language techniques in order to generate a new shorter text that conveys the most important information from the original text.

3. The Use of Video Captioning For Fostering Physical Activity

Summary: This paper proposes a video captioning framework that aims to describe the activities in a video and estimate a person's daily physical activity level. Large amounts of training data, deep learning has made a comeback providing breakthroughs on image recognition, and object detection with many applications. Object detection models are used to extract object information and localization from images and videos. The main contribution of this research is proposing a novel video captioning framework. This framework utilizes the Spatio-Temporal information in a video to generate accurate and coherent captions for filmed physical activities. The captions comprise temporal dynamics of discovered actions that could be used to follow a person's physical activity in daily life.

Methodology:

A. Video Captioning: Recent video captioning techniques often consists of two main parts; an encoder and a decoder.
   Encoder: This part focuses on detecting multiple events recorded in a video by jointly localizing temporal proposals of interest.

   Decoder: the decoder describes the events in natural language sentences.

B. Temporal Activity Proposals: This task generally provides fundamental tools and applications for action detection. For example, Temporal Action Detection is a tool that focuses on localizing the temporal extent of each action for a detected object.

C. Object Detection: Object detection is a computer vision task that tends to collect the related tasks to identify objects in an image or video frame by localizing and classifying them accurately. Deep learning is a must to detect objects here.

4. Key Clips and Key Frames Extraction of Videos Based on Deep Learning:

Summary: A large number of surveillance videos bring more tremendous storage pressure and also increase the difficulty. In this paper, the auto-encoder network has

been used to extract features of video frames. By comparing the feature differences between video frames, automatically select key video clips and keyframe images that contain helpful information to slim down the surveillance video.
Methodology:

A. Loss Function for the Network: The similarity image restoration, SSIM loss function, MS-SSIM loss function to measure the results.
B. Structure of the Network: In order to ensure the speed of extracting video features to reduce the amount of network calculation and the size of features as much as possible. Therefore, the effects of the network in terms of the number of channels, input size, and feature size of the network.
C. Training Method for the Network: After confirming the basic structure of the self-encoding network, we compared the training methods of the network in order to achieve better results on the current network structure.

Finally, the auto-encoder was completed with the goal of extracting key images and keyframes from a video.

5. Self-Supervised Learning to Detect Key Frames in Videos:

Summary: In this paper, we address the problem of automatically annotating and detecting key frames in a video. A video is represented as a sequence of continuous frames and the aim is to automatically annotate a set of frames of interest. By assuming there are training videos without annotated key frames, and the goal is to train a deep neural network that can automatically annotate key frames in training videos. For human action recognition, the key frames are the ones that can well represent the whole action, again for video summarization, the key frames are the set of frames that summarize the video, for video generation, the key frames are the first and last frames of the video clip to be generated, respectively.
Methodology:
Video Annotation Methods: Due to its time-consuming nature, various tools and strategies have emerged to facilitate the annotation task. For object detection in videos and related tasks, many popular annotation tools have been exploited such as ViPER, LabelMe etc.
Key Frame Detection: Earlier the concept of pipeline was used. Later works improved upon this pipeline by using keypoints detection for feature extraction, extracting local features via a SIFT descriptor and pooling the key points to find key frames in videos.
Deep learning: Bi-LSTM, GAN, CNN, LSTM, SLAM

6. Show and Tell: A Neural Image Caption Generator

Summary: We have presented NIC, an end-to-end neural network system that can automatically view an image and generate a reasonable description in plain English. NIC is based on a convolution neural network that encodes an image into a compact

representation, followed by a recurrent neural network that generates a corresponding sentence. The model is trained to maximize the likelihood of the sentence given the image.
Methodology: An “encoder” RNN reads the source sentence and transforms it into a rich fixed-length vector representation, which in turn in used as the initial
hidden state of a “decoder” RNN that generates the target Sentence. Here, they proposed to follow a recipe by replacing the encoder RNN by a deep convolutional neural network (CNN).

They use a CNN as an image “encoder”, by first pre-training it for an image classification task and using the last hidden layer as an input to the RNN decoder that generates sentences
LSTM-based Sentence Generator
Inference There are multiple approaches that can be used to generate a sentence given an image, with NIC. The first one is Sampling where we just sample the first word according to p1, then provide the corresponding embedding as input and sample p2, continuing like this until we sample the special end-of-sentence token or some maximum length. The second one is BeamSearch: iteratively consider the set of the k best sentences up to time t as candidates to generate sentences of size t + 1, and keep only the resulting best k of them.

7. Video Storytelling: Textual Summaries for Events:

Summary: The authors propose a context-aware framework for multimodal embedding learning, where we design a Residual Bidirectional Recurrent Neural Network to leverage contextual information from past and future. The multimodal embedding is then used to retrieve sentences for video clips. Second, we propose a Narrator model to select clips that are representative of the underlying storyline. The Narrator is formulated as a reinforcement learning agent which is trained by directly optimizing the textual metric of the generated story.
Methodology: A. Context-Aware Multimodal Embedding: Encoder: Video- LSTM and CNN. Sentence- RNN. B. Narrator Network: C. Narrator Training

8. Deep Learning-Based Short Story Generation for an Image Using the Encoder-Decoder Structure:

Summary: this study introduces an encoder-decoder framework structure to generate a short story captioning (SSCap) using a common image caption dataset and a manually collected story corpus. This manuscript has three main contributions, which include 1) an unsupervised deep learning-based framework that combines a recurrent neural network (RNN) structure and encoder-decoder model for composing a short story for an image 2) a huge story corpus, which includes two different genres (horror and romantic), manually collected and validated.

Methodology:

A. IMAGE CAPTION GENERATION: RCNN is applied to map the image regions and words into a joint, multimodal embedding in the encoder side. A recurrent neural network (RNN) is then implemented on the decoder side.

1.IMAGE REPRESENTATION - RCNN 2. SENTENCE REPRESENTATION-RNN

B. SKIP-GRAM SENTENCE ENCODER-DECODER