Minor Project Report on

Video Captioning With Attention-based LSTM

Prajna (16IT127)

Muthyam Satwik (16IT119)

Akshay Pandita(16IT151)

Under the Guidance of,

Mr. Dinesh Naik

Department of Information Technology, NITK Surathkal

Date of Submission: 11-04-2019

in partial fulfillment for the award of the degree

of

Bachelor of Technology

In

Information Technology

At

Department of Information Technology

National Institute of Technology Karnataka, Surathkal

April 2019

Department of Information Technology, NITK Surathkal
Minor Project
End Semester Evaluation Report (April 2020)

Course Code : IT 399

Course Title: Minor Project

Project Title: *Video Captioning With Attention-Based LSTM and Semantic Consistency*

Project Group:

Name of the Student	Register No.	Signature with Date
Prajna	16IT127	
Muthyam Satwik	16IT119	
Akshay Pandita	16IT151	

Place:NITK

Date:11-04-2019 *(Name and Signature of Minor Project Guide)*

Abstract

Recent progress in using long short-term memory (LSTM) for image captioning has motivated the exploration of their applications for video captioning. By taking a video as a sequence of features, an LSTM model is trained on video-sentence pairs and learns to associate a video to a sentence. However, most existing methods compress an entire video shot or frame into a static representation, without considering attention mechanism which allows for selecting salient features. Furthermore, existing approaches usually model the translating error, but ignore the correlations between sentence semantics and visual content. To tackle these issues, we propose a novel end-to-end framework named aLSTMs, an attention-based LSTM model with semantic consistency, to transfer videos to natural sentences. This framework integrates attention mechanism with LSTM to capture salient structures of video, and explores the correlation between multimodal representations (i.e., words and visual content) for generating sentences with rich semantic content. Specifically, we first propose an attention mechanism that uses the dynamic weighted sum of local two-dimensional convolutional neural network representations. Then, an LSTM decoder takes these visual features at time t and the word-embedding feature at time t-1 to generate important words. Finally, we use multimodal embedding to map the visual and sentence features into a joint space to guarantee the semantic consistence of the sentence description and the video visual content. Experiments on the benchmark datasets demonstrate that our method using single feature can achieve competitive or even better results than the state-of-the-art baselines for video captioning in both BLEU and METEOR

Keywords: *LSTM, aLSTM, CNN*

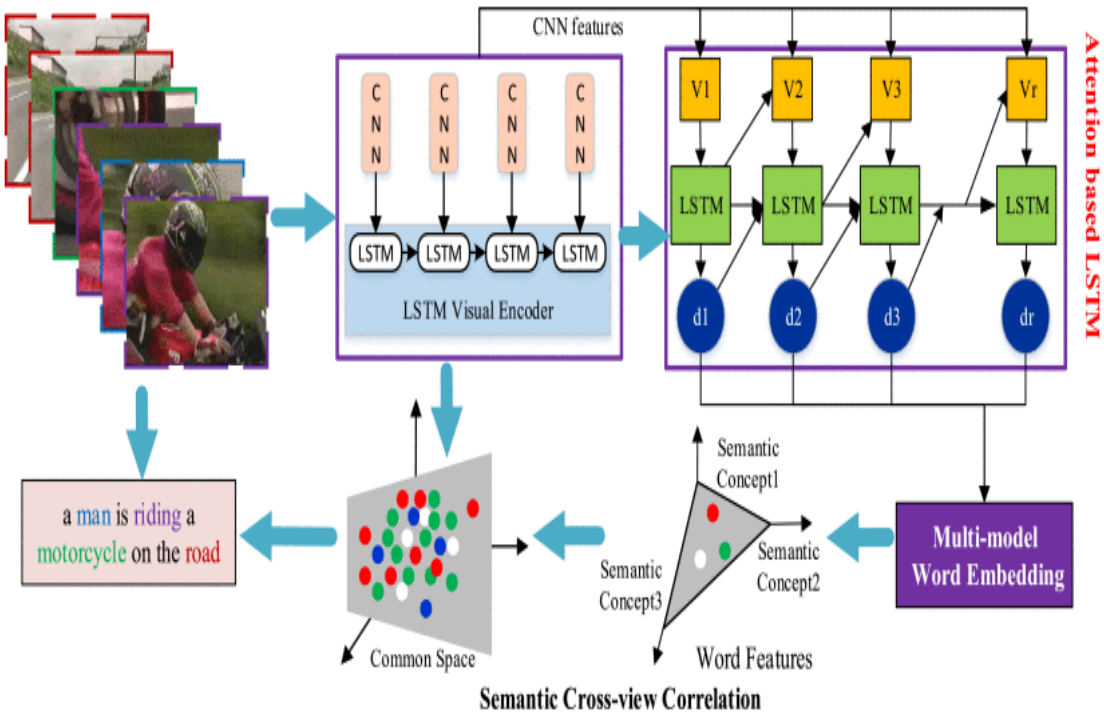
Contents

List of Figures

1 Introduction

Attention networks are currently a standard part of the deep learning toolkit, contributing to impressive results in neural machine translation, visual captioning , and question answering. This approach alleviates the bottleneck of compressing a source into a fixed-dimensional vector by equipping a model with a variable-length memory, thereby providing random access into the source is needed. In addition, attention is implemented as a hidden layer which computes a categorical distribution to make a soft-selection over source elements. Thus we incorporate an attention based LSTM model to capture salient temporal structures of videos. Recently, many applications were proposed to directly connect a visual convolution model to deep LSTM networks. Ideally, video description not only requires modeling and integrating their sequence dynamic temporal attention information into a natural language but also needs to take into account the relationship between sentence semantics and visual content, which to our knowledge has not been simultaneously considered.

Therefore, in this project we propose a unified framework, named aLSTMs, an attention-based LSTM model with semantic consistency. Firstly, to extract more meaningful spatial features, we adopt Inception-v3 neural network which is an extended version of GoogleNet. To exploit temporal information, we introduce one-layer LSTM visual encoder to encode those spatial 2D CNN feature vectors. Then we propose an attention mechanism which takes the dynamic weighted sum of local spatial 2D CNN feature vectors as the input for the LSTM decoder. Finally, we integrate multi-word embedding and cross-view methodology to project the generated words and the visual features into a common space to bridge the semantic gap between videos and the corresponding sentences.


CNN features
C N N
LSTM
LSTM Visual Encoder
V1
V2
V3
Vr
d1
d2
d3
dr
Attention based LSTM
a man is riding a motorcycle on the road
Common Space
Semantic Concept1
Semantic Concept2
Semantic Concept3
Word Features
Multi-model Word Embedding
Semantic Cross-view Correlation

2 Literature Survey

2.1 *Related Work*

1. Image/Video Recognition

   Recognition of image and video is a fundamental and challenging problem in computer vision. Dramatic progress has been achieved by supervised convolutional models on image-based action recognition task. Rapid progress has been made in the past few years, especially in image feature learning, and various pre-trained CNN models are proposed. However, such image based deep features cannot be directly applied to process videos due to the lack of dynamic information. Du et al. propose to learn spatio-temporal features using deep 3D CNN and shows good performance on various video analysis tasks. To effectively learn the spatial-temporal signals and features, Sun et al. propose a new deep architecture, called factorized spatio-temporal convolutional networks, which factorizes the original 3D spatio-temporal convolution kernel learning as a sequential process of learning 2D spatial kernels in the lower network layers. Thanks to the emergence of LSTM, it is able to model sequence data and learn patterns with wider range of temporal dependencies. Donahue et al. integrate CNN and LSTM to learn spatio-temporal information from videos. It extracts 2D CNN features from video frames and then the 2D CNN features are fed into a LSTM network to encode the videos temporal information.

2. Image/Video Captioning

   To further bridge the gap between video/image understanding and natural language processing, generating description for image or video becomes a hot research topic. It aims to generate a sentence to describe the image/video content. Due to the development of Recurrent Neural Network (RNN) and LSTM, researchers have striven to automatically describe an image/video with a correct and novel natural language sentence.

   Inspired by the advantages in multi-modal learning and machine translation, Ryan et al. construct a joint multi-modal embedding approach to project image features

extracted by a deep CNN model and text features encoded by a LSTM network to a common space. Then, a decoder is applied to decode image content into visual sentences using structure-content neural language model. In , they develop a so-called correlation component manifold space learning (CCMSL) to learn a common feature space by capturing the correlations between the heterogeneous databases. Karpathy et al. leverage dataset of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data. It firstly exploits two modalities through a multi-modal embedding to align regions of image and snippets from corresponding sentence. Next, it uses image-sentence and region-snippet pairs to train a multi-modal recurrent neural network to generate novel descriptions of regions and images. In a multi-modal Recurrent Neural Networks (m-RNN) model is proposed for image captioning, which directly models the probability of generating a word given previous words and images. Vinyals et al. propose an end-to-end neural network system to generate sentences for images via integrating LSTM and GoogleNet. In an attribute with high-level concepts is incorporated into a CNN-RNN network as the external input. This work provides a fully trainable attribute-based deep neural network, which yields significantly good performance. Justin et al. introduce a dense captioning approach, which not only detects object region proposals but also generalizes phrases and sentences to describe image region and full image content respectively. It simultaneously takes the object detection and description task into account, and proposes a Fully Convolutional Localization Network (FCLN) architecture that processes an image with a single efficient forward pass, requiring no external regions proposals. This network can be trained end-to-end with a single round of optimization.

Following image captioning, there are several researches focusing on video captioning. In it firstly proposes an end-to-end LSTM-based model for video-to-text generation. This work only leverages the local 2D CNN feature from frame-level, then performs mean pooling over 2D features across each video to form a fixed-dimensional video-level feature. Compared with image content, video has both spatial and temporal structure. In order to efficiently translate video to language, approaches should take both temporal and spatial information into account. In-

spired by this, an end-to-end sequence-to-sequence model is proposed to generate captions for videos. It incorporates a stacked LSTM which firstly reads the sequence of CNN outputs and then generates a sequence of words. Pan et al. propose a novel approach, namely Hierarchical Recurrent Neural Encoder (HRNE), which exploits multiple time-scale abstraction of the temporal information with two-layer LSTMs network.

Attention networks have proven to be an effective approach for embedding categorical inference within a deep neural network. This mechanism, which learns to automatically select the most relevant source data to generate output data, has made a great success in machine translation , visual question answering and video/image captioning. Since not all source words in a sentence are equally salient for machine translation, and also the generated word is usually relevant to a subset of source words, it is important for a model to identify the importances or weights of source words for translation. In two types of attention-based model are proposed for machine translation: a global mechanism in which all source words are attended, and a local one whereby only a subset of source words are considered at a time. Yang et al. introduce a multiple-layer stack attention neural network to answer questions according to an image. In two attention-based LSTM models are proposed. They are capable of well aligning the most relevant visual content to the next word of the sentence. In a set of visual concepts corresponding to each image are firstly obtained by running a set of attribute detectors. Next it learns to selectively attend to semantic concept proposals and feds them into hidden states of recurrent neural networks. Since not all frames in a video are equally salient for a short description generation and an event may last in multiple frames, it is important for a model to identify which frames are more salient. In they propose a content similarity based fast reference frame selection algorithm for reducing the computational complexity of the multiple reference frames based inter-frame prediction. In they propose a temporal attention based LSTM model which combines local temporal modeling to automatically select the most relevant temporal segments to generate the next word.

2.2 *Outcome of Literature Survey*

These are the outcomes expected from doing Literature Survey-

- Assessment of the current state of research on a topic.
- Identification of the experts on a particular topic.
- Identification of key questions about a topic that need further research.
- Determination of methodologies used in past studies of the same or similar topics.

Upon completion of the literature review, we got a solid foundation of knowledge in the area and a good feel for the direction for our project.

2.3 *Problem Statement*

Most of the existing methods compress an entire video shot or frame into a static representation, without considering attention mechanism which allows for selecting salient features. Furthermore, existing approaches usually model the translating error, but ignore the correlations between sentence semantics and visual content. To tackle these issues, we propose a novel end-to-end framework named aLSTMs, an attention-based LSTM model with semantic consistency, to transfer videos to natural sentences

2.4 *Objectives*

1. To generate the feature vectors using 2D convolutional neural network
2. To build vocabulary model
3. To build LSTM encoder-decoder model
4. Add attention to the encoder-decoder model

3 Methodology

Our task was to generate language sentences for videos. In this section, we first defined the terms and notations. Next, we introduce our aLSTMs approach. An objective function is built by integrating two loss functions which simultaneously consider video translation and semantic consistency. Specifically, one loss function aims to guarantee the translation from videos to words, while another loss function tries to bridge the semantic gap with semantic cross-view correlations. The detailed information about solution is given as well.

1. Terms and Notations

   Suppose we have a video V to be described by a textual sentence D=d1,,dNd consisting of Nd words. Let X=x1,,xNxRMNx and D=d1,,dNdRLNd denote the visual and the textual features, where di=Edi is the word representation of a single word di and Nx is the total number of feature vectors. Let M and L denote the dimension of visual feature and textual feature. X is extracted using deep neural networks, which will be described in the experiment.

2. Attention-Based Long Short-Term Memory Decoder

   To date, modeling sequence data with recurrent neural network has been proven successful in the process of machine translation, speech recognition, image/video captioning etc. However, it is still difficult to train a standard RNN due to the vanishing gradient problem. LSTM, an updated version of standard RNN, solved this issue by learning patterns with wider range of temporal dependencies. As videos and natural sentences are both sequential data, LSTM is applied as the basic component for our aLSTMs.

   The main idea of Attention-based Long Short-Term Memory is to integrate attention mechanism into the LSTM. A basic LSTM unit consists of a single memory cell, an input activation function, and three gates (input it, forget ft and output ot). it allows incoming signal to alter the state of the memory cell or block it. ft controls what to be remembered and what to be forgotten by the cell and somehow can avoid the gradient from vanishing or exploding when back propagating through time. Finally, ot allows the state of the memory cell to have an effect on other

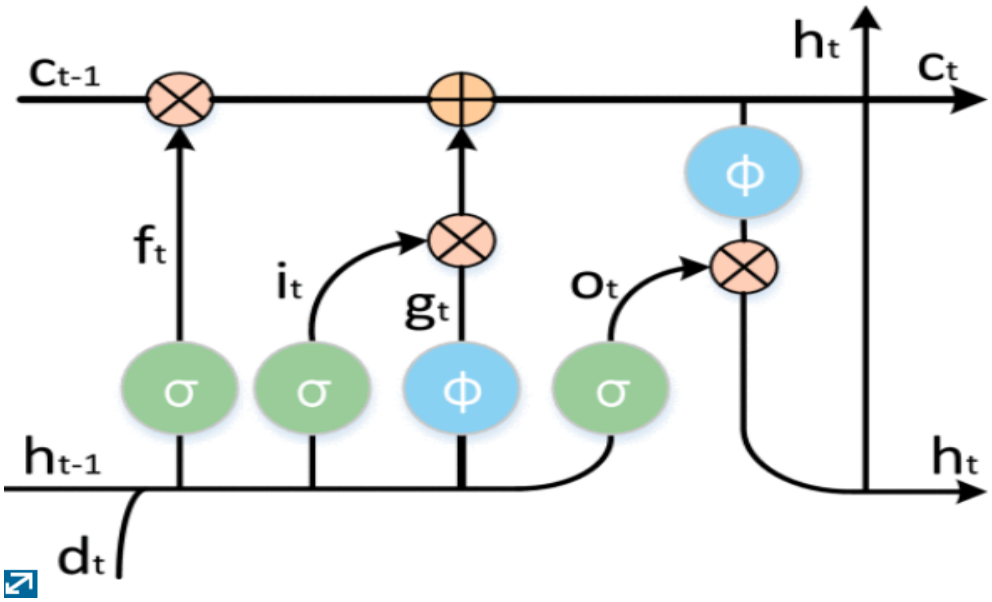
neurons or prevent it. Basically, the memory cell and gates in a LSTM block are defined as follows:

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \phi \end{pmatrix} Z_{L+r,r} \begin{pmatrix} \mathbf{E}d_{t-1} \\ \mathbf{h}_{t-1} \end{pmatrix}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \phi(\mathbf{c}_t)$$

where E denotes an embedding matrix, represents the logistic sigmoid non-linear activation function mapping real numbers to (0,1) and can be thought as knobs that LSTM learns to selectively forget its memory or accept current input, denotes the hyperbolic tangent function tanh, is the element-wise product with the gate value, ZL+r,r denotes the parameters of the LSTM. Let L and r denote the embedding and LSTM dimensionality respectively.



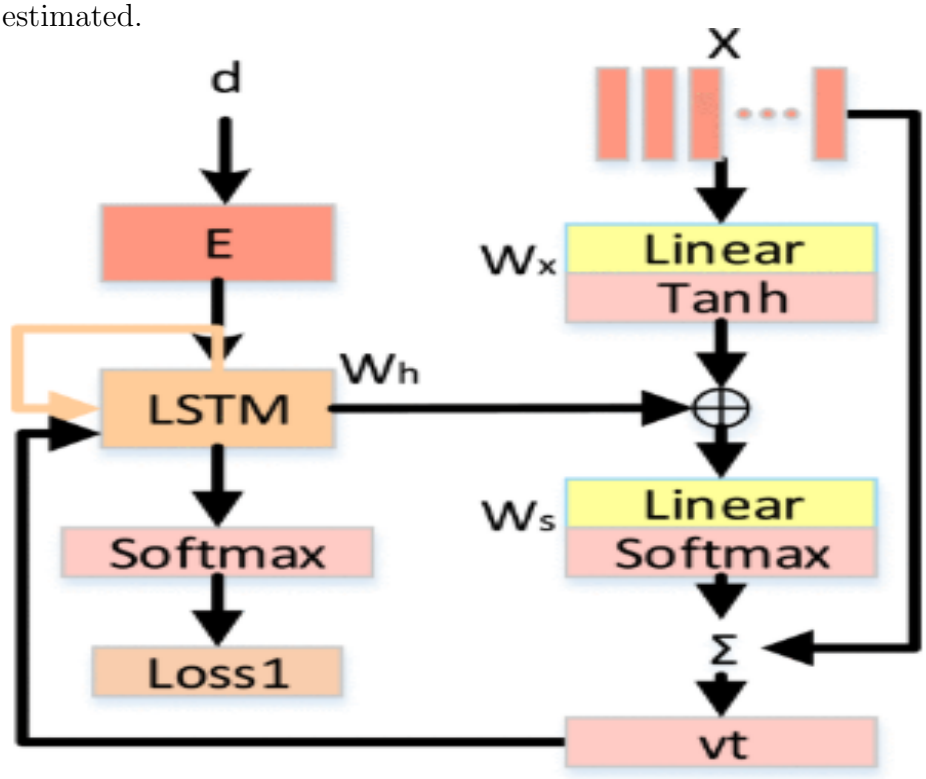
Compared with images, videos contain more complex temporal information which should be aligned to language data. Thus we extend the attention mechanism introduced by to support video captioning. The new form of LSTM is defined as:

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \phi \end{pmatrix} Z_{L+r+M,r} \begin{pmatrix} \mathbf{E}d_{t-1} \\ \mathbf{h}_{t-1} \\ \mathbf{v}_t \end{pmatrix}$$

$$\mathbf{v}_t = \sum_{i=1}^{N_x} \beta_i^t \mathbf{x}_i, \quad s_i^t = W_s \phi \left(W_h \mathbf{h}_{t-1} + W_x \mathbf{x}_i + b_s \right)$$

$$\beta_i^t = \frac{\exp\left(s_k^t\right)}{\sum_{k=1}^{N_x} \exp\left(s_k^t\right)}, \text{s.t,} \quad \sum_{i=1}^{N} \beta_i^t = 1$$

where vt represents context vector which is a dynamic representation of the relevant representation of the video input at time t. M is the dimension of vt . In addition, ti is the attention weights at time t describing the relevance of the i-th feature in the input video. Given the previous hidden state ht1 of the LSTM decoder and the i-th video feature, it returns the unnormalized relevance score sti. Once the relevance scores for all the features X=x1,,xNx are computed, the LSTM is able to obtain ti at each time step t. The Ws , Wh, Wx and bs are the parameters to be estimated.



In addition, to capture rich temporal information, we introduce an one-layer LSTM visual encoder, called LSTM Visual Encoder. The LSTM recently has made a great success in the process of action recognition. Inspired by this, in our framework we propose to integrate the updated GoogleNet with one-layer LSTM visual encoder to encode video temporal information. Specifically, the last output of the LSTM Visual Encoder is used to initialize the first LSTM unit of our attention based LSTM

network to facilitate video captioning.

3. Loss : Translation From Videos to Words

   In the LSTM decoding phase, the LSTM computes context vector vt given an input sequence X=x1,,xNx and the hidden state ht1. Inspired by the principle of translating images, we treat the activation value indexed by a training word dt in the softmax layer of our sentence generator as the likelihood of generating that word

   P(dt—vt;d1,d2,,dt1;E)

   The cost of generating that training word is then defined as the negative logarithm of the likelihood. We further define the cost of generating the words as

   Loss1=t=1Ndlog(P(dt—vt;d1,d2,,dt1;E))

   where Nd denotes the total number of words in sentence and di denotes the i-th word in sentence D. By minimizing Loss1, the contextual relationship among the words in the sentence can be guaranteed, making the sentence coherent and smooth.

4 Work Done

1. For encoding and decoding, the word2vec function from gensim library is used.
2. Then feature vectors are extracted from video frames using sequential model.
3. Then they are fed to LSTM encoder-decoder model built using keras library using tensorflow in the backend.
4. Then video captions are decoded using the vocabulary model built using word2vec.

Dataset used (Video Captoning Dataset):-

The Video Captioning Dataset contains 100K short-videos and 120K sentences describing visual content of the short-videoss. The short-videos have been collected from Tumblr, from randomly selected posts published between May and June of 2015. The sentences are collected via crowdsourcing, with a carefully designed annotation interface that ensures high quality dataset. The dataset shall be used to evaluate video description techniques.

5 Results and Analysis

Result:-

Actual Caption	Predicted Caption
the lady appealed to the man	rotated dobermin gopy ater scraped
a man is packing a food	Hardpaper food scraped
someone is folding a piece of paper	Hardpaper rotated scraped

Analysis:-

1. No. of LSTM units used = 64
2. No. of epochs = 30
3. Loss = mean squared error
4. Optimizer = adadelta

Training on 2638 samples, validate on 10 samples(Next Page)

Epoch no:-	Improved Form	Improved To	Loss
1	inf	0.3326	0.3041
2	0.3326	0.33232	0.3038
3	0.33232	0.33204	0.3035
4	0.33204	0.33176	0.3029
5	0.33176	0.33149	0.3029
6	0.33149	0.33123	0.3026
7	0.33123	0.33097	0.3023
8	0.33097	0.33073	0.3021
9	0.33073	0.33050	0.3018
10	0.33050	0.33028	0.3016
11	0.33028	0.33008	0.3013
12	0.33008	0.32990	0.3011
13	0.32990	0.32973	0.3009
14	0.32973	0.32957	0.3007
15	0.32957	0.32941	0.3006
16	0.32941	0.32926	0.3004
17	0.32926	0.32911	0.3002
18	0.32911	0.32896	0.3001
19	0.32896	0.32880	0.2999
20	0.32880	0.32864	0.2997
21	0.32864	0.32848	0.2995
22	0.32848	0.32831	0.2994
23	0.32831	0.32813	0.2992
24	0.32813	0.32794	0.2990

6 Conclusion & Future Work

In this project, we have built a framework aLSTMs, which is implemented by simultaneously minimizing the relevance loss and semantic cross-view loss. On two popular video description datasets, the results of our experiments demonstrate the success of our approach, which achieves comparable or even superior performance compared with the current state-of-the-art models. In the future, we will modify our model to work on domain-specific datasets, e.g., movies.

References

[1] J. Song et al *Optimized graph learning using partial tags and multiple features for image and video annotation*. IEEE Trans. Image Process., vol. 25, no. 11, pp. 4999-5011, Nov. 2016.

[2] X. Zhu, Z. Huang, J. Cui, H. T. Shen *Video-to-shot tag propagation by graph sparse group lasso*. (German) IEEE Trans. Multimedia, vol. 15, no. 3, pp. 633-646, Apr. 2013.

[3] Z. Pan, Y. Zhang, S. Kwong *Efficient motion and disparity estimation optimization for low complexity multiview video coding* IEEE Trans. Broadcast., vol. 61, no. 2, pp. 166-176, Jun. 2015.