Indian Institute of Technology, Bombay

Seminar Report

Self-supervision for multi-modal representation learning

"Under the guidance of
Dr. Preethi Jyothi and Dr. Ganesh Ramakrishnan"

June 4, 2020

Submitted by

Jayaprakash A

193050050
M. Tech
Computer Science and Engineering
Indian Institute of Technology, Bombay

Acknowledgement

I would like to humbly acknowledge with sincere gratitude, my guide **Prof. Preethi Jyothi** and my co-guide **Prof. Ganesh Ramakrishnan** for giving me an opportunity to work under their mentorship. I am really thankful to them for their able guidance and support throughout the course timeline. I would also like to thank my peers **Rakesh Prasanth**, **Mayur Warialani**, **Abhishek Thakur**, **Mayank Kothyari** and **Varad Bhatnagar** for the healthy conversations we had throughout the process. The discussions and meetings helped me understand all the basic concepts of self-supervision and some other machine learning ideas. I would like to thank my family for supporting me in my life.

Contents

List of Figures

List of Tables

Abstract

Video analysis comprises a large variety of tasks namely caption generation, video segment retrieval, action classification, caption alignment and so on. For all these tasks the most elementary and important aspect is the input representation. Better representations always yield better results. One needs to capture the essence of various channels of information present in video ie, audio and visual features. Better mechanisms are required to represent the input features.

In our task, using supervised learning to learn input representations is quite infeasible as they require huge amounts of labelled training data. Nevertheless we can use the huge unlabelled data in a supervised manner by predicting/regressing over required information, we can achieve autonomous labelling. These techniques called self-supervised learning can be helpful in learning the required multi-modal representations.

Chapter 1

Background

Video Captioning involves generating natural language description of video. Let us first understand how a human being does this. One uses the visual and audio content given to him and also his previous knowledge to analyse them and then generates a caption.

Like in any machine learning task the most important thing is feature representation. How do we represent the text/audio/video/image in machine understandable format? We need a lot of training data to learn better embedding for our input. But collecting such data is very costly. How can make use of the existing data to learn better representation? All these questions gave birth to the concept of self-supervised learning.

As discussed earlier, human beings are not given huge image and caption pairs to be able to do captioning. Image and text are analysed by human brain and using this as prior knowledge humans are able to achieve the task. Self-supervised learning is roughly what we have described now. We will discuss in great detail in the subsequent chapters about self-supervision and its application to learn better visual/audio/textual representations that can be used for video captioning setup.

Chapter 2

Introduction

A machine learning task can be simply stated as predicting the unknown using known data. We shall now discuss about the main basic ingredient of machine learning model ie, input feature representations and how we can efficiently generate features.

2.1 Multi modal representations

In the recent trends, machine learning has evolved several folds. The computers are now able to analyse images and retrieve image based on language description. They can understand the audio files and try to predict the emotion of the person and this list goes on. For all these activities the machine needs to analyse the input. The machines require some representation of the input which they can analyse.

In the context of human–computer interaction, a modality is the classification of a single independent channel of sensory input/output between a computer and a human. In our setup of video analysis, we break down the video to leverage understanding of the input. Video possesses multiple modalities like vision, audio, motion and orientation too. In this report, we provide insights about how to learn useful representations of input that consists of multiple modalities. [5]

2.2 Self supervised learning

2.2.1 Why do we need self-supervision?

For a given task and enough labelled data, we can solve the problem quite efficiently with supervised learning. But it is not always easy to collect labelled data. Manual collection and labelling is quite expensive and also poses scalability issues. But we have a lot of unlabelled data around us. Is there nothing that we can do with such data? The answer is no! We can use this unlabelled data and try to extract the essential meta data about the data and in way autonomously label them. These approaches are what we call self-supervised learning methods.

2.2.2 What is self-supervision?

Self supervised learning is a recent learning technique where the data is not labelled by human interference. The dataset is autonomously labelled. They are labelled by finding relations or correlation between different input signals. In general most self-supervised tasks involve masking/holding back some part of information and the network tries to predict the masked out versions with help from unmasked data. In this way the network tries to learn about the semantic relationship between the data.

Self supervised learning approaches tries to extract the naturally inhibited context and metadata and use them for downstream tasks. For example, to understand the context of a word say ‘set’, one does not require data with its usage and its context. We can learn about the context when it appears near words like ‘tennis’, ‘badminton’, ‘numbers’, ‘data structure’ or ‘alarm’. In a way it is similar as to how human brains learns autonomously. [6]. We do not focus on what is this final performance of this task, but we are rather interested in the intermediate representations that we can make use of.

In this report, we have mostly dealt with self-supervision based approaches for image captioning. We shall discuss about how self-supervised based approaches for multiple modalities can provide better feature representations for our setup. But let us first begin with self-supervised objectives in case of single modality in Chapter 3 and extend the idea further to multiple modalities in Chapter 4.

Chapter 3

Self-supervision for single modality

In this chapter we shall discuss about single modality models that use self-supervised based objectives. We will first discuss about the BERT (Bidirectional Encoder Representations from Transformers) architecture. This model is based on self-supervised objectives to learn better embeddings for the text modality. We will then continue the discussion with PASE (Problem Agnostic Speech Encoder). PASE is used to learn representations for the audio samples.

Unlike the previous language models based on RNNs, BERT [2] is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The underlying architecture used by BERT is based on transformers [1].

3.1 Transformers

The transformers architecture follows an encoder-decoder framework, where encoder maps an input sequence $(x_1, \ldots, x_n)$ to $(z_1, \ldots, z_n)$. The decoder take z as input and generates an output sequence $(y_1, \ldots, y_m)$. Contrary to other encoder-decoder based models where attention is unidirectional (i.e, either left or right attended), transformers make use of self-attention and fully connected layers in both encoder and decoder.

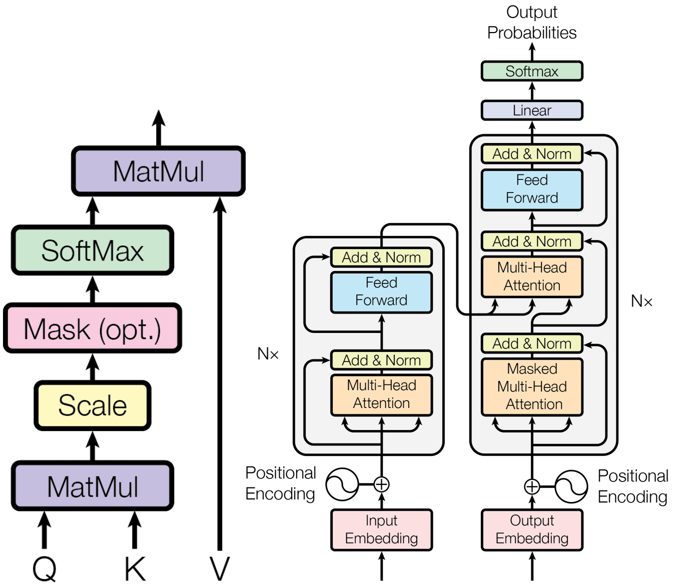
The encoder is composed of identical layers where each layer has a self-attention and feed forward sub layers. The decoder is also composed of identical layers where in each layer we have an additional attention layer(compared to encoder) over the output of encoder layer. As the encoder and decoders have no recurrence in them they use sinusoidal based positional embeddings to inject information about the relative position.

The self-attention is called scaled dot product attention. The input includes queries(Q), keys(K) of dimension d_k and values(V) of dimension d_v.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.1}$$

In the attention layers for both encoder and decoder the three vectors(Q, K, V) are the output of previous layer. Each position can attend to all other positions in the previous layer. However in the encoder-decoder attention layer, queries can from the previous decoder layer and key, value vectors from output of the encoder. This allows every position in the decoder to attend over all positions in the input sequence.

In this way they tried to make use of bidirectional information to learn representations.



(a) Self attention layer

(b) Transformer encoder

Figure 3.1: Transformers [1]

3.2 BERT model

The BERT framework uses transformers as the base architecture. The training process is divided into two tasks: *pre-training* and *fine tuning*.

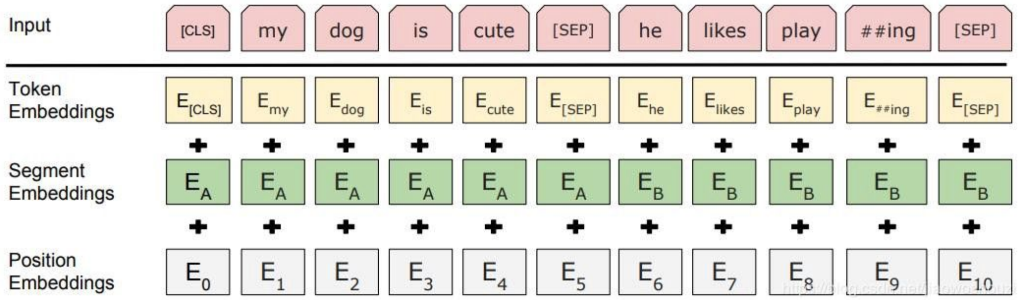

Input
[CLS]
my
dog
is
cute
[SEP]
he
likes
play
##ing
[SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

Figure 3.2: Input representation for BERT [2]

3.2.1 Pre training

Pre training of BERT involves two unsupervised tasks **a. Masked Language Modelling** and **b. Next Sentence Prediction**. The model is trained to minimise a loss based on the both approaches simultaneously.

3.2.1.1 Masked Language Model

Intuitively we can infer that a deep bidirectional model is likely to be more powerful than unidirectional models. But instead, such model has very high bias towards the current word/text, so not much information can be extracted. So in an attempt to get better representations, some parts of text are randomly masked and they are predicted with the context of the remaining unmasked words using cross entropy loss.

3.2.1.2 Next Sentence Prediction

The language modelling task captures the inherent information of word in the context of its sentence. The representation learnt thus far captures the sentence level semantic information. But cannot differentiate between sentences across a big text. So, the paper suggests a next sentence prediction task, where the classifier tries to distinguish if the two sentences are correlated or different. These kind of tasks are useful for downstream tasks like *Question Answering* and *Natural Language Inference*.

3.2.2 Fine tuning

For each NLP downstream task, we simply have a few additional layers and train to minimise such loses on a labelled data. Since our model has been pre-trained extensively, fine tuning requires very little effort and resources too.

With the help of fine tuning, the model was evaluated on eleven different NLP tasks and it has achieved state of the art results. The greatness of the model lies in the fact that we have unified model with very slight variations and yet beat/match the state of the art results. The results were exceptional even in low resource settings. The model was highly successful in demonstrating the role and importance of bidirectional architectures combined with self-supervised techniques.

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4
BERT	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**

Table 3.1: Results on the GLUE benchmark. MNLI: MultiNLI Mismatched, QQP: Quora Question Pairs, QNLI: Question Natural Language Inference and CoLA: The Corpus of Linguistic Acceptability [2]

3.3 Problem Agnostic Speech Encoder(PASE)

PASE [4] is another model that makes use of self-supervised objectives to learn better embeddings for the audio modality. Unlike text, speech signals are high dimensional, long, and variable-length sequences, but also entail a complex hierarchical structure. It is thus hard to find a single self-supervised task that can learn general and meaningful representations able to capture this latent structure. PASE mitigates this issue by jointly training on seven different self-supervised objectives using a ensemble of neural networks. PASE encoder is based on SincNet model.

3.3.1 SincNet

The SincNet [3] convolves the waveform with a set of parametrized *sinc* functions ($\frac{\sin x}{x}$) that implement rectangular band-pass filters. The low and high cut-off frequencies are the only parameters of the filter learned from data. This mode offers considerable flexibility, but forces the network to focus on high-level tunable parameters with broad impact on the shape and bandwidth of the resulting filter. First layer of a standard CNN performs a set of time domain convolutions between the input waveform and some **Finite Impulse Response (FIR)** filters.

$$y[n] = x[n] * h[n] = \sum_{l=0}^{L-1} x[l] \cdot h[n-l] \tag{3.2}$$

where $x[n]$ is a chunk of the speech signal, $h[n]$ is the filter of length L, and $y[n]$ is the filtered output. So instead of learning all the L filter variables we use a function g and try to learn θ.

$$g[n, f_1, f_2] = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \tag{3.3}$$

where $\theta = [f_1, f_2]$, where f_1 and f_2 are the learned low and high cutoff frequencies. SincNet performs the convolution of the raw input waveform with a set of parameterized sinc functions that implement rectangular band-pass filters.

3.3.2 PASE Encoder

The PASE encoder is based on the SincNet model. The subsequent layers are composed of seven convolutional layers that perform time domain convolutions and the input signal is decimated by a factor of 160. The seven different self-supervised objectives are performed using small neural networks called **workers**.

3.3.3 Workers

The workers are fed the encoded representations and solve the seven objectives that include four regression and three binary discrimination tasks. The workers break the audio signal

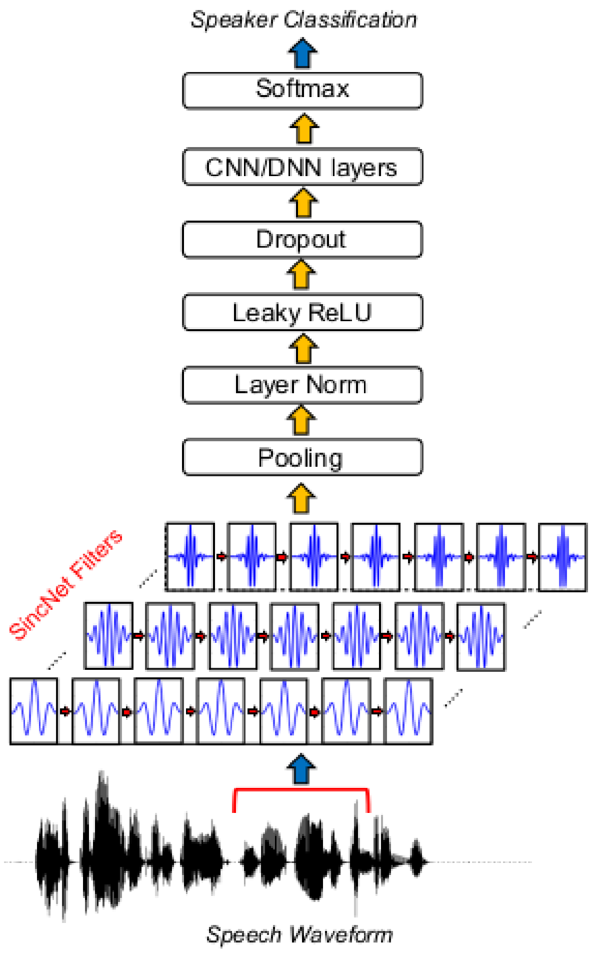


Figure 3.3: SincNet architecture [3]

to various components in an increasing level of abstraction.

The regression workers are trained to minimise the loss between target features and the network predictions. These tasks include prediction of **Waveform**, **Log power spectrum(LPS)**, **Mel-frequency cepstral coefficients (MFCC)** and **Prosody**. The waveform worker tries to reconstruct the waveform based on L1 loss. LPS and MFCC features are the standard features used in any signal processing task. Prosody involves four basic features namely interpolated logarithm of the fundamental frequency, voiced/unvoiced probability, zero-crossing rate, and energy. The prosody features inherit information that is helpful in emotion recogniton.

The binary discrimination involves sampling of anchor, positive and negative sample from already existing collection of PASE features. The workers then try to discriminate between positive and negative samples using binary cross entropy loss. The three different ways of sampling anchor(x_a), positive(x_p) and negative(x_n) samples are as follows

- **Local Info Max(LIM)** involves positive sample from the same sentence of the anchor and a negative sample from another random sentence that likely belongs to a different speaker.

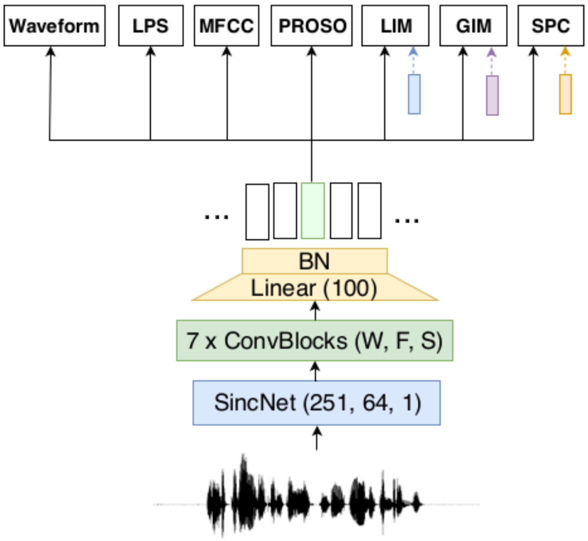


Figure 3.4: PASE encoder architecture [4]

- **Global Info Max(GIM)** The anchor representation is obtained by averaging the PASE encoded frames of a random utterance within a long random chunk of 1 s. The positive sample is derived from another random chunk within the same sentence, while the negative one is obtained from another sentence.

- **Sequence Predicting Coding(SPC)** has a single fixed anchor frame while positive samples are the next 5 consecutive frames and negative samples are the previous 5 frames outside the receptive field of the current frame(150 ms).

3.3.4 Self-supervised training and results

Encoder as well as all the workers are jointly trained by back propagating the total loss of all the workers. To balance the contribution of each worker, the outputs of workers are standardised based on validation statistics. Fine tuning the model with the help of few additional layers for each downstream task, PASE has shown better results as shown in the table 3.2

Model	**Classification accuracy (%)**					
	Speaker-ID		**Emotion**		**ASR**	
	MLP	**RNN**	**MLP**	**RNN**	**MLP**	**RNN**
MFCC	96.9	72.3	90.8	91.1	81.1	84.8
FBANK	98.4	75.1	94.1	92.8	80.9	85.1
PASE-FineTuned	99.3	97.2	97.7	97.0	82.9	85.3

Table 3.2: Comparing results obtained with PASE with state of the art models on few audio based tasks [4]

3.4 Discussion

We have seen that self-supervised based objectives in the case of text and audio modalities have provided ground breaking results. The problem of excessive labelled data has been mitigated, but at the expense of training time which is not much of a issue. We have also found that an unified model with minor changes can give almost state of the art results. One can make use of these architectures while working with tasks that use single modality.

However in our case of video captioning we have three different modalities namely audio, vision and text. So in the next chapter we shall discuss about self-supervised objectives in case of multiple modalities and how we can make use of such ideas.

Chapter 4

Self-supervision for multiple modalities

4.1 Introduction

Self supervised representation learning is not just limited to single modality. One can extend the idea to learn from multiple modalities at the same time. In this chapter we will review the existing work in this field. We shall mostly encounter work from image captioning field. Before we start, let us discuss about object detection models and S3D features that would help in understanding these ideas.

4.1.1 Object detection

For the extraction of image features, Faster R-CNN based object detection model has been used. Faster R-CNN [7] is a recent object detection model based on convolution neural networks. It is composed of three parts.

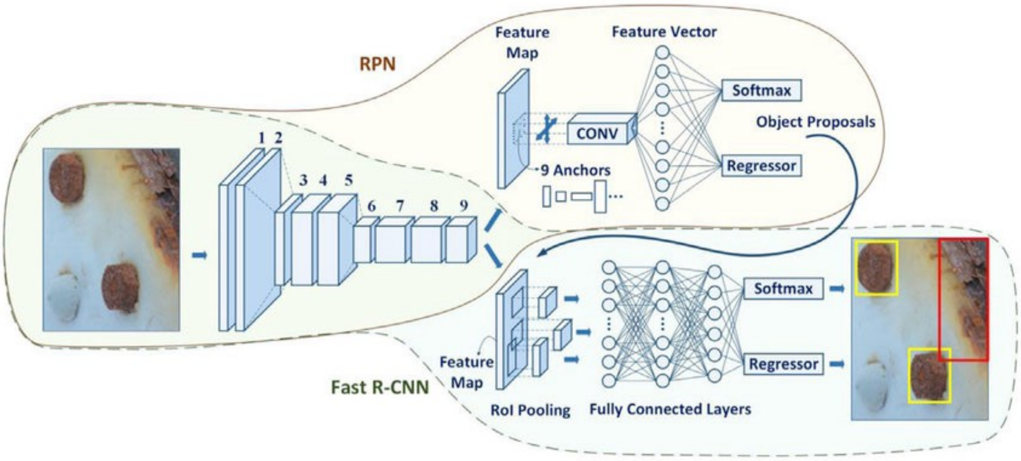


Figure 4.1: Faster R-CNN architecture

1. **Convolution layers** These layers try to learn filters to extract suitable features like colours, shapes that are essential for object detection.

2. **Region Proposal Network (RPN)** RPN is small neural network that uses feature map of the convolution layers and predict whether there is an object or not and also predict the bounding box of such objects.

3. **Classes and Bounding Boxes prediction** A Fully connected neural network that takes as an inupt the regions proposed by the RPN and predict object class.

4.1.2 S3D features

Video analysis has tasted success by using convolutional neural networks. However the gains are not significant compared to static image analysis. The existing 3d convolutional networks say I3D [8] jointly convolve over both spatial and temporal domain.

Recent developments in this field have considered replacing 3D convolutions with spatial and temporal separable 3D convolutions, i.e., replace filters of the form $k_t * k * k$ by $1 * k * k$ followed by $k_t * 1 * 1$, where k_t is the width of the filter in time, and k is the height/width of the filter in space. The resulting model is named S3D [9], which stands for “separable 3D CNN”. S3D has fewer parameters to learn and has shown significant gains in tasks like video classification over the existing architectures.

4.1.3 Optimal Transport(OT)

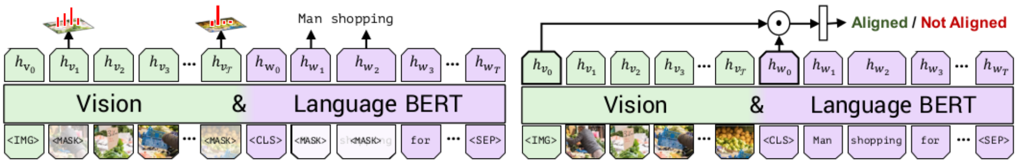
In one of the models to be discussed we use OT loss. So we will discuss in brief about transportation problem. The problem of transportation or distribution arises due to shipment of goods to the destination of their requirement from various sources of the origin [10]. Suppose that there are m sources and n destinations. Let s_i be the number of supply units available at source i (i = 1, 2,..., m) and let d_j be the number of demand units required at destination j (j = 1, 2, ..., n). Let c_{ij} represent the unit transportation cost for transporting the units from sources i to destination j. Let x_{ij} be the units shipped from supply point i to demand destination j. The objective is to determine the number of units to be transported from source i to destination j so that the total transportation cost is minimum. This problem can be expressed as linear program. Optimal solution is the one that minimises the cost.

$$\begin{array}{llll}
\min & \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} & & \\
\text{s.t.} & \sum_{j=1}^{n} x_{ij} & \leq s_i & \forall i = 1,\dots,m \\
 & \sum_{i=1}^{m} x_{ij} & = d_j & \forall j = 1,\dots,n \\
 & x_{ij} & \geq 0 & \forall j = 1,\dots,n; \forall i = 1,\dots,m
\end{array}$$

4.2 Vision-and-Language BERT(ViLBERT)

ViLBERT [11] is a model that tries to learn task agnostic joint representation for image and text modality. ViLBERT extends the idea of BERT to two stream model processing the two modalities separately and interacting them with the help of co-attention layers.

The input for the text stream is the sum of text embedding and its positional encoding. However for the image stream, we use the object detection features extracted using Faster R-CNN model along with spatial location embedded in it. The text stream and input are passed a through a series of transformer blocks(similar to 3.1) before the co-attention layer.


Man shopping
Aligned / Not Aligned
h_{v_0} h_{v_1} h_{v_2} h_{v_3} ... $h_{v_\mathcal{T}}$ h_{w_0} h_{w_1} h_{w_2} h_{w_3} ... h_{w_T}
Vision & Language BERT
<IMG> <MASK> <MASK> <CLS> <MASK> <MASK> for <SEP>
<IMG> <CLS> Man shopping for <SEP>

(a) Masked multi-modal learning (b) Multi-modal alignment prediction

Figure 4.2: ViLBERT architecture

4.2.1 Co-attention transformer block

It is very similar to the standard block. Given intermediate representation $H_v{}^{(i)}$ and $H_w{}^{(j)}$ of the visual and text modalities, the query, key and values vectors are computed as in standard transformer.(**Note.** The difference in intermediate representation i and j is because the image features are already pre-trained and hence require limited context aggregation compared to text features). However the key and value from each modality are passed as input to the other modality's attention block as shown in the figure 4.3.

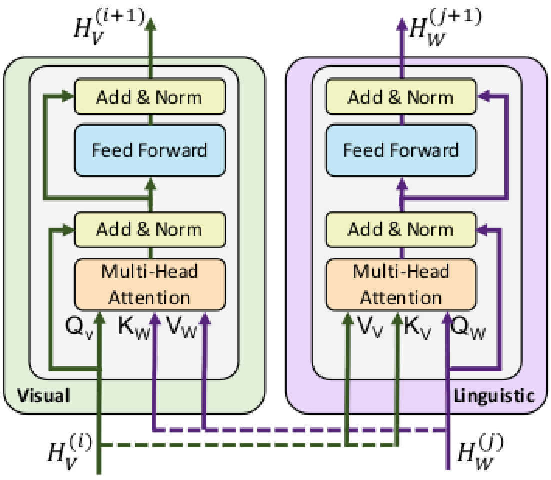


Figure 4.3: Co-attention layer

4.2.2 Training

The pre-training objectives are also very much similar to the ones in BERT. We have two primary objectives

- **Masked multi-modal modelling** The masked multi-modal modelling task follows from the masked language modelling task in standard BERT – masking regions of both image and text streams inputs and tasking the model with reconstructing them with the help of context from both modalities. Rather than directly regressing the masked feature values, the model instead predicts a distribution over semantic classes for the corresponding image region. The model is trained to minimize the KL divergence between these two distributions.

- **Multi-modal alignment prediction** In the multi-modal alignment task imitates the next sentence prediction task, in which an image-text pair is given as input and the model predicts whether the image and text are aligned, i.e. whether the text describes the image.

4.2.3 Discussion

The model is trained with these two tasks and later fine tuned accordingly. We shall discuss the results of the models in section 4.5. The model has performed well compared to SOTA (State of the Art). However there are certain points that we need to explore further. The image features are Faster R-CNN based features, the co-attention between the two modalities is at a very high level of abstraction. Perhaps one can improve the results by having co-attention at a smaller level of abstraction.

Also the masking of image and text features are done at random, so at times masked image region might be the one corresponding to the masked text. In such cases, there is clear misalignment between the two modalities.

4.3 Learning Cross-Modality Encoder Representations from Transformers (LXMERT)

LXMERT [12] framework tries to learn alignment and relationships between visual features and language semantics. LXMERT model consists of two types of encoder, self modality encoder one for each modality (language encoder and object-relationship encoder) and cross modality encoder. The single modality encoders are the same one as discussed in 3.2

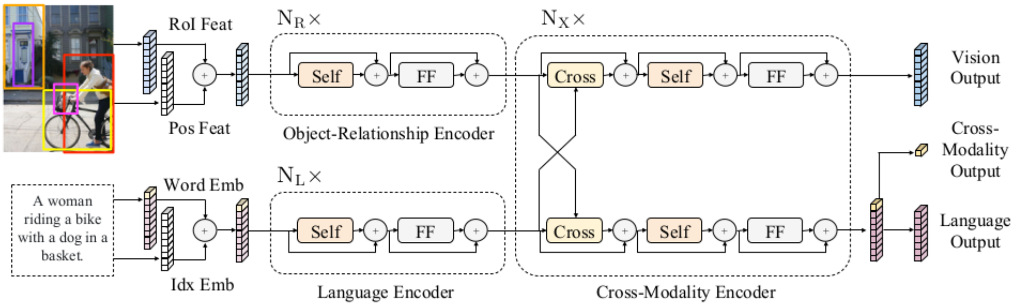


Figure 4.4: LXMERT architecture

4.3.1 Cross modality encoder

The cross modality encoder is based on cross attention layer. It consists of two self attention layers and one bidirectional cross attention layer followed by FF sub layers as shown in image 4.5. The bidirectional cross between the two modalities is a mixture of two unidirectional layers from vision to language and one from language to vision. The context vectors from one modality is given as input to the cross-attention sub layer of the other modality.

LXMERT is very much similar to ViLBERT in the input preprocessing section. Both the text embeddings and visual features are similar. They use object detection features along with bounding box positional encoding as the input to their object relationship encoder and text with its positional encoding as input for language encoder.

4.3.2 Training

The pre-trained objectives are analogous to the earlier approaches seen with slight modifications. The pre-training involves three main tasks.

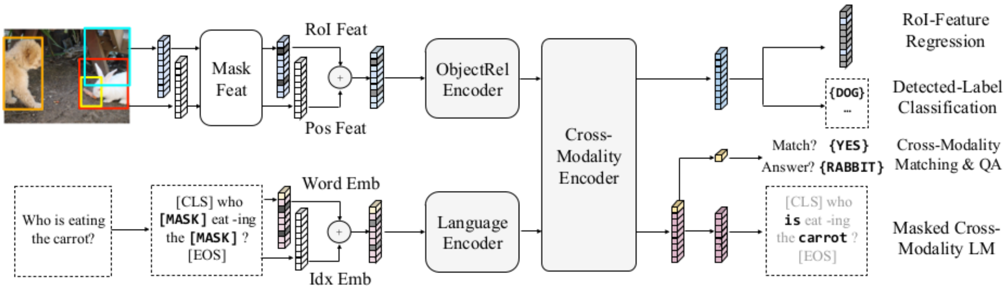


Figure 4.5: Pre-training objectives visualisation

- **Language Task: Masked Cross-Modality LM** Similar to the one in BERT, only that the masked word is now predicted with the help of context from both modalities.
- **Vision Task: Masked Object Prediction** Some objects are randomly masked, and the model tries to learn about it using two sub tasks. RoI-Feature(Region of interest) Regression regresses the object RoI feature with L2 loss, and Detected Label Classification learns the labels of masked objects with cross-entropy loss.
- **Cross-Modality Tasks** It involves two sub tasks that involve both the modalities and help in understanding the mutual relationship between them. Cross modality matching where classifier tries to check if sentence and image match. Image Question Answering sub task that also help to learn cross modality representation.

4.3.3 Discussion

The model is not much different to the one discussed earlier. Also the shortcomings faced earlier have not been resolved. The model however did outperform ViLBERT. We shall discuss about the success of this model in the analysis section 4.5.

4.4 UNiversal Image-TExt Representations (UNITER)

UNITER [13] is a large scale pre-trained model for multimodal embedding. It uses transformer as the core for its model, to leverage its elegant self-attention mechanism designed for learning contextualized representations. Unlike the previous two models that used two different encoders for both the modalities followed by cross modality encoder, UNITER has a different take on the problem.

UNITER takes both the image (object detection) and text embeddings along with positional encodings as input and projects them onto the same unified embedding space. Now using the standard encoder, they have pre-trained the model using the objectives described the following subsection.

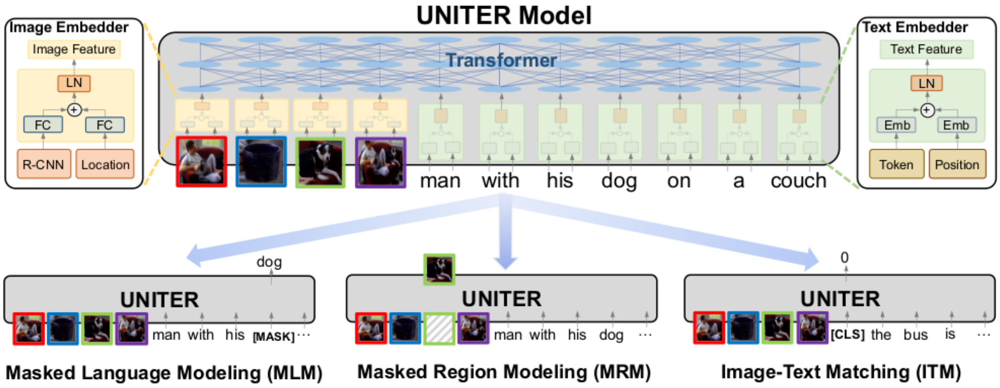


Figure 4.6: UNITER architecture

4.4.1 Training

UNITER involves four main pre-training objectives namely

- **Masked Language Modeling (MLM)** Though the names are similar, but MLM is quite different from the earlier versions of it. Here, only regions of text are masked and predicted using unmasked tasks and the complete image features.

- **Masked Region Modeling (MRM)** Similar to MLM, regions of image are masked and prediction is done with the context from remaining image regions and the complete task.

- **Image-Text Matching (ITM)** ITM is the same as the one described in previous models, where the classfier tries to check if the image corresponds to the text given.

- **Word-Region Alignment (WRA)** An optimal transport based mechanism has been devised to help learn alignment between word(w) and image region(v). Unfortunately, the exact minimization over transport map is computational intractable, and we consider the IPOT [14] algorithm to approximate the OT distance.

4.4.2 Discussion

Both MLM and MRM together try to solve the shortcomings of ViLBERT discussed in earlier section. As only one modality is masked, while keeping the other intact, we would not arise at a situation, where image feature of lets say cat is masked and also cat in the text modality was also masked. So they have a potential solution to the problem of misalignment between modalities. We shall discuss the results of the three models discussed so for in the next section.

4.5 Analysis of the models

The three models ViLBERT, LXMERT and UNITER have their own unique approaches to the problem of learning multi-modal representations. As discussed most of them have some things quite similar between them ie, the use of Encoder architecture and cross between the two modalities. ViLBERT and LXMERT were highly similar in the approach using two encoder for each modality. Although UNITER had a single encoder for both the modalities (see the table 4.1), it can be seen to outperform both the approaches. This is believed to be due to the issue of misalignment between modalities being mitigated by UNITER as discussed before.

Tasks		SOTA	ViLBERT	LXMERT	UNITER
VQA	test-dev	70.63	70.55	72.42	73.24
	test-std	70.90	70.92	72.54	73.40
VCR	Q →A	72.6	73.3	-	77.3
	QA →R	75.7	74.6	-	80.8
	Q →AR	55	54.8	-	62.8
ZS IR(Flickr)	R @ 1	-	31.86	-	65.82
	R @ 5	-	61.12	-	88.88
	R @ 10	-	72.8	-	93.52
IR(Flickr)	R @ 1	48.6	58.2	-	73.66
	R @ 5	77.7	84.9	-	93.06
	R @ 10	85.2	91.52	-	95.98
TR(Flickr)	R @ 1	67.9	-	-	88.2
	R @ 5	90.3	-	-	98.4
	R @ 10	95.8	-	-	99

Table 4.1: Comparing results on few vision language tasks. ZS: Zero-Shot, IR: Image Retrieval and TR: Text Retrieval.

4.6 VideoBERT

VideoBERT [15] tries to learn representations for video and text modality unlike the previous models that learn representations for image and text modalities. It tries to learn bidirectional joint distributions over sequence of visual and linguistic token obtained by vector quantization of the input video data and speech recognition methods. We shall learn more about them in upcoming subsections.

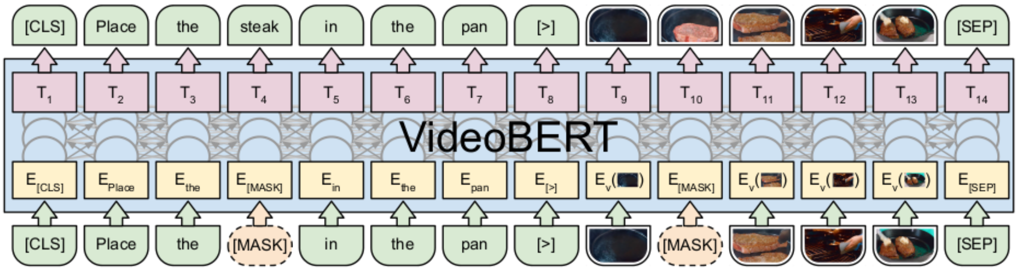
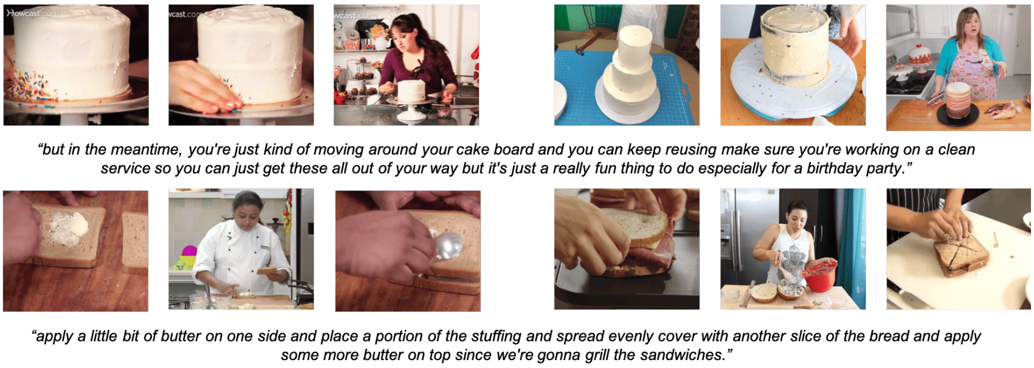

[CLS]
Place
the
steak
in
the
pan
[>]
[SEP]
T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 T11 T12 T13 T14
VideoBERT
E[CLS] E_Place E_the E[MASK] E_in E_the E_pan E[>] E_v( ) E[MASK] E_v( ) E_v( ) E_v( ) E[SEP]
[CLS]
Place
the
[MASK]
in
the
pan
[>]
[MASK]
[SEP]

Figure 4.7: VideoBERT training objectives

4.6.1 Input preprocessing

The text from the videos is obtained by using ASR toolkit provided by YouTube and then further broken into sentences by adding punctuations using LSTM based language model. For visual features, video frames are sampled at 20fps to form clips. S3D network is used to extract the features. These visual features are tokenised using hierarchical K-means, where the hyperparameters are set using visual inspection (See figure 4.8).



"but in the meantime, you're just kind of moving around your cake board and you can keep reusing make sure you're working on a clean service so you can just get these all out of your way but it's just a really fun thing to do especially for a birthday party."

"apply a little bit of butter on one side and place a portion of the stuffing and spread evenly cover with another slice of the bread and apply some more butter on top since we're gonna grill the sandwiches."

Figure 4.8: Quantizing each video segment into a token, and then represent it by the corresponding visual centroid. For each row, we show the original frames (left) and visual centroids (right). We can see that the tokenization process preserves semantic information rather than low-level visual appearance

4.6.2 Training

VideoBERT involves three training tasks: text only, video only and video-text. The text only and video only are the standard mask completion objectives. The video-text objective is the normal linguistic-visual alignment classification. Using a loss which is the weighted sum of the above the three losses, the model is trained.

4.6.3 Results

VideoBERT as a feature extractor has been quite efficient in video captioning setup. The standard practice in machine translation has been followed and BLEU, METEOR, ROUGE-L and CIDEr scores are computed. These metrics to compare generated caption to ground truth caption. Added info about metrics. Read once

- BLEU [16] is useful to compare generated caption with one or more reference captions. The key idea perform n-gram comparison between generated and reference caption and the number of matches is counted. These matches are position-independent.

- CIDEr [17] Average cosine similarity between the generated caption and the reference caption for n-grams of length n. All the n-grams or words in sentence are first converted to their root form and compared how often they occur.

- ROUGE-L [18] Matches the generated caption with set of reference captions by calculating Longest Common Subsequence(LCS). So the key idea is matches need not be consecutive.

- METEOR [19] evaluates the similarity between two sentences by creating word alignment between them. If generated caption is to be compared to multiple reference caption, the generated caption is independently compared with every reference caption and the pair with best score is considered.

Method	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr**
S3D	6.12	3.24	9.52	26.09	0.31
SOTA	7.53	3.84	11.55	27.44	0.38
VideoBERT	**7.59**	**4.33**	**11.94**	**28.80**	**0.55**

Table 4.2: Video captioning performance of YouCook II dataset.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this work on "Self-supervised approaches for learning multi modal representations", we have discussed models that use self-supervised objectives for both uni and multi modal settings. We have discussed about the advantages of using self-supervised approaches and their impact in low resource setting.

We have explored BERT(uni modal) and how its idea can be extended to multi modal case. We have critiqued the various approaches used for image captioning setup. We have also explored PASE(uni modal) that has given state of the art results of audio based tasks.

5.2 Future Work

We can extend the idea of using self-supervised objectives in cases where we the audio and the transcript. One can explore the idea of cross lingual setting, where audio and the transcript belong to different languages.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 5998–6008, Curran Associates, Inc., 2017. iv, 5, 6

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019. iv, 5, 6, 7

[3] M. Ravanelli and Y. Bengio, “Speaker recognition from raw waveform with sincnet,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, 2018. iv, 8, 9

[4] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, “Learning problem-agnostic speech representations from multiple self-supervised tasks,” in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019* (G. Kubin and Z. Kacic, eds.), pp. 161–165, ISCA, 2019. iv, v, 8, 10

[5] Wikipedia, “Modality (human–computer interaction) — Wikipedia, the free encyclopedia.” http://en.wikipedia.org/w/index.php?title=Modality%20(human%E2%80%93computer%20interaction)&oldid=949423718, 2020. [Online; accessed 04-June-2020]. 3

[6] A. Zisserman, “Self-supervised learning.” https://project.inria.fr/paiss/files/2018/07/zisserman-self-supervised.pdf. 4

[7] A. KHAZRI, “Faster rcnn object detection.” https://ai.stackexchange.com/questions/10623/what-is-self-supervised-learning-in-machine-learning. 12

[8] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2017. 13

[9] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, “Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification,” in *Computer Vision - ECCV*

2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), vol. 11219 of *Lecture Notes in Computer Science*, pp. 318–335, Springer, 2018. 13

[10] B. Choudhary, "Optimal solution of transportation problem based on revised distribution method." http://www.ijirset.com/upload/2016/august/109_Optimal.pdf. 13

[11] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems 32*, pp. 13–23, Curran Associates, Inc., 2019. 13

[12] H. Tan and M. Bansal, "LXMERT: learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), pp. 5099–5110, Association for Computational Linguistics, 2019. 15

[13] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," 2019. 17

[14] Y. Xie, X. Wang, R. Wang, and H. Zha, "A fast proximal point method for computing exact wasserstein distance," in *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019* (A. Globerson and R. Silva, eds.), p. 158, AUAI Press, 2019. 18

[15] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7463–7472, 2019. 19

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, (Philadelphia, Pennsylvania, USA), pp. 311–318, Association for Computational Linguistics, July 2002. 20

[17] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation.," in *CVPR*, pp. 4566–4575, IEEE Computer Society, 2015. 20

[18] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, (Barcelona, Spain), pp. 74–81, Association for Computational Linguistics, July 2004. 21

[19] S. Banerjee and A. Lavie, "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, (Ann Arbor, Michigan), pp. 65–72, Association for Computational Linguistics, June 2005. 21