Caption Alignment for Low Resource Audio-Visual Data

Vighnesh Reddy Konda[1], *Mayur Warialani*[1], *Rakesh Prasanth Achari*[1], *Varad Bhatnagar*[1], *Jayaprakash Akula*[1], *Preethi Jyothi*[1], *Ganesh Ramakrishnan*[1], *Gholamreza Haffari*[2] *and Pankaj Singh*[1]

[1]Indian Institute of Technology Bombay, India
[2]Monash University, Australia

{vighnesh,mayurwarialani}@cse.iitb.ac.in

Abstract

Understanding videos via captioning has gained a lot of traction recently. While captions are provided alongside videos, the information about where a caption aligns within a video is missing, which could be particularly useful for indexing and retrieval. Existing work on learning to infer alignments has mostly exploited visual features and ignored the audio signal. Video understanding applications often underestimate the importance of the audio modality. We focus on how to make effective use of the audio modality for temporal localization of captions within videos. We release a new audio-visual dataset that has captions time-aligned by (i) carefully listening to the audio and watching the video, and (ii) watching only the video. Our dataset is audio-rich and contains captions in two languages, English and Marathi (a low-resource language). We further propose an attention-driven multimodal model, for effective utilization of both audio and video for temporal localization. We then investigate (i) the effects of audio in both data preparation and model design, and (ii) effective pretraining strategies (Audioset, ASR-bottleneck features, PASE, *etc.*) handling low-resource setting to help extract rich audio representations.

Index Terms: multimodal models, low-resource audio-visual corpus, caption alignment for videos

1. Introduction

Rooted in video understanding, temporally localizing captions within videos is a relatively new and challenging task where sentences are provided alongside videos, and the task involves predicting start and end times where the sentence best aligns with the video [1, 2, 3, 4]. Grounding the caption within video could be particularly useful for indexing and retrieval applications that extract specific segments within a video corresponding to a sentence. An established approach to tackle the alignment problem is to extract frame-level video features (e.g. from 3-D convolutional neural network encoders), and compare their similarity with sentence level features (e.g. from recurrent neural network encoders). This is based on the idea that in some latent space the most similar video features will be closest to the sentence features. However, these techniques do not exploit the multimodal nature of videos and ignore the audio modality altogether.

In this paper, we aim to improve performance of temporal localization in videos by incorporating audio in an effective way. Even for existing datasets, the audio modality may benefit sentence alignment annotations, e.g. for ActivityNet [5] where the ground truth sentence alignments were created by largely ignoring the audio modality. What if the ground truth sentence alignments were instead created in an audio-sensitive and not an audio-agnostic manner, which is crucial when the audio is largely speech in a specific language? What is the effect of the language of the audio speech and that of the sentence captions on the quality of the output alignment? What are the learnings from existing datasets that can be leveraged for a new language?

We investigate these questions through a new dataset MALTA$_{av}$ (illustrated in Figure 1) which we make available through this work, and MALTA which is our proposed architecture.[1] MALTA$_{av}$ is a dataset consisting of 492 videos, with an average length of 80 seconds and around 7 sentences describing every video in each of two languages, *viz.*, Marathi and English, along with background speech in Marathi rich in content.

The main contributions of this work hinge on the following key points:

(i) The ground truth of MALTA$_{av}$ was generated by instructing close to 10 annotators to pay close attention to the audio as well as the visual streams while aligning the sentence captions with the video. We observe that this process is a lot more intensive than the video-driven and largely audio-agnostic alignment process that has been employed to create erstwhile datasets. We empirically quantify this slowdown to be by a factor of 3 by also having another subset of annotators align captions with a subset of our videos by ignoring audio (as is typically done in benchmark datasets). We refer to this subset as MALTA$_{v}$ and we use it only for evaluation purposes as test data. While we observe (as expected) that the use of the audio stream indeed improves the accuracy of sentence alignment on MALTA$_{av}$ (just as in the case of other English language datasets), we also empiricially show by contrasting performances on MALTA$_{v}$ against those on MALTA$_{av}$ that the gains might get incorrectly dampened if the ground truth in the evaluation data is created in an audio-agnostic manner. We empirically demonstrate that MALTA is effective even when the language of the speech in the videos is different from the language in which the sentences are expressed.

(ii) Another surprising observation on ertswhile datasets is that, when the video and audio modalities are combined using the architecture of MALTA, there is only a slight drop when the audio and corresponding video are deliberately designed to be incongruent. On the other hand, MALTA$_{av}$ clearly demonstrates and underlines the importance of unambiguous annotation semantics – there is a greater drop in accuracy of MALTA when we introduce incongruency or white noise in MALTA$_{av}$.

(iii) Due to the slowdown in the annotation process involving both audio and video modalities, it is somewhat challenging to scale up the number of annotated videos in MALTA$_{av}$. Consequently, to compensate for the relatively smaller (though richer) dataset, our approach in MALTA makes effective use of pre-

[1]MALTA stands for multi-**M**odal **A**nd multi-**L**ingual **T**emporal sentence **A**lignment.

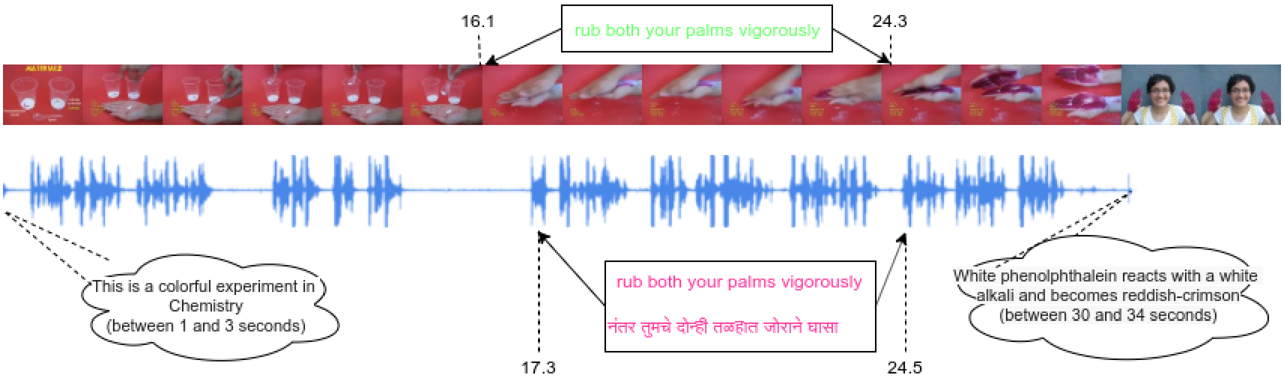

Figure 1: *Illustrating temporal sentence localization based on audio-visual features for a video (https://tinyurl.com/sm7uvfh) from our dataset, showing annotations specific to* MALTA_{av} *(in pink) and* MALTA_{v} *(in green). Please note for the sake of the readers, that the two call outs at the beginning and end are English translations of the original Marathi speech.*

training of the audio (speech) features using existing speech dataset in the new language (Marathi) while leveraging pre-trained visual features using the existing large caption alignment datasets (such as ActivityNet) in English.
(iv) Summarily, MALTA is based on (a) language specific pre-training of the audio modality, (b) mutual co-attention between the three audio, video and text modalities for their effective combination, (c) analyzing the role of audio in performance gain by manipulating the videos to have incongruent audio and using caption-audio from different languages.

Our paper is organised as follows. In Section 3.1, we motivate the design of our MALTA architecture by describing the obvious and subtle aspects of our MALTA_{av} (and MALTA_{v}) dataset. In Section 4 we present detailed experimental analysis. We present related work in Section 2 and conclusions in Section 6

2. Related Work

To match the query and video frame candidates, one approach is to map the visual features of the frame candidates and the textual feature of the caption into a shared space and measure their semantic similarity. This is the basis of Moment Context Network (MCN) [3] and Cross-modal Temporal Regression Localize (CTRL) [6] where the two works differ in the way they build (i) a context-dependent encoding of the video frames, i.e. neighboring local frames in CTRL vs all global frames in MCN, and (ii) an encoding of the sentence caption, i.e. last LSTM state MCN vs skip-thoughts in CTRL. [4] proposed the use of attention mechanism to flexibly adapt to relevant cues in the caption. Most relevant to our work is Attention Based Location Regression (ABLR) [7] which uses a multimodal co-attention mechanism to identify the relevant video frames based on an encoding of the caption.[2] proposes a reinforcement learning based framework by learning an agent which regulates the temporal grounding boundaries progressively based on its policy. [8] proposed Moment Alignment Network (MAN), which unifies the candidate frame encoding and temporal structural reasoning into a single-shot feed-forward network.

Attention-based methods have also been widely employed for the related task of video captioning [9]. [10] generate semantically rich text using an attention-based LSTM network. [11] proposed an attention model that adaptively selects interesting regions from each frame pertinent to the task. [12] proposed the application of attention via fusion across multiple modalities. Among fusion techniques, [13] proposed a multimodal circulant fusion technique and [14] used a novel child-sum fusion technique for improved effectiveness.

Several techniques have been employed in prior work to leverage information from both audio and visual modalities for the task of caption generation. Ramanishka *et. al.* [15] and Jin *et. al.* [16] leveraged multimodality within an encoder-decoder model and obtained a boost in performance. Jin *et. al.* [17] and Hori *et. al.* [18] also used speech features from the audio modality to gain further improvements. On other tasks such as video event classification [19], [20] and [21] have shown improvements by using audio features along with visual features. As for the use of multiple modalities for caption alignment, there is no specific prior work that has come to our attention.

3. MALTA: Dataset and Architecture

Through our dataset MALTA_{av}, we extend the task of temporal localization of sentences within videos for settings where the alignment needs to be performed based on both video and audio modalities. In Figure 1, we illustrate the obvious as well as subtle aspects of such a task on a video segment (from our dataset), whose (rich) audio modality also consists of speech. The speech, in turn might be in a language different from that of the sentence. In our problem setting (*wrt* the Figure 1), we have two synonymous versions of the sentences to be aligned ('rub...vigorously' printed in pink color), *viz.*, English and Marathi, of which we will be provided captions in exactly one of the two languages. The ground truth alignment for the sentence is between 17.3 and 24.5 seconds (pink) when the audio modality is carefully considered. However, going simply by the image frames in the video, the alignment is inferred (by another annotator) to be between 16.1 and 24.3 seconds (green). This discrepancy is simply because the rubbing of the palms is reflected in the speech and sound only starting at 17.3 seconds, though the palms themselves are visible starting at 16.1 seconds. However, such careful annotation, driven by both audio and visual modalities comes at the price of slowdown in the annotation process at least by a factor of 3. In addition to such data being harder and slower to create, the speech (or the text) could be in a language (such as Marathi) with a limited number of available videos. The sentences and speech could either be in the same language or different languages. Our model

MALTA captures both these settings: specifically in the case of the data MALTA_{av} released with this paper, we consider the case in which videos contain audio in the Marathi language and sentences are available in both English and Marathi. Pre-training can help (partially) address both the limited training data and language-mismatch constraints. Examples of benefitting from audio pre-training include use of bottleneck layer or the phoneme probabilities after pre-training an ASR model on the language of the audio-speech or even simpler MFCC features. Pre-trained cross-lingual models such as XLM [22], FastText [23] could also be leveraged on the textual side.

We also make some subtler observations in Figure 1. The speech (call out, translated from Marathi) '*This ... Chemistry*' is an abstract statement between 1 and 3 seconds even before any experiment or any variety of colors are observed! Cross-modal attention could help address this; the image features from the video between 6 and 30 seconds capturing the different colors - white and reddish-crimson as well as the interactions between the powders to yield a different colored output can help attend to the features from the speech segment between 1 and 3 seconds. On the other hand, consider the speech (call out, translated from Marathi) between 30 and 34 seconds: '*White .. reddish-crimson*'. In the video frames corresponding to that speech, neither *Phenolphthalein* nor the *alkali* nor the process of transformation can be seen. This motivates the need for attention from audio to video; the speech features in this segment containing references to *white phenolphthalein* and *white alkali* can help attend to the features from the video frames between 5 seconds and 18 seconds in which the white powders are actually shown.

MALTA is an attention-driven multimodal architecture that supports mutual co-attention between the audio, video and text modalities. The flexible attention schemes between modalities in MALTA allows for a wide spectrum of tasks (e.g., rich in video, rich in audio, audio is in a language different from the captions, etc.) to be easily handled. Section 3.1 describes the specific attention schemes that we use, along with the loss functions that are used to explicitly supervise the modality-specific attentions by treating ground truth alignments as *hard* attentions.

3.1. Attention Based Location Regression

Our multimodal architecture MALTA scaffolds on Attention Based Location Regression (ABLR) [7] It is an end-to-end architecture to convert video and sentence inputs to the temporal coordinates in the output. MALTA comprises three main components as depicted in Figure 2: (i) context-dependent feature encoding of the input audio, video streams and sentence, (ii) multi-modal co-attention interaction highlighting important audio, visual segments in the video and words in the sentence, and (iii) attention based output prediction which can directly regress the temporal coordinates of the target video.

Assume access to a set of N training instances, $\{A_i, V_i, S_i, \tau_i^s, \tau_i^e\}_{i=1}^N$ consisting of audio features A_i, video features V_i and an accompanying sentence describing a segment in the video S_i with start and end times, τ_i^s and τ_i^e respectively. At test time, for a given video and sentence, our task involves predicting the start and end times demarcating the temporal alignment of the sentence within the video.

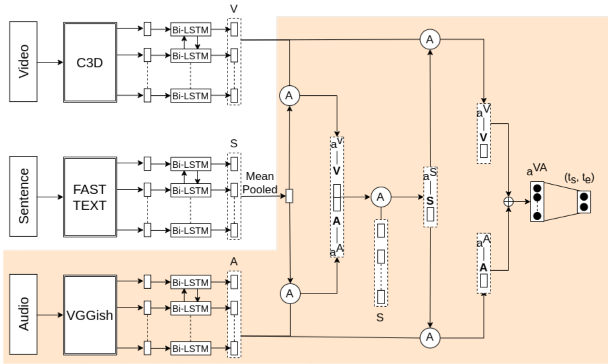

Figure 2: *our proposed* MALTA *(an extension of ABLR)*

3.1.1. Input Feature Representation

Video Feature. A video V is first clipped into M segments, $\{v_1, \ldots, v_i, \ldots, v_M\}$ that are featurized into dense video representations using the well-known C3D network [24]. These feature vectors are subsequently passed as input to a bidirectional LSTM-based encoder so that each video segment's representation is further enhanced with contextual information from the sequence of video segments. Finally, a linear transform is applied to the hidden states from the bi-LSTM network.

$$
\begin{aligned}
\boldsymbol{x}_i &= \text{C3D}(v_i) && (1)\\
\boldsymbol{h}_i &= \text{biLSTM}(\boldsymbol{x}_i, [\overrightarrow{\boldsymbol{h}}_{i-1}; \overleftarrow{\boldsymbol{h}}_{i+1}]^T) && (2)\\
\boldsymbol{v}_i &= \boldsymbol{f}_{\text{enc}}^{(v)}(\mathbf{W}_{\text{enc}}^{(v)}\boldsymbol{h}_i + \boldsymbol{b}_{\text{enc}}^{(v)}) && (3)
\end{aligned}
$$

where $\boldsymbol{f}_{\text{enc}}^{(v)}$ is a nonlinear activation function, and $\mathbf{W}_{\text{enc}}^{(v)}$ and $\boldsymbol{b}_{\text{enc}}^{(v)}$ are the parameter matrix and vector for the video modality, respectively, that projects each $\boldsymbol{h}_i$ to an h_v-sized vector $\boldsymbol{v}_i$. A video $\{v_1, \ldots, v_i, \ldots, v_M\}$ is thus encoded as $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_M] \in \mathbb{R}^{h_v \times M}$.

Audio Feature. For encoding the audio modality, we use VGGish features [25] and linearly map the audio features to video segments. As such, the audio U is also divided into M clips $\{u_1, .., u_j, .., u_M\}$ in sequential order. Finally, these VGGish features are passed to a bidirectional LSTM network to generate context-aware audio feature representations, $[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M] \in \mathbb{R}^{h_a \times M}$. These features are then integrated within ABLR via multimodal co-attention.

Text Feature. A sentence S containing N words, $\{s_1, \ldots, s_N\}$, is encoded using a sequence of steps analogous to the process mentioned above for video encoding. Each token is first encoded as a 300-dimensional (multi-lingual) FastText vector (better than Glove as per supplementary). A sequence of word embeddings is fed as an input to a bidirectional LSTM which contextualizes the sentence features. A final projection step produces a sentence representation, $[\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N] \in \mathbb{R}^{h_s \times N}$. We set $h_s = h_v$ in our experiments.

3.1.2. Co-Attention between Audio, Video and Text Modalities.

we consider sentence-video and sentence-audio interactions independently and compute attention distributions over the video/audio modalities using co-attention. We use the sentence to learn attention on both video and audio modalities separately and then concatenate both attended features to further attend to the sentence. We use the attended sentence features to attend once again to the audio and video modalities separately. We finally sum the attention distributions over both video and audio modalities, normalize it and use the resulting distribution to regress the temporal coordinates of the sentence within the video. (See Figure 2.)

We denote the final attentions representations by $\boldsymbol{a}_U$ for the audio, and $\boldsymbol{a}_V$ afor the video respectively. $a_{V,j}$, $a_{U,j}$ denotes the relative importance of video, audio respectively in the j-th clip for the given sentence.

3.1.3. Training Objective:

Let the ground-truth start and end times of sentence S_i in i^{th} video of duration d_i be $\bar{\tau}_i^s$ and $\bar{\tau}_i^e$, respectively.[2] The predicted start and end times, $\hat{\tau}_i^s$ and $\hat{\tau}_i^e$, are obtained using the sum of final audio and video attention weights and directly regressing the temporal coordinates:

$$(\hat{\tau}_i^s, \hat{\tau}_i^e) = \boldsymbol{f}_{\text{pred}}^{(v)}(\mathbf{W}_{\text{pred}}^{(av)} \boldsymbol{a}_V A^{\intercal} + \boldsymbol{b}_{\text{pred}}^{(av)}) \tag{4}$$

we use a linear interpolation of two losses, L_{reg} and L_{cal}, to supervise the prediction of temporal coordinates of a sentence within a video.

$$L_{\text{reg}} = \sum_{i=1}^{N} [R(\bar{\tau}_i^s - \hat{\tau}_i^s) + R(\bar{\tau}_i^e - \hat{\tau}_i^e)] \tag{5}$$

$$L_{\text{cal}} = -\sum_{i=1}^{N} \frac{\sum_{j=1}^{M} \delta_{i,j} \log{(a_{V_i,j} * a_{A_i,j})}}{\sum_{j=1}^{M} \delta_{i,j}} \tag{6}$$

where $R(\cdot)$ is a smooth L1 function [7], $\delta_{i,j} = 1$ if the j^{th} segment in V_i is within the interval (τ_i^s, τ_i^e) and 0 otherwise. Here, $a_{V,j}$ denotes the relative importance of video in the j-th clip for the given sentence.

3.2. Attending to Audio, Video and Textual Modalities

ABLR uses only video and sentence-based features while completely neglecting audio-based features. We extend the ABLR model to leverage the audio modality. Our claim on the design appropriateness of MALTA is reinforced in two skyline experiments (c.f. §4) wherein we use (i) ground truth based hard attention (instead of attentions inferred from MALTA) and (ii) speech-to-text (ASR) output of the speech channel as the input instead of the audio channel ; the sentences are largely expected to have significant n-gram overlap with the speech transcripts.

3.2.0.1. Audio Encoding:

For encoding the audio modality, we use VGGish features [25] and linearly map the audio features to video segments. As such, the audio U is also divided into M clips $\{u_1, .., u_j, .., u_M\}$ in sequential order. Finally, these VGGish features are passed to a bidirectional LSTM network to generate context-aware audio feature representations, $[\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M] \in \mathbb{R}^{h_a \times M}$. These features are then integrated within ABLR via multimodal co-attention.

3.2.0.2. Co-Attention between Video, Audio and Text:

After creating context-dependent representations of audio, video and text, we devise two multimodal co-attention schemes in MALTA to guide the model towards focusing on relevant parts of different modalities in order to better align the sentence with the video.

(1) In our first scheme, we consider sentence-video and sentence-audio interactions independently and compute attention distributions over the video/audio modalities using co-attention. We use the sentence to learn attention on both video and audio modalities separately and then concatenate both attended features to further attend to the sentence. We use the attended sentence features to attend once again to the audio and video modalities separately. We finally sum the attention distributions over both video and audio modalities, normalize it and use the resulting distribution to regress the temporal coordinates of the sentence within the video. (See Figure 2.) We will refer to a model trained with this scheme as CONC-AV.

(2) In our second scheme, we enable more explicit interactions between the video and audio modalities. Instead of developing coattention mechanisms independently between sentence-video and sentence-audio modalities, we use attended video features (driven by the initial sentence) to learn an attention distribution over the audio features. The attended audio features are subsequently used to learn an attention distribution over the sentence. These *multimodal* attended sentence features are finally used to attend to both the audio and video modalities. (See Figure **??**.) We will refer to a model trained with this scheme as JOINT-VA.

We denote the final attentions representations by $\boldsymbol{a}_U$ for the audio, and $\boldsymbol{a}_V$ afor the video respectively. Similar to $a_{V,j}$ (for video), $a_{U,j}$ denotes the relative importance of audio in the j-th clip for the given sentence.

Summarily, CONC-AV gives equal importance to both the video and audio modalities initially and are independently driven by the sentence. In a second attention step, the audio and video attention vectors are concatenated to further attend to the sentence. In JOINT-VA, the video modality is given higher importance and attended video features (driven by the sentence) are used to learn an attention distribution over the audio features.

4. Experimental Results

We attempt to answer the following questions through our experiments. (i) Do we consistently benefit from attending to multiple modalities? (ii) What is the effect of use of different modalities when the ground truth sentence alignments are created in an audio-video-driven (as against only video-driven) manner? (iii) What is the effect of the language of the speech in the audio and the language of the sentence captions on the quality of the alignment output (iv) How does performance vary using MALTA when we deliberately manipulate videos to have incongruent audio? We investigate (i), (ii), (iii) and (iv) in this section.

4.0.0.1. Datasets.

We conduct experiments on our newly constructed multimodal data MALTA_{av} as well as two standard benchmarks, namely Charades-STA [6] and ActivityNet [5]. We will release our dataset upon acceptance of this paper. MALTA_{av} consists of simple video tutorials of two types: (i) that describes the creation of scientific toys from waste material[3](ii) ATMA_{av} that features farmers describing and demonstrating organic farming techniques. Both video collections have speakers in the background narrating the process in Marathi. These videos are rich

[2] The ground-truth start and end times are normalized by the duration of the video. That is, $(\bar{\tau}_i^s, \bar{\tau}_i^e) = (\frac{\tau_i^s}{d_i}, \frac{\tau_i^e}{d_i})$

[3] We downloaded these videos from http://www.arvindguptatoys.com/toys-from-trash.php and obtained consent from the content creator

in both video and audio content. consists of 492 videos, with an average length of 80 seconds and around 7 sentences describing every video in each of two languages, *viz.*, Marathi and English, along with background speech in Marathi. On the other hand, $\mathrm{ATMA_{av}}$ is relatively smaller, consisting of 95 videos, with an average length of 111 seconds and around 18 sentences describing every video in a single language, *viz.*, Marathi, accompanied by background speech in Marathi.

Charades-STA contains 16128 clip-sentence pairs; we created training/test splits containing 12408/3720 pairs, respectively. ActivityNet is significantly larger containing 20K videos and 100K sentences annotated with start and end times. We used the publicly-available train set for training and the validation set to evaluate our models.

4.0.0.2. Evaluation metric.

Following the metrics adopted in prior work for temporal localization of sentences in videos [6], for each sentence, we calculate the Intersection over Union (IoU) between the predicted and ground truth temporal coordinates. "IoU $= \alpha$" denotes the percentage of the sentence queries which have an IoU larger than α.

4.0.0.3. Implementation Details.

Videos in ActivityNet, Charades-STA and $\mathrm{MALTA_{av}}$ were split into 8922:4369 , 5338:1334 and 389:103 clips for training and testing, respectively. We extracted 4096-dimensional C3D features for each dataset to serve as the video features and 128-dimensional audio features were extracted using VGG. Bidirectional LSTM layers with a hidden state size of 256 were used for each modality. We used the Adam optimizer to train MALTA with a learning rate of 0.001.

4.0.0.4. Pretrained ASR-specific features.

We used the Kaldi toolkit [26] to train a state-of-the-art time delay neural network (TDNN) acoustic model on roughly 100 hours of weakly labelled Marathi spoken tutorials.[4] (These utterances are weakly labelled as we use subtitles in the videos to extract transcriptions for speech.) The TDNN model has 12 layers with a 128-dimensional bottleneck layer before the penultimate layer. We decoded Marathi speech from the videos in $\mathrm{MALTA_{av}}$ using this trained network and extracted bottleneck features. These features will henceforth be referred to as ASR-bnf.

4.0.0.5. PASE features

PASE [27] is a pretrained speech model consisting of multiple workers that are jointly trained to optimize seven different speech-driven self-supervised tasks, including regression tasks that involve predicting the waveform, MFCC [28] and prosody features and binary discrimination tasks that differentiate between positive and negative samples based on an anchor utterance. We do not make use of the speech labels while extracting PASE features.

4.1. Single Modality

In order to systematically analyze the importance of combining modalities, we first investigate systems that only consider co-attention between a single modality (video or audio) and the sentence. Table 1 reports our results on all three datasets;

[4] Available from: https://spoken-tutorial.org/

MODEL	IoU= .1	IoU= .3	IoU= .5	IoU= .7
Activity-Net				
A-ONLY	0.6941	0.5203	0.3373	0.1705
V-ONLY	0.7236	0.5454	0.3571	0.1786
Charades-STA				
A-ONLY	0.6435	0.5002	0.3583	0.1537
V-ONLY	0.5947	0.4910	0.3611	0.1462
$\mathrm{MALTA_{av}}$				
A-ONLY	0.34	0.24	0.14	0.05
ASR-BNF-ONLY	0.3314	0.2459	0.1528	0.0557
V-ONLY	0.32	0.23	0.13	0.04

Table 1: *Results on ActivityNet, Charades-STA,* MALTA_{av} *using a single modality*

A-ONLY refers to using the audio VGG features alone, and V-ONLY refers to using just the C3D video features. We observe that V-ONLY outperforms A-ONLY on ActivityNet. On Charades-STA and $\mathrm{MALTA_{av}}$, A-ONLY on $\mathrm{MALTA_{av}}$ is better than V-ONLY (with the margin being larger for $\mathrm{MALTA_{av}}$). We also observe that the ASR specific features perform even better than the VGG features on $\mathrm{MALTA_{av}}$, thus confirming our claim that $\mathrm{MALTA_{av}}$ benefits from good features encoding the underlying Marathi speech.

4.2. Combining Modalities

Here we investigate question (i) mentioned at the start of Section 4: Does attending to multiple modalities always help? We use two multimodal co-attention schemes, CONC-AV (shown in Figure 2 and JOINT-VA (shown in Figure **??**), to leverage information from both audio and video modalities. Table 2 shows the performance of both multimodal schemes on ActivityNet and Charades-STA. We report consistent improvements in performance with using JOINT-VA over ABLR [7], which is a near state-of-the-art system on both ActivityNet and Charades-STA. JOINT-VA is marginally better than CONC-AV thus highlighting the potential utility of explicitly modeling interactions between the video and audio modalities. The asymmetry in JOINT-VA in using the video modality to attend to audio (rather than going

MODEL	IoU= .1	IoU= .3	IoU= .5	IoU= .7
Activity-Net				
ACRN [29]	0.5037	0.3129	0.1617	-
Xu et al [30]	-	0.4530	0.2770	0.1360
ABLR [7]	0.7330	0.5567	0.3679	-
CONC-AV	0.7137	0.5388	0.3636	0.1872
JOINT-VA	0.7410	0.5523	0.3739	0.1924
Charades-STA				
CTRL [31]	-	-	0.2363	0.0889
Xu et al [30]	-	0.547	0.3560	0.1580
ABLR [7]	0.5947	0.4910	0.3611	0.1462
CONC-AV	0.6510	0.5220	0.3650	0.1490
JOINT-VA	0.6462	0.5238	0.3664	0.1447

Table 2: *Results on ActivityNet, Charades-STA using multi-modal co-attention. "-" corresponds to missing entries for these cells in the respective papers.*

A-feat	IoU= .5	IoU= .7
-	0.1321 ± 0.004	0.0485 ± 0.002
VGG	0.1420 ± 0.002	0.0485 ± 0.005
MFCC	0.1425 ± 0.006	0.0439 ± 0.006
PASE-scratch	0.1387 ± 0.006	0.0474 ± 0.003
PASE-spk scratch	0.1375 ± 0.006	0.0496 ± 0.002
PASE-finetuned	0.1450 ± 0.005	0.0459 ± 0.003
PASE-spk finetuned	0.1459 ± 0.005	0.0484 ± 0.005
PASE-+spk finetuned	0.1478 ± 0.006	0.0462 ± 0.005
PASE ASR-fine tune	0.1450 ± 0.007	0.0451 ± 0.006
ASR-bnf	0.1550 ± 0.005	0.0545 ± 0.005

Table 3: *Results on with multimodal coattention comparing different audio representations. The first row corresponds to the* V-ONLY *model.*

A-feat	IoU= .5	IoU= .7
VGG	0.0476 ± 0.012	0.0065 ± 0.002
MFCC	0.0388 ± 0.003	0.0112 ± 0.004
PASE-ATMA$_{av}$ scratch	0.0382 ± 0.006	0.0147 ± 0.006
PASE-spk finetuned	0.0392 ± 0.007	0.0157 ± 0.003

Table 4: *Results on the* ATMA$_{av}$ *dataset.*

from audio to video) is justified for ActivityNet and Charades-STA where video is much richer in content compared to audio.

Table 3 shows the improvement in performance with using CONC-AV, JOINT-VA and JOINT-AV on our audio-rich MALTA$_{av}$ dataset. We report consistent improvements in performance with using both audio and video modalities. We also see a larger differential in performance between V-ONLY and CONC-AV compared to ActivityNet and Charades-STA, which points to the audio modality being much richer in content in MALTA$_{av}$. We observe the benefits of using speech-aware ASR features with CONC-AV compared to using either VGG or MFCC-based audio features. Also, JOINT-AV is consistently better than JOINT-VA for MALTA$_{av}$ which is justified since MALTA$_{av}$ is rich in speech content.[5]

4.3. Skylines for Audio Modality

Our claim on the appropriateness of the design of MALTA is reinforced in two skyline experiments wherein (i) we use ground truth based "hard" attention (instead of attentions inferred from MALTA) to regress the temporal coordinates for ActivityNet and (ii) we use transcriptions for the speech channel in MALTA$_{av}$ (derived using Google's speech recognition API for Marathi) as an input modality instead of audio features. Table 5 shows results from both these skyline experiments. We expect the ASR-based transcriptions to serve as a skyline because we expect the Marathi transcriptions from Google's API to be largely accurate, in which case the sentences are expected to have significant n-gram overlap with the speech transcriptions.

4.4. Sensitivity to Incongruent Audio

We investigate question (iv) mentioned at the start of Section 4: How does performance vary using MALTA when we deliberately manipulate videos to have incongruent audio extracted randomly from another video? The results are reported in Table

[5] In all subsequent experiments, we use CONC-AV and not JOINT-VA, as the latter is more expensive to train and only marginally better in performance compared to CONC-AV.

MODEL	IoU= .1	IoU= .3	IoU= .5	IoU= .7
ActivityNet				
Hard-Attention	0.9661	0.8934	0.7976	0.3105
MALTA$_{av}$				
Hard-Attention	0.9421	0.8552	0.6526	0.5364
Transcript	0.3342	0.2526	0.1710	0.0736
Video+Transcript	0.3514	0.2585	0.1730	0.0689

Table 5: *Skyline results on ActivityNet using hard attention and using Google transcriptions on* MALTA$_{av}$

6. When the video and audio modalities are combined using MALTA, there is only a slight drop when the audio and corresponding video are deliberately designed to be incongruent on existing datasets such as ActivityNet. On the other hand, MALTA$_{av}$ clearly demonstrates and underlines the importance of unambiguous annotation semantics – there is a greater drop in accuracy of MALTA when we introduce incongruency in MALTA$_{av}$.

Dataset	IoU= .1	IoU= .3	IoU= .5	IoU= .7
ActivityNet	0.7263	0.5408	0.3555	0.1782
MALTA$_{av}$	0.3779	0.2486	0.1336	0.0529

Table 6: *Results on ActivityNet and* MALTA$_{av}$ *with incongruent audio on* CONC-AV *model.*

4.5. Correctly Analysing Gains from Audio

We assess the overlap between the video driven annotations on MALTA$_{v}$ with the more ideal, audio-video driven annotations on MALTA$_{av}$ (without invoking any model). We find IoU=0.1 to be 0.71 and IoU=0.7 to be 0.19. It is interesting to note that the IOU is not that high, reinforcing our claim that alignment using only the video modality will not be very accurate.

In Table 7, we illustrate the somewhat inaccurate assessment of the improvement of CONC-AV over V-ONLY by a less accurately aligned dataset such as MALTA$_{v}$. MALTA$_{av}$ more faithfully represents the gains obtained by CONC-AV compared to V-ONLY (*c.f.* Table 3).

Test on MALTA$_{v}$

MODEL	IoU= .1	IoU= .3	IoU= .5	IoU= .7
V-ONLY	0.3375	0.2063	0.0906	0.0313
CONC-AV	0.3156	0.2109	0.0953	0.0375

Table 7: *Marginal improvement of* CONC-AV *over* V-ONLY *on* MALTA$_{v}$.

4.6. Cross-Lingual Evaluation

In Table 8, we empirically demonstrate on MALTA$_{av}$ that MALTA is effective even when the language of the speech in the videos (Marathi) is different from the language in which the sentences are expressed (English). Even with the mismatch in language, we see a significant improvement in using the audio modality with CONC-AV compared to using only the video modality.

MODEL	IoU= .5	IoU= .7
V-ONLY	0.1428 ± 0.006	0.0452 ± 0.003
CONC-AV with ASR-bnf	0.1490 ± 0.005	0.0566 ± 0.001

Table 8: *Results on* MALTA_{av} *when speech is in Marathi but the captions are in English.*

5. Discussion and Analysis

We observe that the differences between V-only and Conc-AV, A-only and Joint-VA on ActivityNet and Charades-STA are marginal. This is largely because of audio being less prominent in most videos. Among the few videos where audio was more dominant, we observed that our joint audio-video models performed significantly better than the V-only model.

Further, as expected, we observe that the use of the audio modality indeed improves the accuracy of sentence alignment on MALTA_{av}, just as in the case of other popular benchmark datasets. We also empirically find that MALTA is effective even when the language of the speech in the videos is different from the language in which the sentences are expressed.

6. Conclusion

We present an approach MALTA for localizing sentences/captions in videos that leverages both audio and video modalities and that can generalize to new and possibly low-resource language settings. Our approach bootstraps around pre-training of the respective modalities, use of co-attention across the audio-visual and textual modalities and taming of the respective attentions. We present a rich new dataset MALTA_{av}, whose annotation is driven by both audio and visual modalities and which is richer in the audio modality than previous datasets. Further, MALTA_{av} has sentences in two languages (including the language of the speech in the audio modality). We study a state-of-the-art model as well as MALTA on existing monolingual, video-heavy benchmarks as well as on our dataset and present how performance of architectural variations in the model corresponds to the modalities that were used to drive the sentence alignment annotation. We also experimentally validate that speech-heavy audio modality could also benefit sentence alignment when the sentence is in a language different from that of the speech.

7. Acknowledgements

We are grateful to IBM Research, India (specifically the IBM AI Horizon Networks - IIT Bombay initiative) for their support and sponsorship.

8. References

[1] J. Chen, X. Chen, L. Ma, Z. Jie, and T.-S. Chua, "Temporally grounding natural sentence in video," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 162–171. [Online]. Available: https://www.aclweb.org/anthology/D18-1015

[2] D. He, X. Zhao, J. Huang, F. Li, X. Liu, and S. Wen, "Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos," *CoRR*, vol. abs/1901.06829, 2019. [Online]. Available: http://arxiv.org/abs/1901.06829

[3] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. C. Russell, "Localizing moments in video with temporal language," *CoRR*, vol. abs/1809.01337, 2018. [Online]. Available: http://arxiv.org/abs/1809.01337

[4] M. Liu, X. Wang, L. Nie, Q. Tian, B. Chen, and T.-S. Chua, "Cross-modal moment localization in videos," 10 2018, pp. 843–851.

[5] R. Krishna, K. Hata, F. Ren, F. Li, and J. C. Niebles, "Dense-captioning events in videos," *CoRR*, vol. abs/1705.00754, 2017. [Online]. Available: http://arxiv.org/abs/1705.00754

[6] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: temporal activity localization via language query," *CoRR*, vol. abs/1705.02101, 2017. [Online]. Available: http://arxiv.org/abs/1705.02101

[7] T. M. Y. Yuan and W. Zhu, "To find where you talk: Temporal sentence localization in video with attention based location regression. aaai," 2019.

[8] D. Zhang, X. Dai, X. Wang, Y. Wang, and L. S. Davis, "MAN: moment alignment network for natural language moment retrieval via iterative graph adjustment," *CoRR*, vol. abs/1812.00087, 2018. [Online]. Available: http://arxiv.org/abs/1812.00087

[9] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proceedings of ICCV*, 2015.

[10] H. Z. X. X. L. Gao, Z. Guo and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," ser. IEEE, 2017, pp. Trans. Multimedia, vol. 19, no. 9, pp. 2045–2055.

[11] Z. Yang, Y. Han, and Z. Wang, "Catching the temporal regions-of-interest for video captioning," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 146–153.

[12] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Proceedings of ICCV*, 2017.

[13] A. Wu and Y. Han, "Multi-modal circulant fusion for video-to-language and backward," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, ser. IJCAI'18. AAAI Press, 2018, pp. 1029–1035. [Online]. Available: http://dl.acm.org/citation.cfm?id=3304415.3304561

[14] J. Xu, T. Yao, Y. Zhang, and T. Mei, "Learning multimodal attention lstm networks for video captioning," in *Proceedings of the 25th ACM International Conference on Multimedia*, ser. MM '17. New York, NY, USA: ACM, 2017, pp. 537–545. [Online]. Available: http://doi.acm.org/10.1145/3123266.3123448

[15] V. Ramanishka, A. Das, D. H. Park, S. Venugopalan, L. A. Hendricks, M. Rohrbach, and K. Saenko, "Multimodal video description," in *Proceedings of the 24th ACM International Conference on Multimedia*, ser. MM '16. ACM, 2016, pp. 1092–1096.

[16] Q. Jin, J. Liang, and X. Lin, "Generating natural video descriptions via multimodal processing." in *Proceedings of Interspeech*, 2016, pp. 570–574.

[17] Q. Jin, J. Chen, S. Chen, Y. Xiong, and A. Hauptmann, "Describing videos using multi-modal fusion," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1087–1091.

[18] C. Hori, T. Hori, T. K. Marks, and J. R. Hershey, "Early and late integration of audio features for automatic video description," in *Proceedings of ASRU*, 2017, pp. 430–436.

[19] Y.-G. Jiang, Z. Wu, J. Tang, Z. Li, X. Xue, and S.-F. Chang, "Modeling multimodal clues in a hybrid deep learning framework for video classification," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3137–3147, 2018.

[20] X. Long, C. Gan, G. de Melo, X. Liu, Y. Li, F. Li, and S. Wen, "Multimodal keyless attention fusion for video classification," in *Proceedings of AAAI*, 2018.

[21] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 569–576.

[22] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *arXiv preprint arXiv:1901.07291*, 2019.

[23] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *CoRR*, vol. abs/1607.04606, 2016. [Online]. Available: http://arxiv.org/abs/1607.04606

[24] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, 2013. [Online]. Available: https://doi.org/10.1109/TPAMI.2012.59

[25] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *Proceedings of ICASSP*, 2017, pp. 131–135.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," 2011, iEEE Catalog No.: CFP11SRW-USB. [Online]. Available: http://infoscience.epfl.ch/record/192584

[27] *Multi-task self-supervised learning for Robust Speech Recognition*, 01 2020.

[28] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," *Journal of Computer science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[29] M. Liu, X. Wang, L. Nie, X. He, B. Chen, and T.-S. Chua, "Attentive moment retrieval in videos," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: ACM, 2018, pp. 15–24. [Online]. Available: http://doi.acm.org/10.1145/3209978.3210003

[30] H. Xu, K. He, L. Sigal, S. Sclaroff, and K. Saenko, "Text-to-clip video retrieval with early fusion and recaptioning," *CoRR*, vol. abs/1804.05113, 2018. [Online]. Available: http://arxiv.org/abs/1804.05113

[31] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.