BACHELOR'S THESIS

# Multimodal Learning With Limited Supervision

Akkapaka Saikiran
*Department of Computer Science and Engineering*
*Indian Institute of Technology Bombay*
saikiraniitb@gmail.com

*Advisors*

Prof. Preethi Jyothi
*CSE, IIT Bombay*
pjyothi@cse.iitb.ac.in

Prof. Ganesh Ramakrishnan
*CSE, IIT Bombay*
ganesh@cse.iitb.ac.in

*Collaborators*

Shubham Nemani
*CSE, IIT Bombay*
nshubham655@gmail.com

Pranjal Saini
*CSE, IIT Bombay*
pranjalsaini.24@gmail.com

Battepati Karthikeya
*CSE, IIT Bombay*
190050026@iitb.ac.in

Rishabh Dabral
*CSE, IIT Bombay*
rishdabral@gmail.com

# Contents

# 1 Introduction

In recent years we have seen an enormous growth in the volume, reach, and use of multimedia content – text, photo, video, audio, etc. This growth gives rise to various interesting and challenging pertinent problems like image/video captioning, generation (from natural language prompts), retrieval, etc. The task of text to video retrieval, for example, is to find relevant videos from a database given a text query.

In this report, we discuss the problem of audio-text alignment. This problem is of great relevance in situations where a quick audio look-up is needed but there are no checkpoints in the audio. We first lay out the problem setting in some detail. Then we present some approaches towards developing a solution. En route, we examine the challenges that crop up. Finally, we describe a few other threads that were discussed in the thesis.

# 2 Audio-text Alignment

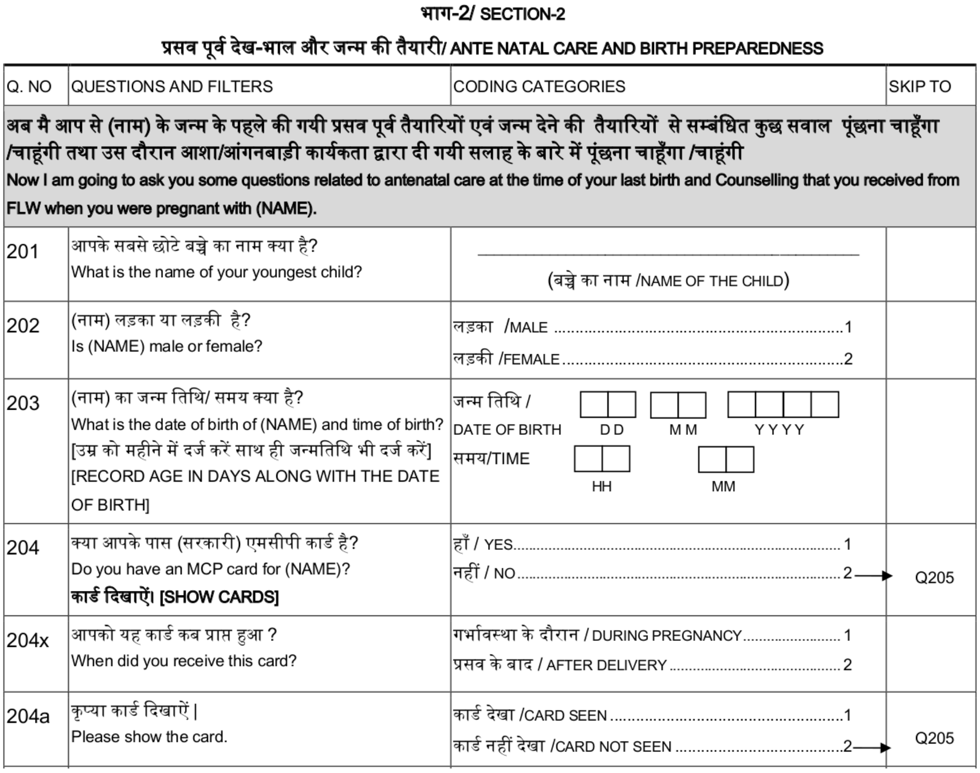
## 2.1 Introduction

Due to a collaboration with IIT Bombay's Koita Centre for Digital Health (KCDH) and CARE India, an Indian NGO, we obtained access to a collection of health surveys. These surveys were available mostly as audio interviews with mothers from rural parts of Bihar. A useful analytical feature on such data is being able to find the time stamp when a particular question was asked in the interview. We pose this as an audio-text alignment problem – the system needs to align an audio audio with a given text query, i.e. locate an audio segment that is similar or relevant to the query. We aim to train a model in a weakly supervised mfashion, i.e. by telling it which questions occur in an audio snippet but by witholding the questions' temporal location. This reduces the burden on annotation and allows us to scale the training data more easily.

There are several additional pieces of information that we utilize. For starters, the interviews are based off a few questionnaires, so we know which questions the interviews can ask. Do note, however, that the interviewers often rephrase the questions, so this is an additional complication for the training. As another source of supervision, we have an ordering on the questions, unlike other temporal alignment datasets such as Charades-STA [2] and DiDeMo [3]. Further, we have response sheets of the interviews. There also exist inter-question dependencies. These are directions based on some answer and are mainly of two kinds – based on the answer to a previous question, skip a question or skip to a question. The questions can be categorized as multiple choice, multiple select, numerical, subjective, or YES/NO. Finally, we have access to the questions in two languages – English and Hindi. Note that the interviews are in Hindi (or local dialects thereof) and we wish for our queries to be in English.

## 2.2 High-level Approach

Let us formally express the problem statement. The inputs to our model are an audio file $a$ and a text query $q$. The model is then asked to output a tuple of start and end (terminal) times $(s, t)$ that localize the query in the audio file. To this end, we encode audio and text into embeddings in high-dimensional spaces. We segment the audio file into several chunks and compute their cosine similarity with the query in the representation space – which works as semantic similarity due to the way the model is trained.



भाग-2/ SECTION-2

प्रसव पूर्व देख-भाल और जन्म की तैयारी/ ANTE NATAL CARE AND BIRTH PREPAREDNESS

| Q. NO | QUESTIONS AND FILTERS | CODING CATEGORIES | SKIP TO |
|---|---|---|---|
| **अब मै आप से (नाम) के जन्म के पहले की गयी प्रसव पूर्व तैयारियों एवं जन्म देने की तैयारियों से सम्बंधित कुछ सवाल पूंछना चाहूँगा /चाहूंगी तथा उस दौरान आशा/आंगनबाड़ी कार्यकता द्वारा दी गयी सलाह के बारे में पूंछना चाहूँगा /चाहूंगी**<br>**Now I am going to ask you some questions related to antenatal care at the time of your last birth and Counselling that you received from FLW when you were pregnant with (NAME).** | | | |
| 201 | आपके सबसे छोटे बच्चे का नाम क्या है?<br>What is the name of your youngest child? | ______________________<br>(बच्चे का नाम /NAME OF THE CHILD) | |
| 202 | (नाम) लड़का या लड़की है?<br>Is (NAME) male or female? | लड़का /MALE ..........................1<br>लड़की /FEMALE ..........................2 | |
| 203 | (नाम) का जन्म तिथि/ समय क्या है?<br>What is the date of birth of (NAME) and time of birth?<br>[उम्र को महीने में दर्ज करें साथ ही जन्मतिथि भी दर्ज करें]<br>[RECORD AGE IN DAYS ALONG WITH THE DATE OF BIRTH] | जन्म तिथि / □□ □□ □□□□<br>DATE OF BIRTH D D M M Y Y Y Y<br>समय/TIME □□ □□<br>HH MM | |
| 204 | क्या आपके पास (सरकारी) एमसीपी कार्ड है?<br>Do you have an MCP card for (NAME)?<br>**कार्ड दिखाऐं। [SHOW CARDS]** | हाँ / YES..........................1<br>नहीं / NO..........................2 → | <br>Q205 |
| 204x | आपको यह कार्ड कब प्राप्त हुआ ?<br>When did you receive this card? | गर्भावस्था के दौरान / DURING PREGNANCY..........1<br>प्रसव के बाद / AFTER DELIVERY..........2 | |
| 204a | कृप्या कार्ड दिखाऐं।<br>Please show the card. | कार्ड देखा /CARD SEEN..........................1<br>कार्ड नहीं देखा /CARD NOT SEEN..........2 → | <br>Q205 |

Figure 1: A few sample questions from the 0-2 age group. Note the bilinguality, the various types of questions, and the directions to skip to a question.

More concretely, given an audio segment $a$, we generate embeddings $\{\mathbf{a}_k\}_{k=1}^{n_a}$ where $n_a$ represents the number of time frames in the segment. Similarly, we compute embeddings $\mathbf{q}$ for the given question $q$. While training, we know that $q$ is asked in $a$. So we use text-guided attention (described below) to aggregate the $n_a$ embeddings into one feature vector $\mathbf{f}$. We view $(\mathbf{f}, \mathbf{q})$ as a positive pair and reward the model for increasing the similarity between them. All non-positive pairs (there will be $n_q^2 - n_q$ of these if the number of questions asked in $a$ is $n_q$) are counted as negative pairs.

## 2.3 Dataset Details

The interviews are part of a quantitative household survey for health and nutrition of mothers in rural Bihar. These were conducted by CARE India over a span of several years. For now, in order to present a proof-of-concept, we operate on a small subset of the data. We have a total of 100 interviews split evenly into five categories based on the age of the mothers' children. Each age group has its own questionannaire. The interviews are long – their duration ranges from 30 to 60 minutes. We set aside one age group (12-23) as the test set and use the other age groups (0-2, 3-5, 6-8, and 9-11) for training. This split helps us assess the generalization power of our trained model.

We augment the test set with rich time-stamp annotations, i.e. we identify and mark the start and end times of every question asked in each audio file. Even if our training approach is weakly supervised, we need these annotations to evaluate the performance of our model.

## 2.4 Audio Processing

For our task, the interviewee's speech is irrelevant. So as an initial preprocessing step, we apply diarization [4] to separate the interviewer speech from that of the interviewee (and noise), i.e. partition the audio track into many small segments, each of which is attributed to one speaker. We then stitch the interviewer segments into one large audio file while also remembering the partition time-stamps.

The audio contains a lot of noise. To tackle this, we use demucs [5], a music source seperation tool. Demucs splits an audio track into multiple tracks which can be superimposed to recover the original audio. These tracks correspond to "vocals", "bass", "drums", and "other" (guitar, piano, etc.). We choose the "vocals" track as the filtered audio file and generate embeddings from it. Even though our audio does not contain music, we found that demucs does a good job at noise removal.

## 2.5 Training

### 2.5.1 A Tale of Two Approaches

We propose two approaches. These differ in the amount of supervision used. In the first approach we don't rely on any annotations. Instead, we process the interview response sheets to ascertain which questions were asked by looking at which questions have answers next to them. This is not foolproof, as occasionally the interviewers fill in answers themselves without asking a question, usually when the answer is obvious from the setting (eg. "do you have an air conditioner?"), but it is a good start. Now we could train using the original audio file. However, we are limited by its huge length (30-60 minutes) and need a way of creating smaller audio chunks.

For this, we turn to speech recognition. Concretely, we perform ASR [6] on the diarized segments $a_j$ of the audio file $a$ to obtain Hindi sentences $h_j$. We generate embeddings $\mathbf{s}_j$ of these sentences using LaBSE [7, 8], and do the same for the English questons from the questionnaire to get embeddings $\mathbf{q}_i$. Note that we need LaBSE's multilingual embeddings as the next step is to compute cosine similarity between $\mathbf{s}_j$ and $\mathbf{q}_i$. This process gives us some matches which we pivot on to partition our audio file into smaller chunks. We call this approach the "Heuristic" approach.

In the second approach, we get annotators to tell us which questions were asked in an audio segment (but not their time-stamps – that would be identical to the test set annotation). We demand around 15 questions per audio segment, and this results in segments whose duration is 1 to 4 minutes. This is manageably small, and also gives us a large enough pool of questions, which helps contrastive learning. This approach is called the "Weakly-supervised" approach.

**Embeddings.** We use Monsoon NLP's Hindi-BERT [9] to generate text embeddings and Wav2Vec 2.0 [10] to generate audio embeddings.

### 2.5.2 Text-guided Attention

Now, suppose that a question $q$ occurs in an audio segment $a$. Next, as described in section 2.2, we compute audio embeddings $\{\mathbf{a}_k\}_{k=1}^{n_a}$ and a question embedding $\mathbf{q}$. We then compute similarity scores between the $k$th time frame of $a$ and the question $q$ as $s_k = \langle \bar{\mathbf{a}}_k, \mathbf{q} \rangle$, where $\bar{\mathbf{a}}_k$ has the same dimensionality as $\mathbf{q}$ and is obtained by applying a fully-connected layer followed by a ReLU [11] and a Dropout layer [12] on $\mathbf{a}_k$. $\langle . \rangle$ here denotes a normalized vector dot product, also called cosine similarity. A softmax along the temporal dimension gives us attention scores $\alpha_k$ as follows.

$$\alpha_k = \frac{\exp(s_k)}{\sum_{k=1}^{n_a} \exp(s_k)}$$

These scores represent local similarity between questions and audio frames. We use these scores to pool (or aggregate) the audio features of a segment and get $\mathbf{f} = \sum_{k=1}^{n_a} \alpha_k \mathbf{a}_k$. This approach of pooling audio features is called text-guided attention (TGA) [13].

### 2.5.3 Loss Function

Observe that $\mathbf{f}$ ($\in \mathbb{R}^V$) and $\mathbf{q}$ ($\in \mathbb{R}^T$) have different dimensions. So we bring them to a shared space by again applying a fully-connected layer and obtain projections $\mathbf{f}^p$ and $\mathbf{q}^p$ (both $\in \mathbb{R}^D$). Let the set of all positive pairs $(\mathbf{f}^p, \mathbf{q}^p)$ be $S_+$ and the set of all non-positive pairs we $S_-$ We optimize the following max-margin contrastive loss $L$ [14]. Here, $m$ is the margin of the max-margin loss, and is a hyperparater of the algorithm.

$$\begin{aligned} L = \sum_{(\mathbf{f}^p, \mathbf{q}^p) \in S_+} \Bigg\{ & \sum_{(\mathbf{f}^p, \mathbf{q}^p_-) \in S_-} \max(0, m - \langle \mathbf{f}^p, \mathbf{q}^p \rangle + \langle \mathbf{f}^p, \mathbf{q}^p_- \rangle) \\ & + \sum_{(\mathbf{f}^p_-, \mathbf{q}^p) \in S_-} \max(0, m - \langle \mathbf{f}^p, \mathbf{q}^p \rangle + \langle \mathbf{f}^p_-, \mathbf{q}^p \rangle) \Bigg\} \end{aligned}$$

## 2.6 Evaluation

We evaluate our performance using a rank-based metric called "R@$K$, IoU=$m$" (Recall at $m$ and Intersection-over-Union $> m$) [2]. This calculates the percentage of test samples for which at least one of the top-$K$ audio segment retrieved for the query sample $q$ has an IoU value larger than $m$. The audio segments obtained after diarization are ranked by averaging the TGA scores of their respective time frames.

# 3 Other Threads

During the course of this thesis we explored several other project ideas. Here are a couple that we spent at least a few of weeks on.

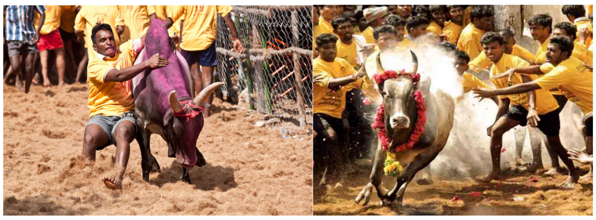
## 3.1 Audio Robustness to Cultural Variations

**Inspiration.** In this paper [15], the authors argue that the concepts and images of most vision-language (V-L) datasets are built on top of ImageNet and its WordNet-based hierarchy. To quote: "the coverage of the original distribution [of these datasets] does not encompass multiple languages and cultures". As a result, models trained on these datasets do not generalize well to tasks which are not in English or are not rooted in American/Western culture. The authors demonstrate this by creating a multilingual multimodal dataset called MaRVL and benchmarking several SoTA V-L models (some monolingual, some multilingual) on this dataset. They also evaluate these models on NLVR2, an Englsh dataset with the same task, to show the relative drop in performance on MaRVL.

**Proposal.** We propose that audio events like natural sounds (fire, rain, storms) and the sounds of things (vehicles, kitchens, tools) are more robust to these linguistic and cultural shifts than natural language. If so, we may expect gains on tasks like text to video retrieval by using audio as an input for the retrieval (A+T → V).

**Comments.** MaRVL's task is called visually grounded reasoning. Given two images as context, agents have to predict whether a natural language statement is true or false. See figure 2 for reference.

There are some concerns with porting MaRVL's idea into the video retrieval domain. For starters, Retrieval may not be a challenging enough task. To quote MaRVL: "We choose this specific task . . . as it requires the integration of information across modalities and deep linguistic understanding, rather than just matching superficial features". Further, MaRVL's analysis is focused solely on transformer based V-L models (ViLBERT [16], UNITER [17], etc.), i.e. models that encode text and images using one multimodal transformer. On the other hand, many video retrieval models (MMV [18], CLIP4Clip [19]) have separate encoders for individual modalities. Finally, how do we effectively mix audio with text to mitigate the fall in performance? This is the whole subproblem of mixup.



(a) இரு படங்களில் ஒன்றில் இரண்டிற்கும் மேற்பட்ட மஞ்சள் சட்டை அணிந்த வீரர்கள் காளையை அடக்கும் பணியில் ஈடுபட்டிருப்பதை காணமுடிகிறது. ("In one of the two photos, more than two yellow-shirted players are seen engaged in bull taming."). Label: TRUE.

Figure 2: An example from MaRVL [15]

**A crude method.** Let's say we have some English video-text dataset D. We can translate the text to a different language L and obtain a dataset D'. We can evaluate the text-to-video retrieval on both datasets and expect a drop in performance (D→D'). We repeat the same experiment but now with audio in the fray. The hypothesis is that the drop now will be smaller. It is worth noting that MaRVL also translates its test data to English to benchmark monolingual models (but the direction of translation is opposite) – the so-called 'translate test' approach. But they do this to enable comparison with NLVR2. In our crude method, there is a linguistic shift but no cultural shift.

### 3.2 Ego4D [1]

**What is it?** Ego4D is a "massive-scale, egocentric dataset and benchmark suite collected across 74 worldwide locations and 9 countries [855 unique camera wearers], with over 3,025 hours of daily-life activity video." People wear cameras (head-mounted) and record activities like walking, cooking, cleaning, etc. These videos are mostly unscripted and 'in the wild'. Due to the diversity in the videos, we considered Ego4D as a test-bed for evaluating audio robustness to cultural variations.

**Motivation.** Computer Vision has come very far by using internet images (and videos) for its datasets (eg. ImageNet, Kinetics, MS COCO, etc.). But these datasets are well-curated, disembodied, in isolated moments of time, and in the third-person view. These limitations hamper progress in robot learning and augmented reality research. Ego4D provides uncu-

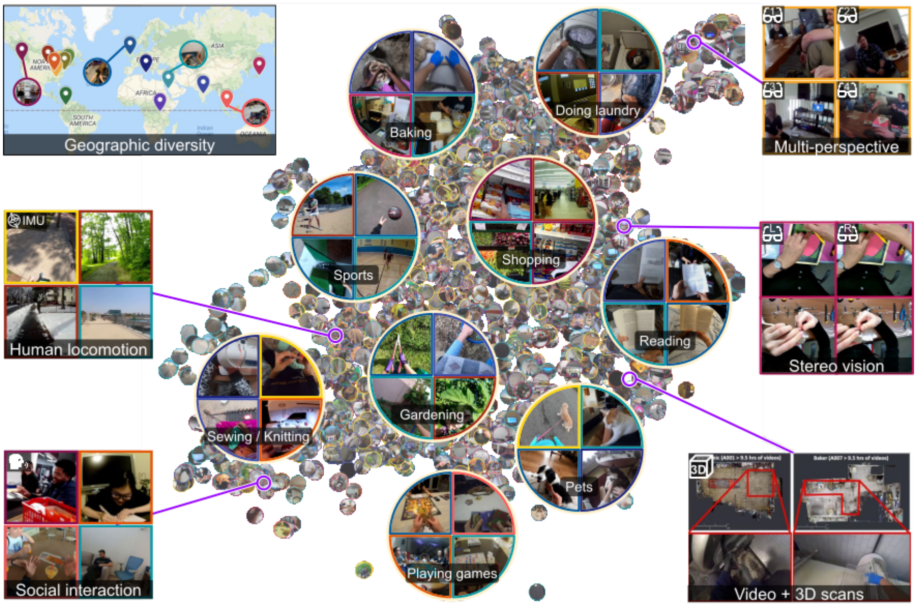


Figure 3: Ego4D captures people from various places doing a wide range of activities from multiple viewpoints [1].

rated, multimodal, first-person streams of data that depend on agents' attention and actions.

**Novelty.** Some existing egocentric datasets: EPIC-Kitchens, UT Ego, EGTEA Gaze+, etc. Ego4D improves on the scale (no. of hours of video as well as no. of participants), diversity (74 locations), and realism.

**Dataset details.** People wore cameras for 1-10 hrs $\implies$ long-form videos capturing many activities (many of which are probably mundane). "Portions of the dataset also provide audio, 3D mesh scans, gaze, stereo, and/or synchronized multi-camera views". Multiple cameras are used, and this variety gives rise to different fields of view (eg. some look down on hands, others look ahead). 2207 hours from the total 3025 hours have audio associated with the videos.

**Narrations.** All videos have temporally dense narrations (on average 13.2 sentences per minute of video) associated with them that describe the activities taking place (on average 7.4 words long). These narrations are manually annotated and amount to 3.85M sentences in total containing 1772 unique verbs (actions) and 4336 unique nouns (objects). Videos are also divided into 5 min clips that have a multiple-sentence summary associated with them. Ego4D can thus be viewed as a large-scale dataset of aligned language and video, similar to HowTo100M. Gives way to many grounded vision-language tasks like retrieval and captioning.

**Benchmarks**. Egocentric vision demands novel approaches of video understanding to account for long-form video, attention cues, person-object interaction, multi-sensory data, and the lack of manual temporal curation inherent to a passively worn camera. The authors also develop baseline models for these tasks using SoTA components. The tasks are Episodic Memory, Hands and Objects, Audio-visual Diarization, Social Interactions, and Forecasting.

# 4 Conclusion

In this report we studied the problem of audio-text alignment in Hindi interviews. This is an application that has potential to assist analysis of health data of people from rural and underrepresented communities; it is an example of AI for social good. There is much work left. A major rate-determining step throughout the project was the annotation process – a very direct experience that further cements the importance of self-supervised approaches.

## Acknowledgements

I am forever indebted to my advisors, Prof. Preethi Jyothi and Prof. Ganesh Ramakrishnan for shaping me and my perspective towards research. I couldn't have asked for more supportive, understanding, yet critical mentors, who guided me at every step of the research process. I am also grateful to my crew-mates: Shubham, Pranjal, Rishabh, and Pranjal for all the invigorating and insightful conversations they gave me, along with the motivation to get things done. I wish them all the very best for their future endeavours. I am grateful to my family for supporting me, especially because most of this project was done from home due to the Covid-19 pandemic. I am also grateful to my friends and batch-mates for making me explain my work to them in layman terms and for keeping my morale up through the thesis timeline. I am grateful to the CSE dept for giving its undergraduates the opportunity to do research under professors through the BTP and R&D courses. Finally, I am grateful to you, the reader. You're the reason this report exists.

## References

[1] K. Grauman *et al.*, "Ego4d: Around the world in 3, 000 hours of egocentric video," *CoRR*, vol. abs/2110.07058, 2021. 1, 6, 7

[2] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "Tall: Temporal activity localization via language query," 2017. 2, 5

[3] L. A. Hendricks, O. Wang, E. Shechtman, J. Sivic, T. Darrell, and B. Russell, "Localizing moments in video with temporal language.," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 2

[4] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," 2019. 4

[5] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021. 4

[6] A. Zhang *et al.*, "Speechrecognition 3.8.1." https://pypi.org/project/SpeechRecognition/, Last accessed on 2022-05-14. 4

[7] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," *CoRR*, vol. abs/2007.01852, 2020. 4

[8] Google, "sentence-transformers/labse." https://huggingface.co/sentence-transformers/LaBSE, Last accessed on 2022-05-14. 4

[9] N. Doiron, “Monsoon nlp - hindi bert,” 2020. https://huggingface.co/monsoon-nlp/hindi-bert, Last accessed on 2022-05-14. 4

[10] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *CoRR*, vol. abs/2006.11477, 2020. 4

[11] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML’10, (Madison, WI, USA), p. 807–814, Omnipress, 2010. 4

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 06 2014. 4

[13] N. C. Mithun, S. Paul, and A. K. Roy-Chowdhury, “Weakly supervised video moment retrieval from text queries,” *CoRR*, vol. abs/1904.03282, 2019. 5

[14] K. Q. Weinberger, J. Blitzer, and L. Saul, “Distance metric learning for large margin nearest neighbor classification,” in *Advances in Neural Information Processing Systems* (Y. Weiss, B. Schölkopf, and J. Platt, eds.), vol. 18, MIT Press, 2005. 5

[15] F. Liu, E. Bugliarello, E. M. Ponti, S. Reddy, N. Collier, and D. Elliott, “Visually grounded reasoning across languages and cultures,” *CoRR*, vol. abs/2109.13238, 2021. 5, 6

[16] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” *CoRR*, vol. abs/1908.02265, 2019. 6

[17] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “UNITER: learning universal image-text representations,” *CoRR*, vol. abs/1909.11740, 2019. 6

[18] J. Alayrac, A. Recasens, R. Schneider, R. Arandjelovic, J. Ramapuram, J. D. Fauw, L. Smaira, S. Dieleman, and A. Zisserman, “Self-supervised multimodal versatile networks,” *CoRR*, vol. abs/2006.16228, 2020. 6

[19] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, “Clip4clip: An empirical study of CLIP for end to end video clip retrieval,” *CoRR*, vol. abs/2104.08860, 2021. 6