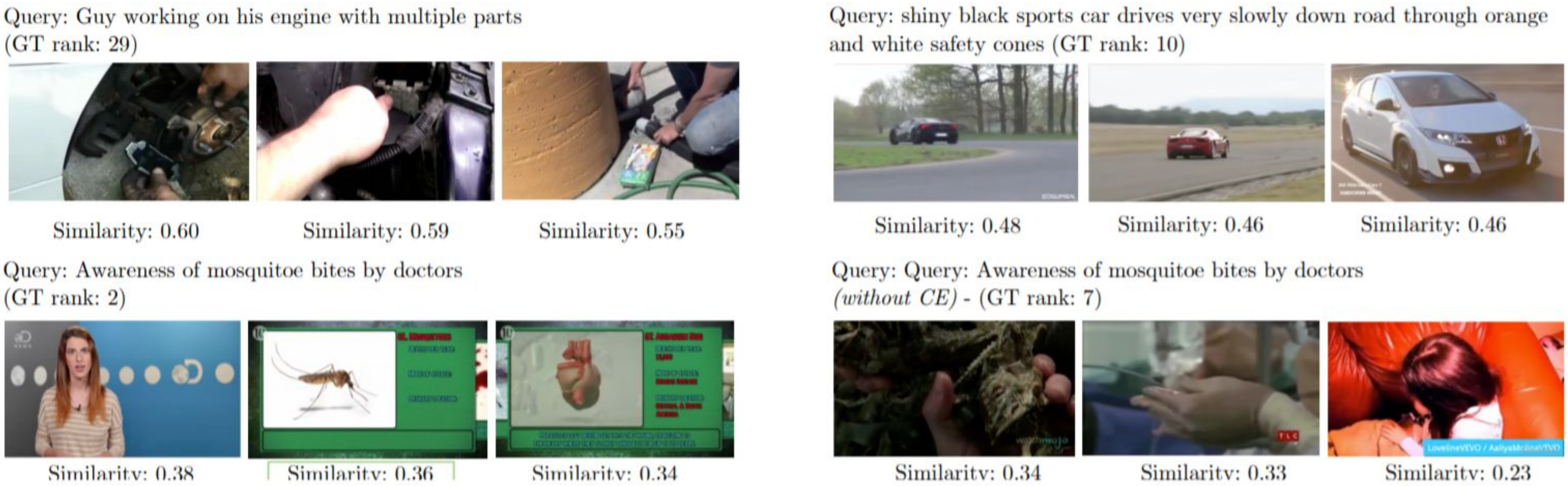
Multi-Modal Multi-lingual Video retrieval

Guided by,
Prof. Preethi Jyothi
Prof. Ganesh Ramakrishnan

Vighnesh Reddy
Mayur Warialani
Achari Rakesh Prasanth
Jayaprakash Akula

Problem Statement

- The video-text retrieval task focuses on returning for each query text, a ranked list of the most likely videos available in dataset and vice-versa.



Video Retrieval Using Representations From Collaborative Experts
https://arxiv.org/pdf/1907.13487.pdf

Dataset- MALTA

Dataset with experiments on science topics and farming. Close to 600 videos in total with average length of ~80 seconds. Approximately 7 descriptions per video.

Query- You need colour paper strips, thin straws, single punch, scissor and Cello-tape.

Description

Caption - Bend small paper strips and glue them to make lower wings.

Identity

Partial

Caption - You need Plastic straws, Woolen cloth, scissor, Single punch and cello-tape.

Caption - You just need some plastic straws and scissor

Caption - The twisted lugs and paper strips act like fan blades.(No relation)

https://www.cse.iitb.ac.in/~malta/

"Caption Alignment for Low Resource Audio-Visual Data", V. R. Konda et al., Submitted to Interspeech 2020.

Regular Contrastive Loss

All the work for cross-modal video-text retrieval till now is based on optimizing variants of contrastive loss.

$$\mathcal{L}_c = \frac{1}{2N}\sum_{i=1}^{N}[(1-y_i)||f_{1,i}-f_{2,i}||_2^2 + (y_i)\{max(0, m-||f_{1,i}-f_{2,i}||_2)\}^2]$$

The labels are either $\mathbf{y_i} = 0$ for positive pairs or $\mathbf{y_i} = 1$ for negative pairs.

Optimal transport distances

Assuming we are given two batches of samples, each batch has ***n*** examples $\boldsymbol{X} \in R^{d\times n}$. Let $x_i \in R^d$ be the representation of the i th shape. Additionally, let ***r*** and ***c*** be the ***n***-dimensional probability vectors for two batches, where $\boldsymbol{r_i}$ and $\boldsymbol{c_i}$ denote the number of times shape i occurs in ***r*** and ***c***

$$\boldsymbol{D}_{\mathrm{OT}}^{\lambda}(\boldsymbol{r},\boldsymbol{c}) = \min_{\boldsymbol{T}\geq 0} \sum\nolimits_{i,j=1}^{n} \boldsymbol{T}_{ij}\boldsymbol{M}_{ij} - \frac{1}{\lambda}h(\boldsymbol{T}_{ij})$$
$$\text{s.t.} \quad \sum\nolimits_{j=1}^{n} \boldsymbol{T}_{ij} = \boldsymbol{r} \quad \text{and} \quad \sum\nolimits_{i=1}^{n} \boldsymbol{T}_{ij} = \boldsymbol{c} \quad \forall i,j.$$

Thus, $\boldsymbol{T}^*$ solved by above equation prefers to assign higher importance values to samples with small ground distances while leaving fewer for others.

Ground Distances

For a pair of similar positive samples:

$$G^{+}_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j; f) = e^{-\gamma||f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)||_2^2},$$

For a pair of negative samples:

$$G^{-}_{ij}(\boldsymbol{x}_i, \boldsymbol{x}_j; f) = e^{-\gamma \max\{0, \varepsilon - ||f(\boldsymbol{x}_i) - f(\boldsymbol{x}_j)||_2^2\}}.$$

γ is a hype-parameter controlling the extent of rescaling.

Batch-wise Optimal Transport Loss

$$\mathbf{T}^* = argmin_{\mathbf{T}} \sum_{i,j=1}^{n} \mathbf{Y}_{i,j}\mathbf{T}_{i,j}\mathbf{G}^{+}_{i,j} + \sum_{i,j=1}^{n} (1 - \mathbf{Y}_{i,j})\mathbf{T}_{i,j}\mathbf{G}^{-}_{i,j}$$

$$Loss = \sum_{i,j=1}^{n} \mathbf{T^*}_{i,j}\mathbf{M}_{i,j}$$

where $\mathbf{Y}_{ij}$ is a binary label assigned to a pair of training batches. Let $\mathbf{Y}_{ij}$ = 1 if sample $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are deemed similar, and $\mathbf{Y}_{ij}$ = 0 otherwise.

Experiments on MSRVTT

Loss	R@1	R@5	R@10	Median	Mean
OT	10.1	29.4	41.6	16	85.8
Max-Margin	10	29	41.2	16	86.8

Results on the Video retrieval given a Text query

Loss	R@1	R@5	R@10	Median	Mean
OT	13.5	37.2	51.1	10.2	46
Max-Margin	15.6	40.9	54.5	8.3	38.1

Results on the Text retrieval given a Video query

Partial Order Contrastive Loss

V, T are set of Videos and Captions and Let m1 < m2 be the margins (hyperparameters)

$$L_{v,t}(\theta) = \sum_{(\mathbf{i},\mathbf{j},\mathbf{k},\mathbf{l}) \epsilon \mathbf{S^{v,t}}} max(m_1 + d(f_{v_i}, g_{t_j}) - d(f_{v_i}, g_{t_k}), 0) + max(d(f_{v_i}, g_{t_k}) - d(f_{v_i}, g_{t_j}) - m_2, 0) + max(m_2 + d(f_{v_i}, g_{t_j}) - d(f_{v_i}, g_{t_l}), 0)$$

$$\mathbf{S^{v,t}} = \{(i, j, k, l) | v_i \epsilon V, t_j \epsilon T_{i+}, t_k \epsilon T_{i\Theta}, t_l \epsilon T_{i-}\}$$

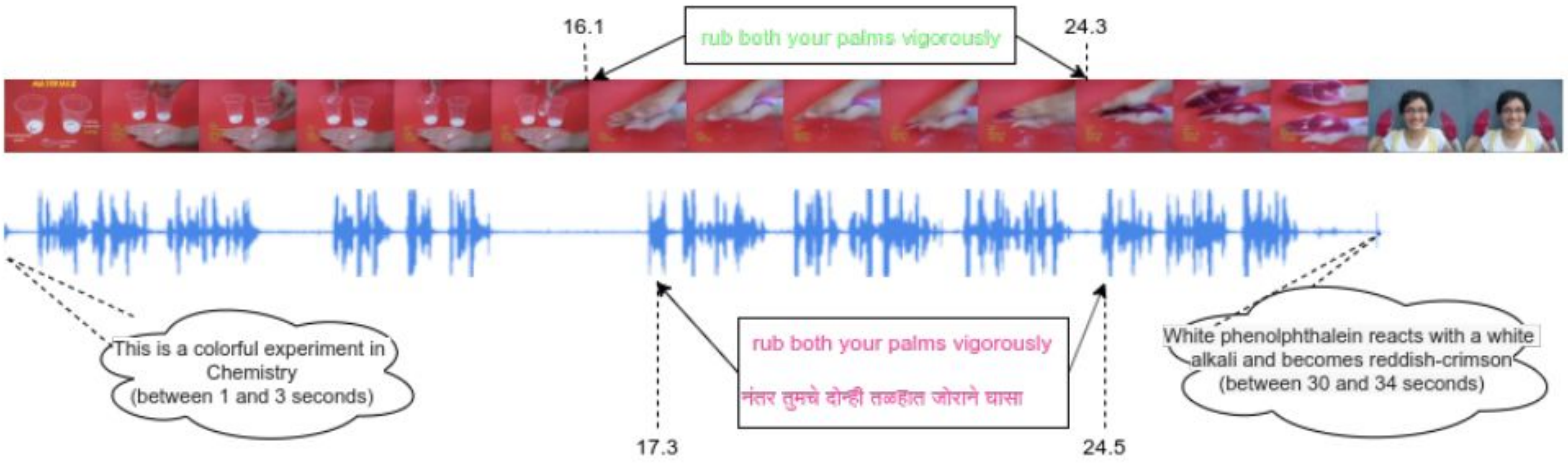
we want d(i,l) to be at least m2 distance away from d(i,j) and d(i,k) to be at most m2 distance away while at least being m1 distance away from d(i,j)

if the above loss improves the rankings then the OT variant of it should also improve the results

Caption Alignment for Low Resource Audio-Visual Data

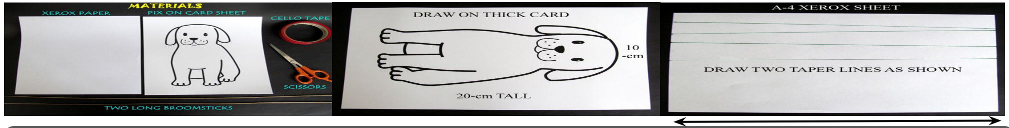
Problem Statement

For a Video, with Audio track, identify the correct start and end times of where information related to a given sentence appears in that video.



Datasets

MALTA- TFTav
-492 videos, average length of 80 seconds, 7 sentences per video
-Educational videos, Marathi Speech and English/Marathi Captions

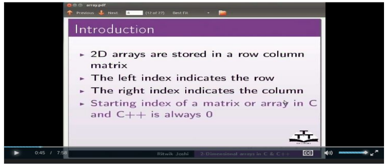


Draw two tapered lines on a xerox sheet as shown.

MALTA- ATMAav
-95 videos, average length of 111 seconds, 18 sentences per video
-Educational videos on Farming, Marathi Speech and Marathi Captions

आपल्याला सेंद्रीय तीसाठी लागणाया निविष्ठांपैकी..

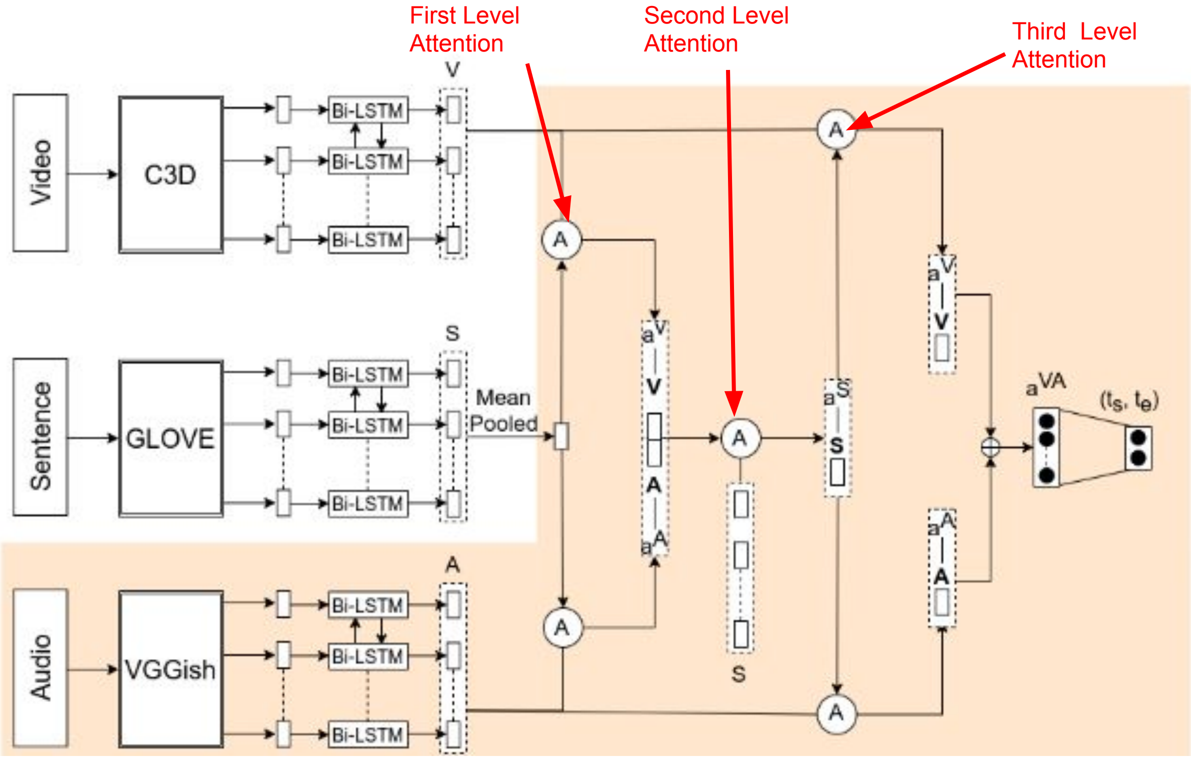
Spoken Tutorials
-100 hours of data in Marathi/English Speech and Transcriptions.
-Educational videos on C an Cpp, Java, Biogas, etc



C-and-C++/C3/Working-With-2D-Arrays/Marathi

Time	Narration
00:01	C आणि C++ मधील **2Dimensional Arrays** वरील स्पोकन ट्यूटोरियल मध्ये आपले स्वागत.
00:08	या ट्युटोरियलमध्ये आपण शिकू,
00:10	2Dimensional array म्हणजे काय आहे?
00:13	आपण यास उदाहरण द्वारे करू.
00:16	हे ट्यूटोरियल रेकॉर्ड करण्यासाठी मी,
00:18	उबुंटु ऑपरेटिंग सिस्टम वर्जन 11.10,
00:22	उबुंटु वर gcc आणि g++ Compiler version 4.6.1 वापरत आहे.
00:29	2 dimensional Array च्या परिचया सह प्रारंभ करूया.

Modifications to the baseline

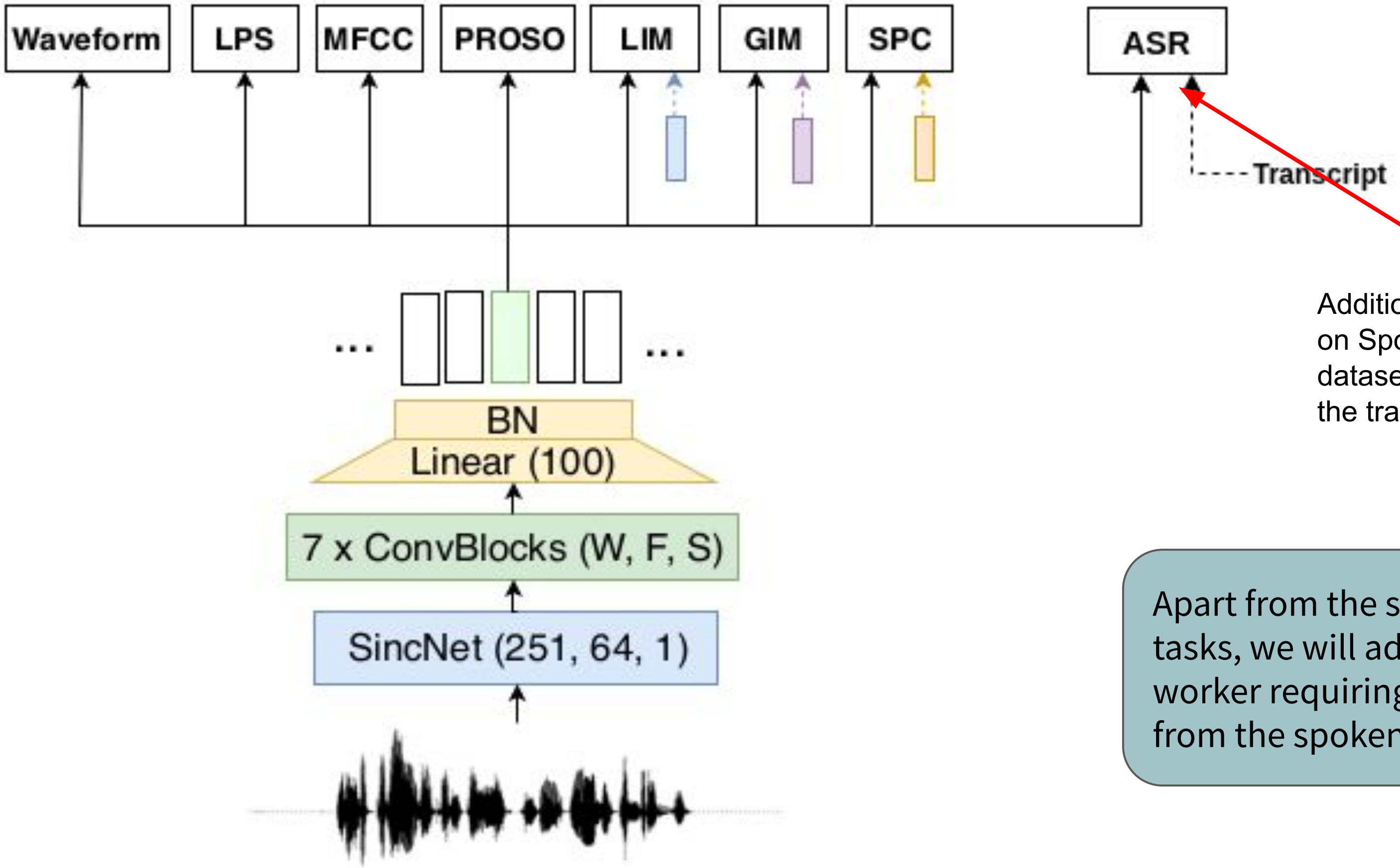


conc-AV Model

Here, 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.

Adding new worker to pase



Additional Pase worker on Spoken Tutorial dataset, as we have the transcripts.

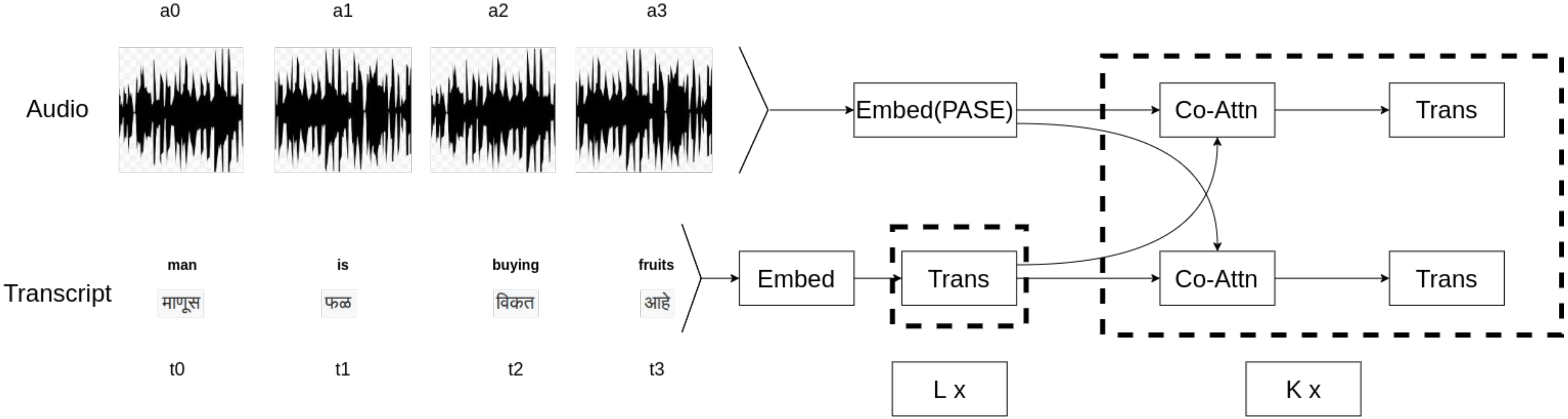
Apart from the seven, unsupervised tasks, we will add an additional worker requiring supervised data from the spoken tutorial dataset.

Experimental Results

Audio - Feature	IOU>=0.5	IOU>=0.7
-	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 tft scratch	0.1387±0.006	0.0474±0.003
PASE spk scratch	0.1375±0.006	0.0496±0.002
PASE tft+spk finetuned	0.1478±0.006	0.0462±0.005
ASR-bnf	0.1550±0.005	0.0545±0.005

Results on TFT av (A sub part of MALTA dataset)

Using transcript as additional input to pase



The input features (audio/transcript) along with its positional encodings is fed as input to the model.

The audio features ie PASE features are at a very high level of abstraction, so we pass transcript through a series of L transformers

ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations https://arxiv.org/pdf/1907.13487.pdf

References

- To find where you talk: Temporal sentence localization in video with attention based location regression. https://arxiv.org/pdf/1804.07014.pdf
- Temporal grounding of natural language sentence in video. https://www.aclweb.org/anthology/D18-1015/
- Structured Optimal Transport - https://arxiv.org/abs/1712.06199
- Hierarchical Optimal Transport for Multimodal Distribution Alignment
- ToysFromTrash http://www.arvindguptatoys.com/toys.html
- Video Retrieval Using Representations From Collaborative Experts https://arxiv.org/pdf/1907.13487.pdf
- Learning with Batch-wise Optimal Transport Loss for 3D Shape Recognition http://openaccess.thecvf.com/content_CVPR_2019/papers/Xu_Learning_With_Batch-Wise_Optimal_Transport_Loss_for_3D_Shape_Recognition_CVPR_2019_paper.pdf
- Learning Problem-agnostic Speech Representations https://arxiv.org/pdf/1904.03416.pdf

Adding new worker to PASE