Learning Sample Importance for Cross-Scenario Video Temporal Grounding

Peijun Bao, Yadong Mu
Peking University,
{peijunbao, myd}@pku.edu.cn

Abstract

The task of temporal grounding aims to locate video moment in an untrimmed video, with a given sentence query. This paper for the first time investigates some superficial biases that are specific to the temporal grounding task, and proposes a novel targeted solution. Most alarmingly, we observe that existing temporal ground models heavily rely on some biases (*e.g.*, high preference on frequent concepts or certain temporal intervals) in the visual modal. This leads to inferior performance when generalizing the model in cross-scenario test setting. To this end, we propose a novel method called Debiased Temporal Language Localizer (Debias-TLL) to prevent the model from naively memorizing the biases and enforce it to ground the query sentence based on true inter-modal relationship. Debias-TLL simultaneously trains two models. By our design, a large discrepancy of these two models' predictions when judging a sample reveals higher probability of being a biased sample. Harnessing the informative discrepancy, we devise a data re-weighing scheme for mitigating the data biases. We evaluate the proposed model in cross-scenario temporal grounding, where the train / test data are heterogeneously sourced. Experiments show large-margin superiority of the proposed method in comparison with state-of-the-art competitors.

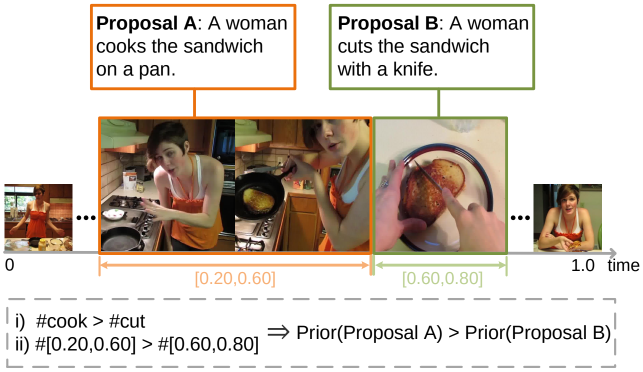

Proposal A: A woman cooks the sandwich on a pan.
Proposal B: A woman cuts the sandwich with a knife.
0
1.0
time
[0.20,0.60]
[0.60,0.80]
i) #cook > #cut
ii) #[0.20,0.60] > #[0.60,0.80]
⇒ Prior(Proposal A) > Prior(Proposal B)

Figure 1: Illustration of video uni-modal bias in ActivityNet. Even without knowing the sentence query, the prior of video moment proposal A to be grounded is larger than B, mainly due to two facts in the annotations: i) the visual concept "cook" contained in A are much more frequently queried than the concept "cut" in B; ii) the temporal interval $[0.2, 0.6]$ of A appears more frequently annotated than B. # denotes the frequency.

1 Introduction

Given a sentence query and an untrimmed video, the goal of temporal grounding [Anne Hendricks *et al.*, 2017; Gao *et al.*, 2017] is to localize video moment described by the sentence query. In recent years, a list of promising models [Wang *et al.*, 2020; Ghosh *et al.*, 2019; Rodriguez *et al.*, 2020; Wang *et al.*, 2019a; Hendricks *et al.*, 2018; Zhang *et al.*, 2019b; Liu *et al.*, 2018a; Stroud *et al.*, 2019; Yuan *et al.*, 2019; Zhang *et al.*, 2020] have been designed to tackle this task. Despite remarkable research progress, we empirically find that these models are heavily affected by some superficial bias of the data, leading to inferior generalization performance on cross-scenario testing data. In one of our pilot experiments, we take some well-trained state-of-the-art temporal grounding model and zero the feature vector of all testing queries. This boils down to using only the visual information in the temporal grounding. Surprisingly, the performance under such a setting is comparable to many models that normally read both queries and videos during testing. It also significantly outstrips random guess.

To make the point more clear, let us provide some concrete empirical observations. We have identified two sorts of dominant biases that a model can exploit for over-fitting the training scenarios, namely the visual content bias and temporal interval bias. In detail, a few visual concepts and temporal intervals are more frequently queried by the sentence than others in the training dataset. Although the issue of dominant biases have been previously reported as language bias in visual-question answering tasks [Ramakrishnan *et al.*, 2018; Cadene *et al.*, 2019; Chen *et al.*, 2020], the preference of visual contents and temporal intervals is specific to the temporal grounding task and our report here is the first. For instance, the concept "run" is largely frequent than "sit" in the benchmark of ActivityNet. Likewise, certain temporal intervals are more likely to be grounded. As illustrated in Figure 1, the interval $[0.20, 0.60]$ statistically has more annotations than $[0.60, 0.80]$. If above biases were sufficiently strong, a fully uni-modal input (such as zeroing query's features) can still achieve good performance in this multi-modal task. However, when generalizing the learned model into other unseen scenarios, these superficial biases between video moments and ground-truth may disappear, which adversely impacts the cross-scenario performance.

To this end, we propose a novel method called Debiased Temporal Language Localizer (Debias-TLL) to prevent the model from naively learning the video moment bias and enforce it to ground sentence in the video. Our key idea is to simultaneously train two twined models, with one of them aiming to learn video moment bias from the data and further to debias the other model. The models have an identical backbone. One of them reads only the video input, and the other normally has access to the full video-query input. The first model is designed to learn the video moment bias and predict the localization results only from visual modality. As illustrated in Figure 2, we then use the prediction of the first model to reweigh the importance of training samples for the second model and adjust the loss function accordingly. During this process, those training samples with high probability to being biased is suppressed. In this way, the training data is adaptively re-weighed in order to mitigate video moment bias in the second model. At the inference stage, we drop the first model and only use the second one for final prediction.

Note that the weakness of video modality biased model cannot be reflected by existing standard evaluation process because the training and testing data share a similar distribution of video moment correlation. To fairly evaluate the model, we propose a novel cross-scenario setting for the video temporal grounding task. In specific, we conduct the training and evaluation processes across two data distributions where video moment correlation cannot transfer from one data distribution to another. Under such settings, a model which makes prediction utilizing video moment bias would fail to perform well on the testing data.

Our contributions are summarized as follows:

1) To the best of our knowledge, we are the first to investigate the video moment bias in the video temporal grounding task, which adversely affects the generalization ability of the model. Two specific sorts of video moment biases in the data (visual content bias and temporal interval bias) are studied.

2) We propose a novel two-model-based methods to reweigh the training data via learning sample importance and delineate the video moment bias for a temporal grounding model.

3) A cross-scenario evaluation setting is proposed to reveal the weakness of video-moment biased temporal grounding. Our proposed method beats the state-of-the-art competitors with a clear margin under the cross-scenario settings.

2 Related Work

2.1 Temporal Grounding

The task of temporal grounding in video is recently introduced by [Anne Hendricks *et al.*, 2017; Gao *et al.*, 2017], which aims to determine the start and end time of the video moment described by a sentence query. [Anne Hendricks *et al.*, 2017] proposes a moment context network to jointly model text query and video clips. [Gao *et al.*, 2017] proposes cross-modal localizer to regress action boundary for candidate video clips. [Liu *et al.*, 2018b; Liu *et al.*, 2018c] advice to apply attention mechanism to highlight the crucial part of visual features or query contents. [Wang *et al.*, 2019b] then develops a semantic matching reinforcement learning framework to reduce the large visual-semantic discrepancy between video and language.

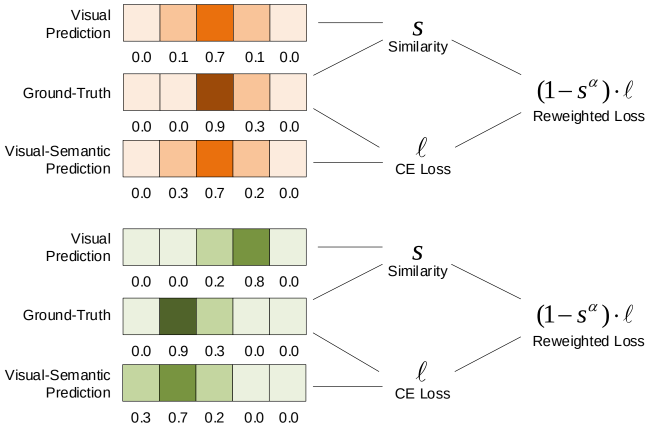


Figure 2: Sample importance reweighing. The visual localizer adjusts the loss function for the visual-semantic localizer, *i.e.*, suppressing the importance of training sample with high relevance to video moment bias (upper figure) and augmenting the weight of the irrelevant one (lower figure).

Several recent works [Zhang *et al.*, 2019a; Zhang *et al.*, 2020; Wang *et al.*, 2020] propose to model temporal dependencies within sentence to closely integrate language and video representation via graph convolution or non-local modules. And [Zhang *et al.*, 2019b; Stroud *et al.*, 2019; Zhang *et al.*, 2019c] further utilize compositional property of query sentence and decompose sentence as multiple components for better temporal reasoning.

2.2 Unbiased Cross-Modal Understanding

[Ramakrishnan *et al.*, 2018; Cadene *et al.*, 2019; Chen *et al.*, 2020] study language bias in visual question answering (VQA) caused by answer prior. [Chen *et al.*, 2020] proposes a model-agnostic counterfactual samples-synthesizing training scheme to reduce the language biases, which generates numerous counterfactual training samples by masking critical objects in images or words in questions. [Ramakrishnan *et al.*, 2018] introduces a question-only model, and then pose training as an adversarial game which discourages the VQA model from capturing language biases in its question.

Recently, [Tang *et al.*, 2020] studies the predicate bias on the task of scene graph generation from image. And [Qi *et al.*, 2020] investigates the dialog history bias in the visual dialog and proposes two causal principles for improving the quality of visual dialog. To the best of our knowledge, the video moment bias is specific to the temporal grounding task and is never explored before.

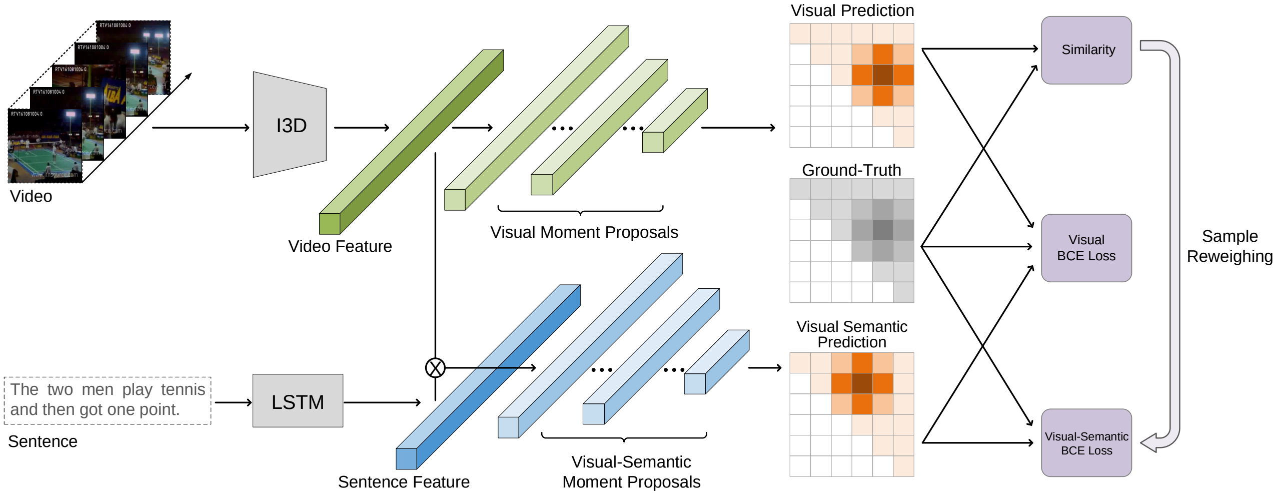


Figure 3: Our proposed network consists of four main components: a language encoder, a video encoder, a visual localizer, and a visual-semantic localizer. The visual localizer learns the video moment bias from the data with a single input of video modality. Then it adjusts the loss function for the visual-semantic localizer, *i.e.*, suppressing the importance of training sample with high relevance to video moment biases.

3 Methods

3.1 Problem Formulation

Given an untrimmed video V and a sentence description S, the goal of temporal grounding is to localize temporal moments T described by the sentence. More specifically, the video is presented as a sequence of frames $V = \{v_i\}_{i=1}^{L_V}$ where v_i is the feature of i-th frame and L_V is the frame number of the video. The sentence description S is presented as $S = \{s_i\}_{i=1}^{L_S}$ where s_i represents i-th word in the sentence and L_S denotes the total number of words. The temporal moment T is defined by the start and end time points of the moment in the video.

3.2 Debiased Temporal Language Localizer

As illustrated in Figure 3, our proposed Network consists of four main components: a language encoder, a video encoder, a visual-semantic localizer, and a visual localizer. This section will elaborate on the details of each component.

Video Encoder

Video encoder aims to obtain high-level visual representations of video moment proposals from raw input frames. Specifically, the input video is first segmented into small clips, with each video clip containing T frames. A fixed-interval sampling is performed to obtain N video clips. For each sampled video clip, we extract a sequence of ordinary spatio-temporal features $V = \{v_i\}_{i=1}^{N}$ with a pretrained I3D Network [Carreira and Zisserman, 2017].

The visual feature embeddings for moment proposals are constructed from these basic I3D features. For a moment proposal (a, b) with start point at a and end point at b, we apply boundary-matching (BM) operation [Lin *et al.*, 2019] over all I3D features covered by this proposal to get the feature embedding:

$$\tilde{f}^{V_{ab}} = \mathrm{BM}(\{v_i\}_{i=a}^{b}). \tag{1}$$

The boundary-matching operation can efficiently generate proposal-level feature from basic clip-level feature, through a series of bilinear sampling and convolutional operations. More algorithmic details are omitted here and can be referred to [Lin *et al.*, 2019]. $\tilde{f}^{V_{ab}}$ is passed through a fully-connected layer to obtain the final feature embedding $f^{V_{ab}} \in \mathbb{R}^{d^V}$ for the moment proposal (a, b). Essentially, this extracted feature $f^{V_{ab}}$ summarizes spatial-temporal patterns from raw input frame and thus represents the visual structure of the moment proposal.

Language Encoder

Given an input of a natural language sentence query, the goal of language encoder is to encode the sentence such that moments of interest can be effectively retrieved in the video. Our language encoder extracts feature embedding f^S of the sentence descriptions S.

Instead of encoding each word with a one-hot vector or learning word embeddings from scratch, we rely on word embeddings obtained from a large collection of text documents. In more details, each word s_i in S is first encoded into Glove word embedding [Jeffrey Pennington and Manning, 2014] as w_i. Then the sequence of word embedding $\{w_i\}_{i=1}^{L_S}$ is fed to an LSTM [Hochreiter and Schmidhuber, 1997]. The last hidden state of LSTM is passed to a single fully-connected layer to extract the final sentence feature $f^S \in \mathbb{R}^{d^S}$.

Visual Localizer and Visual-Semantic Localizer

We design two twined models i.e. visual localizer and visual-semantic localizer, with the visual localizer aiming to learn video moment bias from the data and further to debias the visual-semantic localizer. The visual localizer reads only the video input, and the visual-semantic localizer normally has access to the full video-query input.

The visual-semantic localizer first constructs visual-semantic features of moment proposals for the sentence query and then localizes the described moments. In specific, video moment feature $f^{V_{ab}}$ for all possible moment proposals are computed according to Eq. (1) where $1 \leq a \leq b \leq N$. The features of the visual modality and language modality are fused to generate visual-semantic features for each moment proposals. To interact the language feature f^S with video moment feature $f^{V_{ab}}$, we multiply f^S with video moment clip feature $f^{V_{ab}}$ and then normalize the fused feature $\hat{M}_{ab}$ with its $\mathcal{L}_2$ norm, namely

$$\begin{aligned} \hat{M}_{ab} &= f^{V_{ab}} \odot f^S, \\ M_{ab} &= \hat{M}_{ab} / ||\hat{M}_{ab}||_2, \end{aligned} \tag{2}$$

where $\odot$ denotes the Hadamard product.

Finally we pass the visual-semantic features $\{M_{ab}\}$ to a fully-connected layer and a sigmoid layer to generate the visual-semantic score map $\{p_{ab}\}$. Each value $p_{a,b}$ in the visual-semantic score map denotes the predicted matching score of the temporal moment (a, b) for the sentence query. The maximum of the score map p corresponds to the grounding result for the sentence query.

The visual localizer directly guesses the most interested moments based on the visual feature of moment proposals $\{f^{V_{ab}}\}$ without the input of sentence query. Specifically, the video moment features $\{f^{V_{ab}}\}$ are directly passed to a fully-connected layer and sigmoid layer to generate a visual score map $\{p'_{ab}\}$, which represents the predicted prior of video moment (a, b) to be grounded.

3.3 Sample Importance Reweighing

Each training sample consists of an input video V, a sentence query S and the temporal annotation T associated with the query. During training, we need to determine which temporal moment in the temporal-sentence score map corresponds to the annotations and train the model accordingly. Instead of hard label, we assign each moment proposal with a soft label according to its overlap with the annotations. Specifically, for each moment in the temporal-sentence score map, we compute the IoU score IoU_{ab} between its temporal boundary (a, b) and the annotation T. Then a soft ground truth label gt_{ab} is assigned to it according to IoU_{ab}:

$$gt_{ab} = \begin{cases} 0 & IoU_{ab} \leq \mu_{min}, \\ \frac{IoU_{ab} - \mu_{min}}{\mu_{max} - \mu_{min}} & \mu_{min} < IoU_{ab} < \mu_{max}, \\ 1 & IoU_{ab} \geq \mu_{max}, \end{cases} \tag{3}$$

where μ_{min} and μ_{max} are two thresholds to customize the distribution of soft labels.

The visual localizer's goal is to learn the video moment bias and predict the localization results only from visual modality. For each training sample, the visual localizer can be trained with a binary cross entropy loss, which is defined as:

$$\mathcal{L}_v = -\sum_{(a,b) \in \mathcal{C}} \mathrm{gt}_{ab} \log(p'_{ab}) + (1 - \mathrm{gt}_{ab}) \log(1 - p'_{ab}), \tag{4}$$

where $\mathcal{C} = \{(a, b) | 1 \leq a \leq b \leq N\}$ is the set of all valid moment proposal boundaries and p'_{ab} is the prediction output of visual localizer.

To train the visual-semantic localizer, previous works commonly train it with binary cross entropy loss similar to Eq. 4 as

$$\mathcal{L}_{vs} = -\sum_{(a,b) \in \mathcal{C}} \mathrm{gt}_{ab} \log(p_{ab}) + (1 - \mathrm{gt}_{ab}) \log(1 - p_{ab}), \tag{5}$$

where p_{ab} is the prediction of visual-semantic localizer.

Due to the superficial bias between video moments and ground-truth, the temporal grounding model trained with the cross entropy loss tends to simply exploit the video modality to make a prediction, rather than jointly understand both video and language as claimed before. To this end, we use the prediction output p'_{ab} of the visual localizer to reweigh the importance of training sample and adjust the loss function $\mathcal{L}_{vs}$ for the the visual-semantic localizer accordingly. Specifically, we first compute the cosine similarity s of visual localizer prediction $p' = \{p'_{ab}\}$ and ground-truth $\mathrm{gt} = \{\mathrm{gt}_{ab}\}$ as

$$s = \frac{p' \cdot \mathrm{gt}}{||p'||_2 ||\mathrm{gt}||_2} \tag{6}$$

Then the adjusted loss $\mathcal{L}'_{vs}$ for visual-semantic localizer is reweighted by s as follows

$$\mathcal{L}'_{vs} = (1 - s^{\alpha}) \cdot \mathcal{L}_{vs}, \tag{7}$$

where α is a hyper-parameter to control the weight decay. Intuitively, high value of s implies large probability of inferring the ground truth merely from the visual modal, implying tight relevance to the video moment biases. Following this intuition, Eq. 7 suppresses their sample weight.

Last, we define the total loss function for Debias-TLL as:

$$\mathcal{L}_t = \mathcal{L}_v + \mathcal{L}'_{vs}, \tag{8}$$

which consists of the loss $\mathcal{L}_v$ for visual localizer and the adjusted loss $\mathcal{L}'_{vs}$ for visual-semantic localizer.

With the final loss function $\mathcal{L}_t$, Debias-TLL can be trained in an end-to-end manner to mitigate video moment bias. At the inference stage, we drop the visual localizer and only use the visual-semantic localizer.

4 Experiment

4.1 Dataset

ActivityNet Captions. It consists of 19,209 untrimmed videos with the annotation of sentence description and moment boundary. The contents of the videos are diverse. It is originally built for dense-captioning events [Krishna *et al.*, 2017] and lately introduced for temporal grounding. It is the largest existing dataset in the field of temporal grounding. There are 37,417, 17,505, and 17,031 moment-sentence pairs in the training, validation and testing set, respectively.

Charades-STA. It contains 9,848 videos of daily indoors activities. It is originally designed for action recognition and localization. Gao et al. [Gao *et al.*, 2017] extend the temporal annotation (*i.e.*, labeling the start and end time of moments) of this dataset with language descriptions and name it as Charades-STA. There are 3,720 moment-sentence pairs in the testing set.

DiDeMo. It was recently proposed in [Hendricks *et al.*, 2018], specially for natural language moment retrieval in open-world videos. DiDeMo contains 10,464 videos with 4,021 annotated moment query pairs in the testing set.

4.2 Evaluation Metrics

The commonly-adopted evaluation metric in temporal grounding is known to be "Recall@N,IoU=θ ". For each sentence query we calculate the Intersection over Union (IoU) between a grounded temporal segment and the ground truth. "Recall@N,IoU=θ " represents the percentage of top N grounded temporal segments that have at least one segment with higher IoU than θ. Following previous works [Zhang *et al.*, 2020; Yuan *et al.*, 2019], we report the results as $N \in \{1, 5\}$ with $\theta \in \{0.5, 0.7\}$ for ActivityNet Captions, Charades-STA and DiDeMo dataset.

4.3 Baseline Methods

We compare our methods with several state-of-the-art methods listed as followings: **CTRL** [Gao *et al.*, 2017]: Cross-model Temporal Regression Localizer. **PFGA** [Rodriguez *et al.*, 2020]: Proposal-free Temporal Moment Localization using Guided Attention. **SCDM** [Yuan *et al.*, 2019]: Semantic Conditioned Dynamic Modulation. **2D-TAN** [Zhang *et al.*, 2020]: 2D Temporal Adjacent Networks. We further consider following methods: **random**: Randomly select the moment proposals. **TLL**: The model with identical archetecture to Debias-TLL, but trained by commonly used binary cross entropy loss.

4.4 Implementation Details

we use pretrained CNN [Carreira and Zisserman, 2017] as previous methods to extract I3D video features on all datasets And we use Glove [Jeffrey Pennington and Manning, 2014] word embeddings pretrained on Common Crawl to represent each word in the sentences. A three layer LSTM is applied to word-embeddings to obtain the sentence representation. The channel numbers of sentence feature and video proposal feature d^S, d^V are all set to 512 .The number of sampled clips N is set to 32. For BM operations in the video encoder, we set sampling number of each proposals to 32.

During training, We use Adam [Kingma and Ba, 2014] with learning rate of 1×10^{-4}, a momentum of 0.9 and batch size of 4. $\alpha = 1.0$ in Eq. 7. During inference, we choose the moment proposals with the highest confidence score for the sentence query as the final result. If it is desired to select multiple moment locations per sentence (*i.e.*, for R@5), non-maximum suppression (NMS) with a threshold of 0.4 is applied to remove redundant candidates.

4.5 Analysis of Video Moment Biases

To show the video moment bias in the temporal grounding model, in this experiment, we mask all the words of the sentence input and evaluate existing models with single video input (marked as "video-only") on the ActivityNet Captions. The results are summarized in Table 1, where all of these methods achieve better performance than the random method with large margins. This shows that the model can heavily exploit the superficial correlation between video moment and ground-truth to provide correct localization results.

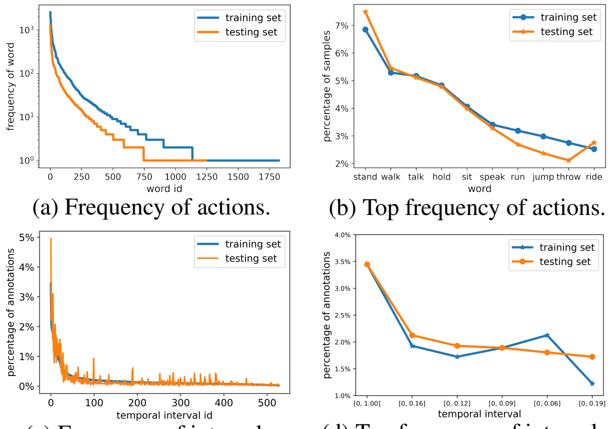
Here we further identify two sorts of specific biases (visual content bias and temporal interval bias) that the model can exploit to localize the target moment when ignoring the sentence query. Figure 4a presents the frequencies of action concepts in all the sentence queries in both training and testing data of the ActivityNet Captions. The action concept distribution follows a long-tail distribution, and some actions concept are much more frequently queried than others. We further illustrate the top frequency of action concepts in Fig. 4b, showing that training and testing data share a similar distribution. A similar conclusion also exists in the object concepts in the sentence query. Furthermore, we illustrate the frequency distribution and the top frequency of temporal intervals of the queried moments in Figure 4c and 4d. Like the video content, the temporal intervals also follow long-tail distributions shared by the training and testing data. The temporal information of the moment can be captured by the modern temporal grounding model with temporal context modeling using recurrent neural networks, non-local blocks, etc. This means video moment proposals that contain certain video concepts and temporal intervals are more likely to be localized as positive, and the model can then infer the localization results only

Table 1: Performance evaluation results on the ActivityNetCap.

Input	Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
	random	13.99	4.69	44.69	17.64
video & query	CTRL	29.01	10.34	59.17	37.54
	PFGA	33.04	19.26	-	-
	SCDM	36.75	19.86	64.99	41.53
	2D-TAN	44.51	26.54	77.13	61.96
	TLL	44.24	27.01	75.22	60.23
video-only	PFGA	21.69	12.56	-	-
	SCDM	23.84	12.93	51.66	32.36
	2D-TAN	27.56	13.93	61.65	36.78
	TLL	28.10	13.96	59.07	36.25

Table 2: Cross-scenario performance of video-only model on the AcNet2Charades and AcNet2DiDeMo.

Dataset	Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
Charades	random	11.88	3.76	46.64	16.88
	video-only	6.68	0.41	56.68	25.55
DiDemo	random	8.36	2.59	37.53	11.79
	video-only	4.87	1.22	38.74	16.54



(a) Frequency of actions. (b) Top frequency of actions.
(c) Frequency of intervals. (d) Top frequency of intervals.

Figure 4: Video moment bias analysis on ActivityNet Captions

Table 3: Performance evaluation results on the AcNet2Charades.

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
random	11.88	3.76	46.64	16.88
video-only	6.68	0.41	56.68	25.55
PFGA	5.75	1.53	-	-
SCDM	15.91	6.19	54.04	30.39
2D-TAN	15.81	6.30	59.06	31.53
Debias-TLL	**21.45**	**10.38**	**62.34**	**32.90**

Table 4: Performance evaluation results on the AcNet2DiDeMo.

Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
random	8.36	2.59	37.53	11.79
video-only	4.87	1.22	38.74	16.54
PFGA	6.24	2.01	-	-
SCDM	10.88	4.34	43.30	18.40
2D-TAN	12.50	5.50	44.88	20.73
Debias-TLL	**13.11**	**7.70**	**44.98**	**21.32**

according to the video, irrespective of the sentence query.

4.6 Cross-Scenario Evaluation

To quantify the effect of all afore-mentioned task-specific biases, we evaluate under a cross-scenario setting for temporal grounding in video. The training and evaluation processes are conducted on two different distributions where video moment biases cannot be transferred from one data distribution to another. Specifically, we train the model on the ActivityNet Captions considering its large-scale data amount and the diversity of scenes and activities. Then we test the model on the Charades-STA and DiDeMo (dubbed as AcNet2Charades and AcNet2DiDeMo repectively). As shown in Table 2, under the cross-scenario setting, the "video-only" TLL model fails to perform well on the testing data and is even inferior than the random guess model on the metrics of R1.

4.7 Performance Comparison

The results of the proposed Debias-TLL and the baselines on ACNet2Charades and ACNet2DiDeMo are summarized in Table 3 and 4 respectively. Our algorithm outperforms all the competing methods with a clear margin. It is noticeable that the proposed technique surpasses the state-of-the-art performances by 8.64% and 4.08% points in terms of R1@0.5 and R1@0.7 metric, respectively. This verifies the effectiveness of the sample importance reweighing in cross-scenario temporal grounding. The prevailing solutions for temporal grounding can be grouped into two categories i.e. top-down and bottom-up approach. We note that the top-down method PFGA achieves much inferior results than the top-down methods SCDM and 2D-TAN, which suggests the superiority of top-down design compared to the bottom-up one under the cross-scenario setting. We suspect that this is because the bottom-up approach directly predicts each frame's probabilities as ground-truth interval boundary and more easy to overfit to the temporal intervals bias.

Table 5: Ablation study on AcNet2Charades and AcNet2DiDeMo.

Dataset	Method	R@1 IoU=0.5	R@1 IoU=0.7	R@5 IoU=0.5	R@5 IoU=0.7
Charades	TLL	14.76	6.12	60.41	31.89
	Debias-TLL	**21.45**	**10.38**	**62.34**	**32.90**
DiDeMo	TLL	10.25	5.13	44.73	19.49
	Debias-TLL	**13.11**	**7.70**	**44.98**	**21.32**

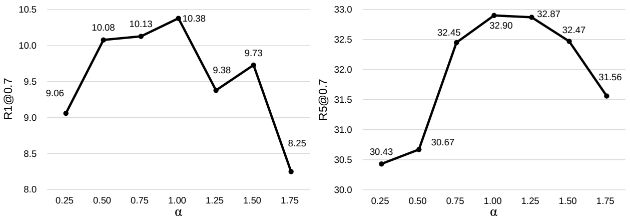


Figure 5: Impact of α on the AcNet2Charades.

4.8 Ablation Study

Impact of Importance Reweighing To study the impact of sample importance reweighing in Debias-TLL, we substitute the two-model based adjusted loss Eq. 7 with the commonly used binary cross entropy loss (marked as TLL), accordingly train the model TLL on ActivityNet Captions. The evaluation results on both Chrades-STA and DiDeMo are listed in Table 5. As expected, without the sample importance reweighing, the TLL model gets inferior results than Debias-TLL with a clear margin on both datasets, verifying the effectiveness of the proposed technique under the setting of cross-scenario temporal grounding.

Impact of Hyperparameter α The hyperparameter α plays a key role in controlling the magnitude of sample importance reweighing. And when $\alpha \to \infty$, $1 - s^{\alpha}$ approximates 1 and then the adjusted loss Eq. 7 approximates to binary cross entropy loss Eq. 5. An appropriate setting of α is required to rebalance the heavily video moment biased data. Here we study the effect of different settings of α. Figure 5 illustrates the performance R1@0.7 and R5@0.7 of the Debias-TLL model on AcNet2Charades with α set to 0.25 to 1.0. We found that the performances increase until $\alpha = 1.0$ and then decrease afterward. This shows that $\alpha = 1$ is a proper selection to balance the distribution of training samples and achieve satisfactory results. Note that even when $\alpha = 1.75$, the performance is still much superior to the TTL baseline without sample importance reweighing.

5 Conclusion

In this paper, we show that temporal grounding models are heavily affected by video moment bias of the data, limiting the generalization performance on cross-scenario testing data. To prevent the model from naively memorizing the biases and enforce it to ground the query sentence based on true cross-modal understanding, we propose a novel Debiased Temporal Language Localizer with a two-model based data-reweighing mechanism. Experiments show large-margin superiority of the proposed method in comparison with state-of-the-art competitors in cross-scenario temporal grounding.

References

[Anne Hendricks *et al.*, 2017] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, 2017.

[Cadene *et al.*, 2019] Remi Cadene, Corentin Dancette, Matthieu Cord, Devi Parikh, et al. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*, 2019.

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[Chen *et al.*, 2020] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. 2020.

[Gao *et al.*, 2017] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017.

[Ghosh *et al.*, 2019] Soham Ghosh, Anuva Agarwal, Zarana Parekh, and Alexander Hauptmann. Excl: Extractive clip localization using natural language descriptions. *arXiv preprint arXiv:1904.02755*, 2019.

[Hendricks *et al.*, 2018] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with temporal language. In *EMNLP*, 2018.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[Jeffrey Pennington and Manning, 2014] RichardSocher Jeffrey Pennington and ChristopherD Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Krishna *et al.*, 2017] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017.

[Lin *et al.*, 2019] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, 2019.

[Liu *et al.*, 2018a] Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Temporal modular networks for retrieving complex compositional activities in videos. In *ECCV*, 2018.

[Liu *et al.*, 2018b] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. Attentive moment retrieval in videos. In *SIGIR*, 2018.

[Liu *et al.*, 2018c] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. Cross-modal moment localization in videos. In *ACM MM*, 2018.

[Qi *et al.*, 2020] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *CVPR*, 2020.

[Ramakrishnan *et al.*, 2018] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *NeurIPS*, 2018.

[Rodriguez *et al.*, 2020] Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, Hongdong Li, and Stephen Gould. Proposal-free temporal moment localization of a natural-language query in video using guided attention. In *WACV*, 2020.

[Stroud *et al.*, 2019] Jonathan C Stroud, Ryan McCaffrey, Rada Mihalcea, Jia Deng, and Olga Russakovsky. Compositional temporal visual grounding of natural language event descriptions. *arXiv preprint arXiv:1912.02256*, 2019.

[Tang *et al.*, 2020] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020.

[Wang *et al.*, 2019a] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, 2019.

[Wang *et al.*, 2019b] Weining Wang, Yan Huang, and Liang Wang. Language-driven temporal activity localization: A semantic matching reinforcement learning model. In *CVPR*, 2019.

[Wang *et al.*, 2020] Jingwen Wang, Lin Ma, and Wenhao Jiang. Temporally grounding language queries in videos by contextual boundary-aware prediction. In *AAAI*, 2020.

[Yuan *et al.*, 2019] Yitian Yuan, Lin Ma, Jingwen Wang, Wei Liu, and Wenwu Zhu. Semantic conditioned dynamic modulation for temporal sentence grounding in videos. In *NeurIPS*, 2019.

[Zhang *et al.*, 2019a] Da Zhang, Xiyang Dai, Xin Wang, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *CVPR*, 2019.

[Zhang *et al.*, 2019b] Songyang Zhang, Jinsong Su, and Jiebo Luo. Exploiting temporal relationships in video moment localization with natural language. In *ACM MM*, 2019.

[Zhang *et al.*, 2019c] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. Cross-modal interaction networks for query-based moment retrieval in videos. In *SIGIR*, 2019.

[Zhang *et al.*, 2020] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *AAAI*, 2020.