iPerceive: Applying Common-Sense Reasoning to Multi-Modal Dense Video Captioning and Video Question Answering

Aman Chadha
Stanford University / Apple Inc.
amanc@stanford.edu

Abstract

Most of the previous works in visual understanding rely solely on understanding the "what" (e.g., object recognition) and "where" (e.g., event localization), which in some cases, fails to describe correct contextual relationships between events or leads to incorrect underlying visual attention. Part of what defines us as human and fundamentally different from machines is our instinct to seek causality behind any association, say an event Y that happened as a direct result of event X. To this end, we propose iPerceive, a framework capable of understanding the "why" between events in a video by building a common-sense knowledge base using contextual cues. We demonstrate the effectiveness of our technique using the dense video captioning (DVC) and video question answering (VideoQA) tasks. Furthermore, while most prior art in DVC and VideoQA relies solely on visual information, other modalities such as audio and speech are vital for a human observer's perception of an environment. We formulate DVC and VideoQA tasks as machine translation problems that utilize multiple modalities. By evaluating the performance of iPerceive DVC and iPerceive VideoQA on the ActivityNet Captions and TVQA datasets respectively, we show that our approach furthers the state-of-the-art.

1. Introduction

Today's computer vision systems are good at telling us the "what" (e.g., classification [20], segmentation [10]) and "where" (e.g., detection [30], localization [49], tracking [59]). Common-sense reasoning [41], which leads to the interesting question of "why", is a thinking gap in today's pattern-learning-based systems which rely on the likelihood of observing object Y given object X, $P(Y|X)$.

Failing to factor in causality leads to the unfortunate conclusion that the co-existence of objects X and Y might be attributed to spurious observational bias [12, 35]. For e.g., if a keyboard and mouse are often observed on a table, the model learns to develop an "association" between the two. The underlying common-sense that the keyboard and mouse are parts of a computer would not be inferred, and in fact the duo would be wrongly associated as being part of a table.

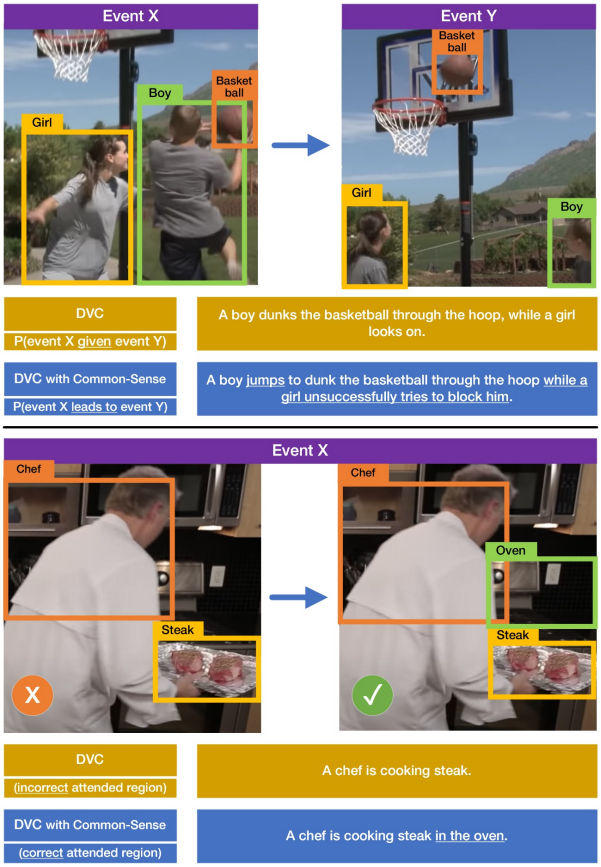

Figure 1. Top: An example of a cognitive error in DVC. While the girl tries to block the boy's dunking attempt, him *jumping* (event X) eventually *leads* to him dunking the basketball through the hoop (event Y). Bottom: An example of an incorrect attended region where conventional DVC approaches correlate a chef and steak to the activity of cooking *without even attending* to the nearby oven. We used [18] as our DVC baseline as it is the current state-of-the-art.

In the event that a keyboard and mouse are observed outside of a tabular setting, the model can commit a cognitive error. To address these limitations, we propose iPerceive, a framework that is able to build a common-sense knowledge base from one of the most common ways humans consume information, i.e., videos.

Video content has become an important source for humans to acquire information and knowledge. Video understanding is particularly important given the prevalence of visual information. Dense video captioning (DVC) [26] aims to temporally localize events from an untrimmed video and describe them using natural language. On the other hand, video question answering (VideoQA) is another challenging task in computer vision which requires enough expressive power from the model to distill visual events and their relations using linguistic concepts.

The vast majority of research in the field of DVC [26, 60, 54] and VideoQA [58] generates captions purely based on visual information. However, given the fact that auditory feedback is an essential aspect of human communication, unsurprisingly, almost all videos include an audio track and some also include a speech track, both of which could provide vital cues for understanding the context of the event. Inspired by the implementation in [18], our DVC model consumes the video, audio and speech modality for the caption generation process. Similarly, our VideoQA implementation utilizes video and text (in the form of dense captions, subtitles and QA).

The task of DVC can be decomposed into two parts: event detection and event description. Existing methods tackle this using a module for each of these sub-tasks, and either train the two modules independently [18] or in alternation [26, 54, 5]. This restricts model "wiggle", i.e., since events in a video sequence and the generated language are closely related, the language information should ideally be able to help localize events in the video. We address this by performing end-to-end training of our DVC model. While [18] presents a detailed study of the merits of using multiple modalities for DVC, they do not implement an end-to-end trainable system and train the captioning module on ground-truth event proposals. To this end, we utilize an end-to-end trainable model similar to [60] – this enforces consistency between the content in the proposed video segment and the semantic information in the language description. We thus blend multi-modal DVC with end-to-end learning.

We present iPerceive, a framework that generates common-sense features based on visual input. We offer hands-on evaluation of iPerceive using the tasks of DVC and VideoQA as case-studies. iPerceive DVC is a system that utilizes common-sense features and offers an end-to-end trainable multi-modal architecture that enables coherent dense video captions. Next, we propose an enhanced multi-modal architecture called iPerceive VideoQA that utilizes common-sense feature generation using iPerceive and dense captions using iPerceive DVC as its building blocks.

Our key contributions are centered around common-sense reasoning for videos, which we envision as a step towards human-level causal learning. [55] tackles the issue of observational bias in the context of images using common-sense generation, but applying a similar set of ideas to videos comes with its own set of challenges distinct from the image case. One observation is that events in videos can range across multiple time scales and can even overlap. Also, events can have causal relationships between themselves $(X \Rightarrow Y)$ that humans subconsciously perceive without any visible acknowledgment/feedback. Humans naturally learn common sense in an unsupervised fashion by exploring the physical world, and until machines imitate this learning path, there will be a "gap" between man and machine. This requires us to build a knowledge base and acquire contextual information from temporal events in a video sequence to determine inherent causal relationships. These "context-aware" features can improve both the accuracy of contextual relationships as well as steer attention to the correct entities. Furthermore, videos are generally challenging to process compared to images owing to the sheer size of data they contain.

We modify the approach proposed in [55] to suit the aforementioned nuances specific to videos (cf Section 3.3 for details) and apply our proposed technique to the tasks of DVC and VideoQA. Furthermore, since our common-sense features can be generated in a self-supervised manner, they can be easily adapted for other high-level vision tasks such as scene understanding [16], panoptic segmentation [24], etc.

2. Related Work

2.1. Visual Common-Sense

Current research in the field of building a common-sense knowledge base mainly falls into two categories: (i) learning from images [57, 52, 62] and (ii) learning actions from videos [9]. While the former limits learning to human-annotated knowledge which restricts its effectiveness and outreach, the latter is essentially learning from correlation.

2.2. Causality in Vision

There has been a recent surge of interest in coupling the complementary strengths of computer vision and causal reasoning [41, 40]. The union of these fields has been explored in several contexts, including image classification [32, 4], reinforcement learning [37], adversarial learning [25], visual dialog [42], image captioning [61] and scene/knowledge graph generation [48, 38]. While these methods offer limited task-specific causal inference, [55] offers a generic feature extractor for images.

2.3. Video Captioning

With the success of neural models in translation systems [47], similar methods became widely popular in video captioning [56, 53]. The core rationale behind this approach is to train two Recurrent Neural Networks (RNNs) in an encoder-decoder fashion. Specifically, an encoder inputs a set of video features and accumulates its hidden state, which is passed on to a decoder for captioning.

2.4. Dense Video Captioning

A significant milestone in the domain of video understanding was reached when Krishna et al. [26], inspired by the idea of the dense image captioning task [19], introduced the problem of DVC. They also released a new dataset called ActivityNet Captions which propelled research in the field [60, 54, 31]. Further, [54] adopted the idea of the context-awareness [26] and generalized the temporal event proposal module to utilize both past and future contexts as well as an attentive fusion to differentiate captions from highly overlapping events. Given the success of Transformers [51] in the machine translation task, it was inevitable for them to enter the similarly-complex task of video understanding. Zhou et al. [60] adopted the Transformer for DVC to alleviate the limitations of RNNs when modeling long-term dependencies in videos.

2.5. Multi-Modal Dense Video Captioning

Several attempts have been made to incorporate additional cues like audio and speech [43, 14] for dense video captioning task. Rahman et al. [43] utilized the idea of cycle-consistency to build a model with visual and audio in- puts. However, due to weak supervision, the system did not yield high performance. Hessel et al. [14] and Shi et al. [45] employ a Transformer architecture to encode both video frames and speech segments to generate captions for instructional cooking videos. While they achieve stellar results, their model fails to generalize to real-world videos where speech and captions can have a "gap" with the visual inputs whereas in instructional videos, speech and the captions are usually well-aligned with the visual content [36]. [18] tackles the problem of multi-modal DVC using a Transformer-based architecture and renders great results, however, does not incorporate the concept of end-to-end training introduced in [60].

2.6. Visual/Video Question Answering

The allied tasks of Visual Question Answering (VQA) and VideoQA involve the important ability of understanding visual information conditioned on language. While QA based on a single image has been well explored [17, 2, 46, 1, 33, 34], the field of VideoQA has received relatively little attention [58, 21, 7, 29, 27] owing to the inherent complexity involved in the task. Answering questions using videos requires an understanding of temporal information as well as spatial information, so it is more challenging than single image question answering.

Within VideoQA, [58] have explored using non-dense image captions. However, there exists limited research that utilizes dense captions to help improve the temporal localization of videos to extract object-level details of image frames. [21] explore this using dense image captioning and we take it a step further by utilizing DVC with common-sense features.

3. iPerceive DVC: Proposed Framework

3.1. Top Level View

Fig. 3 outlines the goals of iPerceive DVC: (i) temporally localize a set of events in a video, (ii) build a knowledge base for common-sense reasoning and, (iii) produce a textual description using audio, visual and speech cues for each event. To this end, we apply a three-stage approach. Since we limit our discussion of implementational details in this study to the blocks that we adapt to suit the common-sense reasoning aspect of iPerceive DVC, we refer the curious reader to [18], which we use as our baseline for the DVC task, for details of the building blocks of our architecture.

3.2. Event Proposal Module

We localize the temporal events in the video using the bidirectional single-stream (Bi-SST) network proposed in [54]. Bi-SST applies 3D convolutions (C3D) [50] to video frames and passes on the extracted features to a Bi-directional LSTM [15] network. The LSTM accumulates visual cues from past and future context over time and predicts the endpoints of each event in the video along with its confidence scores. The output of the LSTM feeds the common-sense reasoning module to convey the set of events in the input video to extract common-sense features for.

3.3. Common-Sense Reasoning Module

Common-sense reasoning enables self-supervised feature representation learning to serve as an improved visual region encoder for subsequent processing. In the context of developing a common-sense knowledge base, one of the biggest challenges involved is determining causal context: how do you figure out the cause and effect relationship between objects and thus re-align context for the model? As such, the training objective of the common-sense reasoning module is the proxy task of predicting the contextual objects of an event. Common-sense reasoning is based on causality which relies on $P(Y|do(X))$ [41]. This is fundamentally different from what prior work in the domain of DVC or VideoQA, or video understanding in general uses – the conventional likelihood $P(Y|X)$.

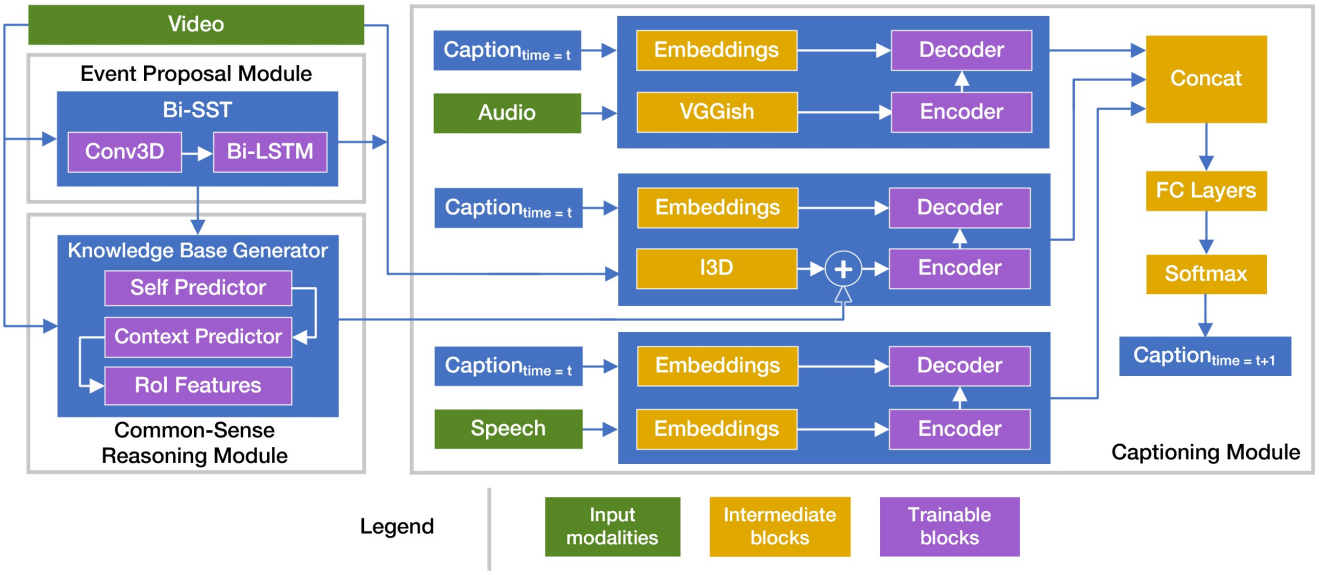

Figure 2. Architectural overview of iPerceive DVC. iPerceive DVC generates common-sense vectors from the temporal events that the proposal module localizes (left). Features from all modalities are sent to the corresponding encoder-decoder Transformers (middle). Upon fusing the processed features we finally output the next word in the caption using the distribution over the vocabulary (right).

Building upon the approach in [55], we carry out the following deliberate "borrow-put" experiment for a given frame with an event identified using the proposal module: (1) "borrow" non-local context, say an object Z from another event, (2) "put" Z in the context of object X and object Y, and then (3) test if object X still causes the existence of object Y given Z. This experiment helps determine if the chance of Z is independent on X or Y. Therefore, by using $P(Y|do(X))$ as the learning objective instead of $P(Y|X)$, the observational bias from the "apparent" context could be alleviated.

Our visual world contains several confounding agents $z \in Z$ that add spurious observational bias around objects X and Y and hinder common-sense development. This limits the model's learning using the traditional likelihood $P(Y|X)$, which can be formally defined [55] using Bayes' rule as:

$$P(Y|X) = \sum_{z} P(Y|X,z)P(z|X) \tag{1}$$

where the confounder z introduces the observational bias via $P(z|X)$.

Since we can hardly identify all confounders in the real world, we approximate the confounder set Z to a fixed confounder dictionary in the shape of a $N \times d$ matrix for practical use, where N is the category size in the dataset (e.g., 80 in MS-COCO) and d is the feature dimension of each RoI. Each entry $z \in Z$ is the averaged RoI feature, obtained using Faster R-CNN, of the i^{th} category samples in dataset.

We similarly define the "do" operation by disrupting the causal link between z and X (and thus de-biasing X) as,

$$P(Y|do\,(X)) = \sum_{z} P(Y|X,z)P(z) \tag{2}$$

Each RoI is then fed into two sibling branches: a self-predictor to predict the class of the "center" object $x \in X$, and a context-predictor to predict the "center" object's context labels, $y \in Y$, using "do" calculus. Since the self-predictor outputs a probability distribution p over N categories. On the other hand, the context-predictor outputs a probability distribution for a pair of RoIs, a center object X and Y is one of the K context objects. The last layer of the network thus performs label prediction using the Softmax layer:

$$P(Y|do\,(X)) = \mathbb{E}_z(Softmax(f(x,z))) \tag{3}$$

where, $f(\cdot)$ calculates the logits for N categories and $\mathbb{E}_z$ requires expensive sampling of z over the set of confounder objects Z. We utilize the Normalized Weighted Geometric Mean (NWGM) to approximate the above expectation. A detailed discourse for NWGM has been provided in [55].

Note that since the architecture proposed in [55] essentially serves as an improved visual region encoder given a region of interest (RoI) in an image, it assumes that an RoI exists and is available at test time. This greatly limits its usability to models that extract RoIs as part of their flow, and thus reduces its effectiveness with new datasets that the model has never seen before. We extend their work by utilizing a pre-trained Mask R-CNN [11] model to generate RoIs for frames within each event that has been localized by the event proposal module.

To make the task of common-sense feature generation for videos tractable, we only generate common-sense features for a frame when we detect a change in the environmental "setting" going from one frame to the next in a particular localized event. Specifically, we check for changes in the set of object labels in a scene and only generate common-sense features if a change is detected; if not, we re-use the common-sense features from the last frame.

These common-sense features derived for each localized event are then stacked with the corresponding visual features and sent down-stream for further processing.

3.4. Captioning Module

Given an event proposal and its common-sense vectors, the captioning module generates a caption using audio, visual and speech modalities. We formulate the captioning task as a machine translation problem and adapt the Transformer-based encoder-decoder architecture from [18]. We use inflated 3D convolutions (I3D) [3] to process visual modalities and the VGGish network [13] for audio modalities. We deploy an automatic speech recognition (ASR) system [8] to extract temporally-aligned speech transcriptions. These are juxtaposed alongside the video frames and the corresponding audio track, and fed in as input to our model. Features from each modality are fed to individual Transformer models along with the words of a caption from the previous time steps. The output of each Transformer is fused and a probability distribution is estimated over the vocabulary. After sampling the next word, the process is repeated until a special end token is obtained.

A note-worthy point is that for models that contain a self-attention architecture such as the one proposed in [18], we discard blocks that carry out self-attention as the self-attentive operation inherently has an indiscriminate correlation against our new learning objective based on the do-expression [55]. Put differently, self-attention implicitly applies conventional likelihood $P(Y|X)$ which contradicts causal reasoning $P(Y|do(X))$. Furthermore, given that the computation of self-attention is expensive, especially in the case of multiple heads, early concatenation of common-sense features significantly slows down training. Thus, we omit the self-multi-headed attention component in the encoder of [18].

3.5. Loss Functions

iPerceive DVC uses a four-fold training loss: (i) cross entropy MCE as proposal loss L_p to balance positive and negative proposals, (ii) multi-task common-sense reasoning loss L_{cs}, (iii) binary cross entropy BCE as mask prediction loss L_m and, (iv) cross entropy MCE across all words in every sentence as captioning loss L_c.

3.5.1 Proposal Loss

$$L_p = MCE(c, t, X, y) \tag{4}$$

where, c is the prediction score at time t, X is the input video and y is the ground truth label with an acceptable intersection over union (IoU).

3.5.2 Common-Sense Reasoning Loss

For a "center" object $x \in X$ in the video frame at time t, the self-predictor loss L_{self} can be defined using negative log likelihood as,

$$L_{self}(p, x^c, t) = -log(p[x^c]) \tag{5}$$

where, p is the probability distribution output of the self-predictor over N categories for X; x^c is the ground-truth class of RoI X.

Similarly, for a "context" object $y_i \in Y$ in the video frame at time t, the context-predictor loss L_{cxt} can be defined for a pair of RoI feature vectors x and y_i using negative log likelihood as,

$$L_{cxt}(p_i, y_i^c, t) = -log(p_i[y_i^c]) \tag{6}$$

where, y_i^c is the ground-truth label for y_i; p_i is calculated using $p_i = P(Y_i|do(X))$ in Eq. 2 and $p_i = (p_i[1], ..., p_i[N])$ is the probability over N categories.

The overall multi-task loss for each RoI X is,

$$L_{cs} = L_{self} + \frac{1}{K}\sum_i L_{cxt} \tag{7}$$

3.5.3 Mask Prediction Loss

$$L_m = BCE(Bin(S_a, E_a, t), f_M(S_a, E_p, S_a, E_a, t)) \tag{8}$$

where, $Bin(.)$ is 1 if $t \in [S_a, E_a]$ and 0 otherwise; f_M is a differentiable mask for time t; S_p and E_p are the start and end times of the event; S_a and E_a are the start and end positions of the anchor.

3.5.4 Captioning Loss

$$L_c = MCE(w'_t, w_t) \tag{9}$$

where w'_t is the ground truth word at time t.

3.5.5 Overall Loss Formulation

The final loss L is a combination of the individual losses associated with common-sense reasoning, event proposal localization and captioning:

$$L = \lambda_1 L_p + \lambda_2 L_{cs} + \lambda_3 L_m + \lambda_4 L_c \tag{10}$$

where λ_{1-4} weigh the individual loss components.

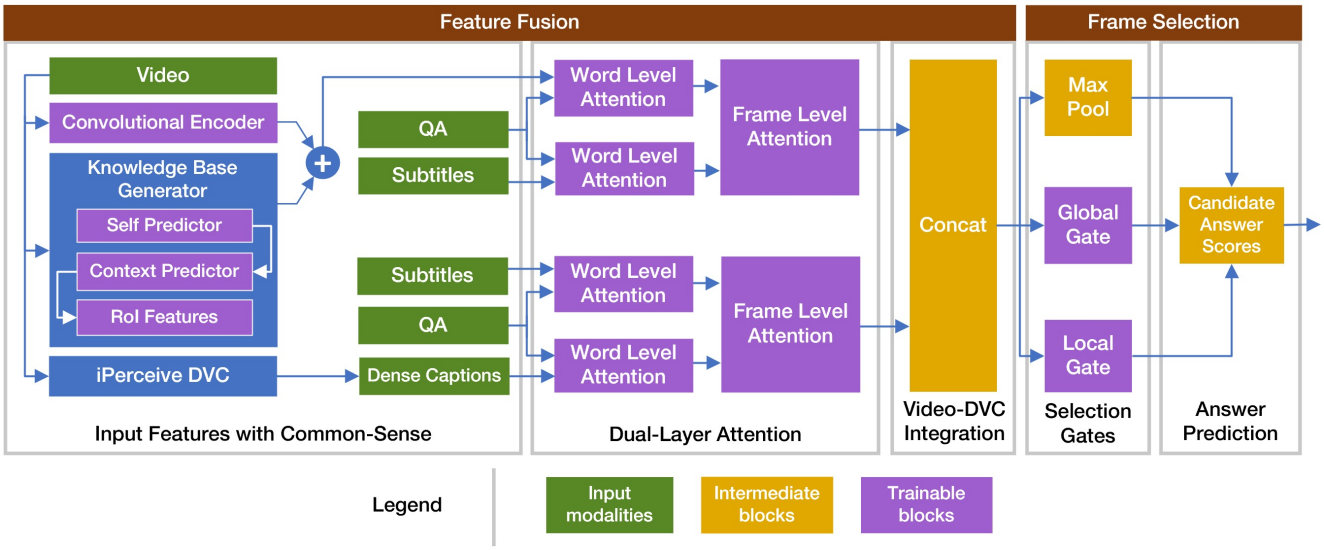

Figure 3. Architectural overview of iPerceive VideoQA. Our model consists of two main components: feature fusion and frame selection. For feature fusion, we encode features using a convolutional encoder, generate common-sense vectors from the input video sequence, and use iPerceive DVC for dense captions (left). Features from all modalities (video, dense captions, QA and subtitles) are then fed to dual-layer attention: word/object and frame-level (middle). Upon fusing the attended features, we calculate frame-relevance scores (right).

4. iPerceive VideoQA: Proposed Framework

4.1. Top Level View

Building upon the architecture proposed by [21], we propose iPerceive VideoQA, a model that uses common-sense knowledge to perform VideoQA. We utilize dense captions using iPerceive DVC to offer the model additional telemetry to correlate objects identified from video frames and their salient actions expressed through dense captions. Since we limit our discussion of implementational details in this study to the blocks that we adapt to suit the common-sense reasoning aspect of iPerceive VideoQA, we refer the curious reader to [21], which we use as our baseline for the VideoQA task, for details of the building blocks of our architecture.

Fig. 3 outlines the goals of iPerceive VideoQA: (i) build a common-sense knowledge base, (ii) extract features from multiple modalities: video and text (in the form of dense captions, subtitles and QA) and, (iii) implement the relevant-frames selection problem as a multi-label classification task. As such, we apply a two-stage approach.

4.2. Feature Fusion

4.2.1 Feature Generation Module

Leveraging the approach in [21], we extract features from multiple modalities viz. video, dense captions, question-answer (QA) pairs and subtitles. To generate dense captions, we utilize iPerceive DVC and operate it at a frame-level to derive dense captions for the current frame. We create five hypotheses by concatenating a question feature with each of five answer features, and we pair each visual frame feature with temporally neighboring subtitles. We encode all the features using a convolutional encoder.

4.2.2 Common-Sense Reasoning Module

We utilize the common-sense generation module proposed earlier to generate common-sense vectors corresponding to each frame of the input video. iPerceive VideoQA builds a common-sense knowledge base, concatenates common-sense features with the features extracted from the convolutional encoder and sends the output downstream.

4.2.3 Dual-Layer Attention

Word/Object-Level Attention: For each frame, the visual features are combined with the textual features namely, the QA features and sub-title features, using word/object-level attention following the approach in [21]. Separately, we also combine DVC features with QA features and sub-title features in a similar manner. To this end, we calculate similarity matrices [44] from (i) QA/subtitle and QA/visual features and, (ii) QA/subtitle and QA/DVC features, respectively. Attended subtitle features are obtained from the similarity matrices.

Frame-Level Attention: The fused features from word/object-level attention are integrated frame-wise via frame-level attention. Similar to the idea behind word/object-level attention, a similarity matrix is calculated from which attended frame-level features are calculated.

4.2.4 Video-DVC Integration Module

[21] implements self-cross attention to amalgamate information from the dual-layer attended visual and dense caption features, both of which have been fused with QA and subtitles. As discussed in Section 3.3, due to the challenges associated with generating common-sense features using a model that implements self-attention, we carry out concatenation of video and dense caption features.

4.3. Frame Selection

4.3.1 Selection Gates

Similar to the approach in [21], we utilize gates to selectively control the flow of information and ensure relevant information propagates through to the classifier. As such, we use a fully-connected layer to get frame-relevance scores that indicate how appropriate each frame is for answering a particular question. Using the logits for the five candidate answers, we choose the highest value as our prediction.

4.4. Loss Functions

iPerceive VideoQA uses a four-fold training loss: (i) multi-task common-sense reasoning loss L_{cs}, (ii) softmax cross-entropy loss as answer selection loss L_{ans}, (iii) balanced binary cross entropy as frame-selection loss L_{fs} and, (iv) in-and-out frame score margin L_{io}.

4.4.1 Common-Sense Reasoning Loss

We adopt a similar multi-task common-sense reasoning loss L_{cs} as in Section 3.5.2.

4.4.2 Answer Selection Loss

We use softmax cross-entropy loss to select the correct answer from five candidates.

$$L_{ans} = -\log\left(\frac{e^{s_{gt}}}{\sum_k e^{s_k}}\right) \tag{11}$$

where, s_{gt} is the logit of ground-truth answer.

4.4.3 Frame-Selection Supervision Loss

We consider frame selection as a multi-label classification task. We label each frame with a score corresponding to whether it lies within the time span needed to offer the correct answer. Since negative examples would dominate a typical training setting, to account for the imbalance between negative and positive examples, we utilize Balanced Binary Cross-Entropy (BBCE) [21].

$$L_{fs} = -\left(\sum_i^{N_{in}} \frac{log(s_i^{in})}{N_{in}} + \sum_j^{N_{out}} \frac{log(1-s_j^{out})}{N_{out}}\right) \tag{12}$$

where, s_i^{in} and s_i^{out} are i^{th} in-frame score and j^{th} out-frame score respectively; N_{in} and N_{out} are the number of in-frames and out-frames respectively.

4.4.4 In-and-Out Frame Score Margin

Drawing upon the novel loss function proposed in [21], we implement in-and-out frame score margin loss L_{io} as,

$$L_{io} = 1 + avg(OFS) - avg(IFS) \tag{13}$$

where, OFS (Out Frame Score) and IFS (In Frame Score) are the average scores for frames whose labels are '0' and '1' respectively.

4.4.5 Overall Loss Formulation

The final loss L is a combination of the individual losses,

$$L = \lambda_1 L_{cs} + \lambda_2 L_{ans} + \lambda_3 L_{fs} + \lambda_4 L_{io} \tag{14}$$

where λ_{1-4} weigh the individual loss components.

5. Experiments

6. iPerceive DVC

6.0.1 Dataset

We train and assess iPerceive DVC using the ActivityNet Captions [26] dataset, using a train/val/test split of 0.5:0.25:0.25, ActivityNet Captions contains 20k videos from YouTube, and on an average each video has 3.65 events which are each 2 minutes long and are annotated by two different annotators using 13.65 words each. We report all results using the validation set (since no ground truth is available for the test set).

6.0.2 Metrics

We evaluate the performance of iPerceive DVC using BLEU@N [39] and METEOR [6]. We use the official evaluation script provided in [26].

6.0.3 Comparison with Baseline Methods

Tab. 1 shows a comparison of iPerceive DVC with the state-of-the-art. Algorithms were split into the ones which "saw" all training videos and others which trained on partially available data (since some YouTube videos which were part of the ActivityNet Captions dataset are no longer available). Fig. 4 shows a qualitative comparison between different models. We compared [60] and [55] because they are the best performing baselines for captioning and event localization respectively.

Figure 4. Qualitative sampling of iPerceive DVC. Captioning results for a sample video from the ActivityNet Captions validation set show better performance owing to common-sense reasoning and end-to-end training.

Table 1. Evaluation of iPerceive DVC on the ActivityNet Captions validation set using BLEU@N (B@N) and METEOR (M).

Method	GT Proposals			Learned Proposals		
	B@3	B@4	M	B@3	B@4	M
Seen full dataset						
Krishna et al. [26]	4.09	1.60	8.88	1.90	0.71	5.69
Wang et al. [54]	-	-	10.89	2.55	**1.31**	5.86
Zhou et al. [60]	5.76	2.71	11.16	2.42	1.15	4.98
Li et al. [31]	4.55	1.62	10.33	2.27	0.73	6.93
Seen part of the dataset						
Rahman et al. [43]	3.04	1.46	7.23	1.85	0.90	4.93
Iashin et al. [18]	4.12	1.81	10.09	2.31	0.92	6.80
iPerceive	5.23	2.34	11.77	2.59	1.07	7.29
Iashin et al. (all modalities)	5.83	2.86	11.72	2.60	1.07	7.31
iPerceive (all modalities)	**6.13**	**2.98**	**12.27**	**2.93**	1.29	**7.87**

6.0.4 Ablation Analysis

Tab. 2 shows ablation studies for iPerceive DVC to assess the impact of common-sense reasoning and end-to-end training as design decisions.

Table 2. Ablation analysis for iPerceive DVC.

Common-Sense Reasoning	End-to-End Training	METEOR
✗	✗	7.31
✗	✓	7.42
✓	✗	7.51
✓	✓	7.87

7. iPerceive VideoQA

7.0.1 Dataset

We train and evaluate iPerceive VideoQA using the TVQA dataset [28] which consists of video frames, subtitles and QA pairs from six TV shows. The train/val/test splits for TVQA are 0.84/0.12/0.06. Each example has five candidate answers with one of them the ground-truth. TVQA is thus a classification task which can be evaluated based on the accuracy metric.

7.0.2 Comparison with Baseline Methods

Tab. 3 shows a comparison of iPerceive VideoQA with the state-of-the-art.

Table 3. Evaluation of iPerceive VideoQA on the TVQA dataset.

Method	Test-Public (%)							Val (%)
	Mean	BBT	Friends	HIMYM	Grey	House	Castle	
Lei et al. [28]	66.46	70.25	65.78	64.02	67.20	66.84	63.96	65.85
Kim et al. [23]	66.77	-	-	-	-	-	-	
Lei et al. [22]	67.05	-	-	-	-	-	-	
Kim et al. [21]	74.09	74.04	73.03	74.34	73.44	74.68	74.86	74.20
iPerceive VideoQA	**75.15**	**75.32**	**74.22**	**75.14**	**74.42**	**75.22**	**75.77**	**76.97**

7.0.3 Ablation Analysis

Tab. 2 shows ablation studies for iPerceive VideoQA to assess the impact of common-sense reasoning and iPerceive DVC as design decisions.

Table 4. Ablation analysis for iPerceive VideoQA.

Common-Sense Reasoning	iPerceive DVC	Val (%)
✗	✗	74.20
✗	✓	75.42
✓	✗	75.55
✓	✓	76.97

8. Conclusion

We proposed iPerceive, a portable framework that enables common-sense learning for videos. We demonstrated the effectiveness of iPerceive on the tasks of DVC and VideoQA using the ActivityNet Captions and TVQA datasets respectively. Furthermore, iPerceive DVC blends multi-modal DVC with end-to-end Transformer-based learning. iPerceive VideoQA leverages iPerceive DVC to offer state-of-the-art performance. Using ablation studies, we showed that these common-sense features help the model better perceive relationships between events in videos, leading to improved performance on challenging video tasks that need cognition.

9. Broader Impact

Given that humans perceive their immediate world by understanding their surroundings, we feel that video understanding in general is important to close the "gap" between man and machine. Our work propels the idea of causal reasoning for machines and bring us one step closer to the ultimate goal of visual-linguistic causal reasoning which is one of the distinct qualities that make us human. Since our work is easily portable, we hope that the promising results in our work would encourage researchers to further explore the domain of common-sense reasoning and apply it to new applications in the field of video and language understanding.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 3

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2017. 5

[4] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*, 2014. 2

[5] Shizhe Chen, Yuqing Song, Yida Zhao, Jiarong Qiu, Qin Jin, and Alexander Hauptmann. Ruc+ cmu: System report for dense captioning events in videos. *arXiv preprint arXiv:1806.08854*, 2018. 2

[6] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 7

[7] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1999–2007, 2019. 3

[8] Google. YouTube Data API Video Captions. https://developers.google.com/youtube/v3/docs/captions, 2008. [Accessed May 21 2020]. 5

[9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, volume 1, page 5, 2017. 2

[10] Matthias Grundmann, Vivek Kwatra, Mei Han, and Irfan Essa. Efficient hierarchical graph-based video segmentation. In *2010 ieee computer society conference on computer vision and pattern recognition*, pages 2141–2148. IEEE, 2010. 1

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[12] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018. 1

[13] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017. 5

[14] Jack Hessel, Bo Pang, Zhenhai Zhu, and Radu Soricut. A case study on combining asr and visual features for generating instructional video captions. *arXiv preprint arXiv:1910.02930*, 2019. 3

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[16] Anthony Hu, Fergal Cotter, Nikhil Mohan, Corina Gurau, and Alex Kendall. Probabilistic future prediction for video scene understanding. *arXiv preprint arXiv:2003.06409*, 2020. 2

[17] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 804–813, 2017. 3

[18] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. *arXiv preprint arXiv:2003.07758*, 2020. 1, 2, 3, 5, 8

[19] Justin Johnson, Andrej Karpathy, and Fei-Fei Li. Densecap: fully convolutional localization networks for dense captioning. corr abs/1511.07571 (2015). *arXiv preprint arXiv:1511.07571*, 2015. 3

[20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 1

[21] Hyounghun Kim, Zineng Tang, and Mohit Bansal. Dense-caption matching and frame-selection gating for temporal localization in videoqa. *arXiv preprint arXiv:2005.06409*, 2020. 3, 6, 7, 8

[22] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Gaining extra supervision via multi-task learning for multi-modal video question answering. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 8

[23] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8337–8346, 2019. 8

[24] A Kirillov, K He, R Girshick, C Rother, and P Dollar. Panoptic segmentation. arxiv. *arXiv preprint arXiv:1801.00868*, 1(3):5, 2018. 2

[25] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017. 2

[26] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 2, 3, 7, 8

[27] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. *arXiv preprint arXiv:2002.10698*, 2020. 3

[28] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 8

[29] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019. 3

[30] Liyuan Li, Weimin Huang, Irene YH Gu, and Qi Tian. Foreground object detection from videos containing complex background. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 2–10, 2003. 1

[31] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7492–7500, 2018. 3, 8

[32] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6979–6987, 2017. 2

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 3

[34] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016. 3

[35] Varun Manjunatha, Nirat Saini, and Larry S Davis. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9562–9571, 2019. 1

[36] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2630–2640, 2019. 3

[37] Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751*, 2019. 2

[38] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. *arXiv preprint arXiv:2003.13942*, 2020. 2

[39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 7

[40] Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014. 2

[41] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 1, 2, 3

[42] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. *arXiv preprint arXiv:1911.10496*, 2019. 2

[43] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8908–8917, 2019. 3, 8

[44] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016. 6

[45] Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. Dense procedure captioning in narrated instructional videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6382–6391, 2019. 3

[46] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 3

[47] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 3

[48] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. *arXiv preprint arXiv:2002.11949*, 2020. 2

[49] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 1

[50] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 3

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[52] Ramakrishna Vedantam, Xiao Lin, Tanmay Batra, C Lawrence Zitnick, and Devi Parikh. Learning common sense through visual abstraction. In *Proceedings of the IEEE international conference on computer vision*, pages 2542–2550, 2015. 2

[53] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015. 3

[54] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7190–7198, 2018. 2, 3, 8

[55] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. *arXiv preprint arXiv:2002.12204*, 2020. 2, 4, 5, 7

[56] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015. 3

[57] Mark Yatskar, Vicente Ordonez, and Ali Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, 2016. 2

[58] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2, 3

[59] Da Zhang, Hamid Maei, Xin Wang, and Yuan-Fang Wang. Deep reinforcement learning for visual object tracking in videos. *arXiv preprint arXiv:1701.08936*, 2017. 1

[60] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748, 2018. 2, 3, 7, 8

[61] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. More grounded image captioning by distilling image-text matching model. *arXiv preprint arXiv:2004.00390*, 2020. 2

[62] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, pages 408–424. Springer, 2014. 2