Robust Change Captioning

Dong Huk Park, Trevor Darrell, Anna Rohrbach

University of California, Berkeley

Abstract

Describing what has changed in a scene can be useful to a user, but only if generated text focuses on what is semantically relevant. It is thus important to distinguish distractors (e.g. a viewpoint change) from relevant changes (e.g. an object has moved). We present a novel Dual Dynamic Attention Model (DUDA) to perform robust Change Captioning. Our model learns to distinguish distractors from semantic changes, localize the changes via Dual Attention over “before” and “after” images, and accurately describe them in natural language via Dynamic Speaker, by adaptively focusing on the necessary visual inputs (e.g. “before” or “after” image). To study the problem in depth, we collect a CLEVR-Change dataset, built off the CLEVR engine, with 5 types of scene changes. We benchmark a number of baselines on our dataset, and systematically study different change types and robustness to distractors. We show the superiority of our DUDA model in terms of both change captioning and localization. We also show that our approach is general, obtaining state-of-the-art results on the recent realistic Spot-the-Diff dataset which has no distractors.

1. Introduction

We live in a dynamic world where things change all the time. Change detection in images is a long-standing research problem, with applications in a variety of domains such as facility monitoring, medical imaging, and aerial photography [17, 46, 50]. A key challenge in change detection is to distinguish the relevant changes from the irrelevant ones [49] since the former are those that should likely trigger a notification. Existing systems aim to sense or localize a change, but typically do not convey detailed semantic content. This is an important limitation for a realistic application, where analysts would benefit from such knowledge, helping them to better understand and judge the significance of the change. Alerting a user on every detected difference likely will lead to a frustrated operator; moreover, it is desirable to have a change detection system that does not output a binary indicator of change/no-change, but instead outputs a concise description of what has changed, and where.

Expressing image content in natural language is an active area of Artificial Intelligence research, with numerous approaches to image captioning having been recently proposed [3, 12, 38, 59]. These methods have the benefit of conveying visual content to human users in a concise and natural way. They can be especially useful, when tailored to a specific task or objective, such as e.g. explaining the model’s predictions [20, 45] or generating non-ambiguous referring expressions for specific image regions [40, 61].

In this work we investigate robust Change Captioning, where an important scene change has to be identified and conveyed using language in the presence of distractors (where only an illumination or viewpoint change occurred). We aim to generate detailed and informative descriptions that refer to the changed objects in complex scenes (see Figure 1).

To distinguish an irrelevant distractor from an actual change (e.g. an object moved), one needs to “compare” the two images and find correspondences and disagreements. We propose a Dual Dynamic Attention Model (DUDA) that
learns to localize the changes via a specialized attention mechanism. It consists of two components: Dual Attention that predicts a separate spatial attention for each image in the “before”/“after” pair, and a Dynamic Speaker that generates a change description by semantically modulating focus among the visual features relayed from the Dual Attention. Both components are neural networks that are trained jointly with only caption-level supervision, i.e., no information about the change location is used during training.

In order to study Change Captioning in the presence of distractors, we build a CLEVR-Change Dataset. We rely on the image generation engine by [26], which allows us to produce complex compositional scenes. We create pairs of “before” and “after” images with: (a) only illumination/viewpoint change (distractors), and (b) illumination/viewpoint change combined with a scene change. We consider 5 scene change types (color/material change, adding/dropping/moving an object), and collect almost 80K image pairs. We augment the image pairs with automatically generated change captions based on templates (see Figure 3). Note that in the recently proposed Spot-the-Diff dataset [25], the task also is to generate change captions for a pair of images. However, their problem statement is different from ours in that: 1) they assume a change in each image pair while our goal is to be robust to distractors, 2) the images are aligned (no viewpoint shift), 3) change localization can not be evaluated as ground-truth is not available in [25].

We first evaluate our novel DUDA model on the CLEVR-Change dataset, and compare it to a number of baselines, including a naive pixel-difference captioning baseline. We show that our approach outperforms the baselines in terms of change caption correctness as well as change localization. The most challenging change types to describe are object movement and texture change, while movement is also the hardest to localize. We also show that our approach is general, applying it to the Spot-the-Diff dataset [25]. Given the same visual inputs as [25], our model matches or outperforms their approach.

2. Related Work

Here we discuss prior work on change detection, task-specific image captioning, and attention mechanism.

Change detection One popular domain for image-based change detection is aerial imagery [35, 53, 62], where changes can be linked to disaster response scenarios (e.g., damage detection) [17] or monitoring of land cover dynamics [29, 54]. Prior approaches often rely on unsupervised methods for change detection, e.g., image differencing, due to high cost of obtaining ground-truth annotations [9]. Notably, [17] propose a semi-supervised approach with human in the loop, relying on a hierarchical shape representation.

Another prominent domain is street scenes [1, 28]. Notably, [50] propose a Panoramic Change Detection Dataset, built on Google Street View panoramic images. In their follow-up work, [51] propose an approach to change detection which relies on dense optical flow to address the difference in viewpoints between the images. In a recent work, [43] rely on 3D models to identify scene changes by re-projecting images on one another. Another line of work targets change detection in video, e.g., using a popular CD-net benchmark [16, 58], where background subtraction is a successful strategy [8]. Instead of relying on costly pixel-level video annotation, [30] propose a weakly supervised approach, which estimates pixel-level labels with a CRF.

Other works address a more subtle, fine-grained change detection, where an object may change its appearance over time, e.g., for the purpose of a valuable object monitoring [14, 24]. To tackle this problem, [52] estimate a dense flow field between images to address viewpoint differences.

Our DUDA model relies on an attention mechanism rather than pixel-level difference or flow. Besides, our task is not only to detect the changes, but also to describe them in natural language, going beyond the discussed prior works.

Task-specific caption generation While most image captioning works focus on a generic task of obtaining image relevant descriptions [3, 12, 57], some recent works explore pragmatic or “task-specific” captions. Some focus on generating textual explanations for deep models’ predictions [19, 20, 45]. Others aim to generate a discriminative caption for an image or image region, to disambiguate it from a distractor [4, 10, 40, 39, 55, 61]. This is relevant to our work, as part of the change caption serves as a referring expression to put an object in context of the other objects. However, our primary focus is to correctly describe the scene changes.

The most related to ours is the work of [25], who also address the task of change captioning for a pair of images. While we aim to distinguish distractors from relevant changes, they assume there is always a change between the two images. Next, their pixel-difference based approach assumes that the images are aligned, while we tackle viewpoint change between images. Finally, we systematically study different change types in our new CLEVR-Change Dataset. We show that our approach generalizes to their Spot-the-Diff dataset in subsection 5.3.

Attention in image captioning Attention mechanism [6] over the visual features was first used for image captioning by [59]. Multiple works have since adopted and extended this approach [15, 36, 47], including performing attention over object detections [3]. Our DUDA model relies on two forms of attention: spatial Dual Attention used to localize changes between two images, and semantic attention, used by our Dynamic Speaker to adaptively focus on “before”, “after” or “difference” visual representations.
3. Dual Dynamic Attention Model (DUDA)

We propose a Dual Dynamic Attention Model (DUDA) for change detection and captioning. Given a pair of “before” and “after” images \( I_{bef} \) and \( I_{aft} \), respectively, our model first detects whether a scene change has happened, and if so, locates the change on both \( I_{bef} \) and \( I_{aft} \). The model then generates a sentence that not only correctly describes the change, but also is spatially and temporally grounded in the image pair. To this end, our model includes a Dual Attention (localization) component, followed by a Dynamic Speaker component to generate change descriptions. An overview of our model is shown in Figure 2.

We describe the implementation details of our Dual Attention in subsection 3.1, and our Dynamic Speaker in subsection 3.2. In subsection 3.3, we detail our training procedure for jointly optimizing both components using change captions as the only supervision.

3.1. Dual Attention

Our Dual Attention acts as a change localizer between \( I_{bef} \) and \( I_{aft} \). Formally, it is a function \( f_{loc}(X_{bef}, X_{aft}; \theta_{loc}) = (l_{bef}, l_{aft}) \) parameterized by \( \theta_{loc} \) that takes \( X_{bef} \) and \( X_{aft} \) as inputs, and outputs feature representations \( l_{bef} \) and \( l_{aft} \) that encode the change manifested in the input pairs. In our implementation, \( X_{bef}, X_{aft} \in \mathbb{R}^{C \times H \times W} \) are image features of \( I_{bef}, I_{aft} \), respectively, encoded by a pretrained ResNet [18].

We first subtract \( X_{bef} \) from \( X_{aft} \) in order to capture semantic difference in the representation space. The resulting tensor \( X_{diff} \) is concatenated with both \( X_{bef} \) and \( X_{aft} \) which are then used to generate two separate spatial attention maps \( a_{bef}, a_{aft} \in \mathbb{R}^{1 \times H \times W} \). Following [41], we utilize element-wise sigmoid instead of softmax for computing our attention maps to avoid introducing any form of global normalization. Finally, \( a_{bef} \) and \( a_{aft} \) are applied to the input features to do a weighted-sum pooling over the spatial dimensions:

\[
X_{diff} = X_{aft} - X_{bef} \tag{1}
\]
\[
X_{bef}' = [X_{bef}; X_{diff}], X_{aft}' = [X_{aft}; X_{diff}] \tag{2}
\]
\[
a_{bef} = \sigma(\text{conv}_2(\text{ReLU}(\text{conv}_1(X_{bef})))) \tag{3}
\]
\[
a_{aft} = \sigma(\text{conv}_2(\text{ReLU}(\text{conv}_1(X_{aft})))) \tag{4}
\]
\[
l_{bef} = \sum_{H,W} a_{bef} \odot X_{diff}, l_{bef} \in \mathbb{R}^{C} \tag{5}
\]
\[
l_{aft} = \sum_{H,W} a_{aft} \odot X_{diff}, l_{aft} \in \mathbb{R}^{C} \tag{6}
\]

where \([;]\), conv, \(\sigma\), and \(\odot\) indicate concatenation, convolutional layer, elementwise sigmoid, and elementwise multiplication, respectively. See Figure 2 for the visualization of Dual Attention component.

This particular architectural design allows the system to attend to images differently depending on the type of a change and the amount of a viewpoint shift, which is a capability crucial for our task. For instance, to correctly describe that an object has moved, the model needs to localize and match the moved object in both images; having single attention that locates the object only in one of the images is likely to cause confusion between e.g. moving vs. adding an object. Even if there is an attribute change (e.g. color) which does not involve object displacement, single attention might not be enough to correctly localize the changed object under a viewpoint shift. Unlike [60, 42, 37, 31, 45], DUDA utilizes Dual Attention to process multiple visual inputs separately and thereby addresses Change Captioning in the presence of distractors.

3.2. Dynamic Speaker

Our Dynamic Speaker is based on the following intuition: in order to successfully describe a change, the model should not only learn where to look in each image (spatial attention, predicted by the Dual Attention), but also when to look at each image (semantic attention, here). Ideally, we
would like the model to exhibit dynamic reasoning, where it learns when to focus on “before” ($l_{\text{bef}}$), “after” ($l_{\text{aft}}$), or “difference” feature ($l_{\text{diff}} = l_{\text{aft}} - l_{\text{bef}}$) as it generates a sequence of words. For example, it is necessary to look at the “after” feature ($l_{\text{aft}}$) when referring to a new object added to a scene. Figure 2 illustrates this behaviour.

To this end, our Dynamic Speaker predicts an attention $\alpha_i(t)$ over the visual features $l_i$’s at each time step $t$, and obtains the dynamically attended feature $l_{\text{dyn}}(t)$:

$$l_{\text{dyn}}(t) = \sum_i \alpha_i(t) l_i$$

where $i \in \{\text{bef}, \text{diff}, \text{aft}\}$. We use the attentional Recurrent Neural Network [5] to model this formulation.

Our Dynamic Speaker consists of two modules, namely the dynamic attention module and the caption module. Both are recurrent models based on LSTM [21]. At each time step $t$, the LSTM decoder in the dynamic attention module takes as input the previous hidden state of the caption module $h_{\text{c}}(t-1)$ and some latent projection $v$ of the visual features $l_{\text{bef}}, l_{\text{diff}},$ and $l_{\text{aft}}$ to predict attention weights $\alpha_i(t)$:

$$v = \text{ReLU}(W_{d1}[l_{\text{bef}}; l_{\text{diff}}; l_{\text{aft}}] + b_{d1})$$

$$u(t) = [v; h_{\text{d}}(t-1)]$$

$$h_{\text{d}}(t) = \text{LSTM}_d(u(t); h_{\text{d}}(0:t-1))$$

$$\alpha(t) \sim \text{Softmax}(W_{h1}h_{\text{d}}(t) + b_{d2})$$

where $h_{\text{d}}(t)$ and $h_{\text{c}}(t)$ are LSTM outputs at decoder time step $t$ for dynamic attention module and caption module, respectively, and $W_{d1}, b_{d1}, W_{d2},$ and $b_{d2}$ are learnable parameters. Using the attention weights predicted from Equation (11), the dynamically attended feature $l_{\text{dyn}}(t)$ is obtained according to Equation (7). Finally, $l_{\text{dyn}}(t)$ and the embedding of the previous word $w_{t-1}$ (ground-truth word during training, predicted word during inference) are input to the LSTM decoder of the caption module to begin generating distributions over the next word:

$$x(t-1) = E\mathbb{1}_{w_{t-1}}$$

$$c(t) = [x(t-1); l_{\text{dyn}}(t)]$$

$$h_{\text{c}}(t) = \text{LSTM}_c(h_{\text{c}}(t-1)\mid c(t); h_{\text{c}}(0:t-1))$$

$$w_t \sim \text{Softmax}(W_c h_{\text{c}}(t) + b_c)$$

where $\mathbb{1}_{w_{t-1}}$ is a one-hot encoding of the word $w_{t-1}$, $E$ is an embedding layer, and $W_c, b_c$ are learned parameters.

<table>
<thead>
<tr>
<th># Img Pairs</th>
<th># Captions</th>
<th># Bboxes</th>
</tr>
</thead>
<tbody>
<tr>
<td>39,803</td>
<td>199,015</td>
<td>-</td>
</tr>
<tr>
<td>7,958</td>
<td>58,850</td>
<td>15,916</td>
</tr>
<tr>
<td>7,963</td>
<td>58,946</td>
<td>15,926</td>
</tr>
<tr>
<td>7,966</td>
<td>59,198</td>
<td>7,966</td>
</tr>
<tr>
<td>7,961</td>
<td>58,843</td>
<td>7,961</td>
</tr>
<tr>
<td>7,955</td>
<td>58,883</td>
<td>15,910</td>
</tr>
<tr>
<td>79,606</td>
<td>493,735</td>
<td>64,679</td>
</tr>
</tbody>
</table>

Table 1: CLEVR-Change Dataset statistics: number of image pairs, captions, and bounding boxes for each change type: DISTRACTOR (DI), COLOR (C), TEXTURE (T), ADD (A), DROP (D), MOVE (M).

### 3.3. Joint Training

We jointly train the Dual Attention and the Dynamic Speaker end-to-end by maximizing the likelihood of the observed word sequence. Let $\theta$ denote all the parameters in DUDA. For a target ground-truth sequence $(w_1^t, \ldots, w_T^t)$, the objective is to minimize the cross entropy loss:

$$L_{XE}(\theta) = - \sum_{t=1}^{T} \log(p_\theta(w_t^t \mid w_1^t, \ldots, w_{t-1}^t))$$

Similar to [41], we apply $L_1$ regularization to the spatial attention masks generated by our Dual Attention in order to minimize unnecessary activations. We also use an entropy regularization over the attention weights generated by our Dynamic Speaker to encourage exploration in using visual features. The final loss function we optimize is as follows:

$$L(\theta) = L_{XE} + \lambda_{L1} L_1 - \lambda_{ent} L_{ent}$$

where $L_1$ and $L_{ent}$ are $L_1$ and entropy regularization, respectively, and $\lambda_{L1}$ and $\lambda_{ent}$ are hyperparameters. Note, that the Dual Attention component receives no direct supervision for change localization. The only available supervision is obtained through the Dynamic Speaker, which then directs the Dual Attention towards discovering the change.

### 4. CLEVR-Change Dataset

Given a lack of an appropriate dataset to study Change Captioning in the presence of distractors, we build the CLEVR-Change Dataset, based on the CLEVR engine [26]. We choose CLEVR, inspired by many works that use it to build diagnostic datasets for various vision and language tasks, e.g. visual question answering [26], referring expression comprehension [22, 34], text-to-image generation [13] or visual dialog [33]. As Change Captioning is an emerging task we believe our dataset can complement existing datasets, e.g. [25], which is small, always assumes the presence of a change and lacks localization ground-truth.

First, we generate random scenes with multiple objects in them, which serve as “before” images. Note, that in domains such as satellite imagery [35, 53, 62] or surveillance/street scenes [1, 28, 43], typical distractors include
changes in camera position/zoom or illumination. Motivated by these applications we approach distractor construction accordingly. For each “before” image we create two “after” images. In the first one, we change the camera position leading to a different angle, zoom, and/or illumination. We use a specific allowed range for the transformation parameters: for each \((x, y, z)\) camera location, we randomly sample a number from the range between \(-2.0\) and \(2.0\), and jitter the original coordinates by the sampled amount. In the second “after” image, we additionally introduce a scene change. We consider the following types of scene changes: (a) an object’s color is changed, (b) an object’s texture is changed, (c) a new object is added, (d) an existing object is dropped, (e) an existing object is moved. In the following we refer to these as: COLOR, TEXTURE, ADD, DROP, MOVE, and DISTRACTER for no scene change. In total, we generate 39,803 “before” images with respectively 79,606 “after” images. We make sure that the number of data points for each scene change type is balanced. The dataset is split into 67,600, 3,976, and 7,970 training/validation/test image pairs, respectively.

Based on the created “before” and “after” scenes, we further augment them with change captions. Each caption is automatically constructed with two parts: the referring part (e.g. “A large blue sphere to the left of a red object”) and the change part (e.g. “has appeared”). Note that for all the change types except ADD, the referring part is generated based on the “before” image, while for ADD, the “after” image is used. To get the change part, we rely on a set of change specific templates (see supplemental for details). However, note that the proposed DUDA model is not limited to templated language as we further demonstrate on a dataset with natural language descriptions.

Finally, we obtain spatial locations of where each scene change took place, so that we can evaluate the correctness of change localization. Specifically, we obtain bounding boxes for all the objects affected by a change, either in one image or in both (“before”/“after”), depending on the change type. The overall dataset statistics are shown in Table 1, and some examples of distractors vs. scene changes with their descriptions and bounding boxes are shown in Figure 3.

5. Experiments

In this section, we evaluate our DUDA model on the Change Captioning task against a number of baselines. First, we present quantitative results for the ablations and discuss their implications on our new CLEVR-Change Dataset. We also provide qualitative analyses of the generated captions, examine attention weights predicted by DUDA, and assess its robustness to viewpoint shift. Finally, we test the general effectiveness of our approach on the Spot-the-Diff [25], a realistic dataset with no distractors.

5.1. Experimental setup

Here, we detail our experimental setup in terms of implementation and evaluation schemes.

Implementation Details. Similar to [23, 27, 48], we use ResNet-101 [18] pretrained on ImageNet [11] to extract visual features from the images. We use features from the convolutional layer right before the global average pooling, obtaining features with dimensionality of 1024 x 14 x 14. The LSTMs used in the Dynamic Speaker have a hidden state dimension of 512. The word embedding layer is trained from scratch and each word is represented by a 300-dim vector. We train our model for 40 epochs using the Adam Optimizer [32] with a learning rate of 0.001 and a batch size of 128. The hyperparameters for the regularization terms are \(\lambda_{l_1} = 2.5e^{-03}\) and \(\lambda_{ent} = 0.0001\). Our code and dataset will be made publicly available at github.com/Seth-Park/RobustChangeCaptioning.

Evaluation. To evaluate change captioning, we rely on BLEU-4 [44], METEOR [7], CIDEr [56], and SPICE [2] metrics which measure overall sentence fluency and similarity to ground-truth. For change localization, we rely on the Pointing Game evaluation [63]. We use bilinear interpolation to upsample the attention maps to the original image size, and check whether the point with the highest activation “falls” in the ground-truth bounding box.

5.2. Results on CLEVR-Change Dataset

Pixel vs. representation difference [25] utilize pixel difference information when generating change captions under the assumption that the images are aligned. To obtain insights into whether a similar approach can still be effective when a camera position changes, we introduce the following baselines: Capt-Fix-Diff is a model that directly utilizes...
Table 2: Change Captioning evaluation on our CLEVR-Change Dataset. Our proposed model outperforms all baselines on BLEU-4 (B), CIDEr (C), METEOR (M), and SPICE (S) in each setting (i.e. Total, Scene Change, Distractor).

<table>
<thead>
<tr>
<th>Approach</th>
<th>Total</th>
<th>Scene Change</th>
<th>Distractor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>C</td>
<td>M</td>
</tr>
<tr>
<td>Capt-Pix-Diff</td>
<td>30.2</td>
<td>75.9</td>
<td>23.7</td>
</tr>
<tr>
<td>Capt-Rep-Diff</td>
<td>33.5</td>
<td>87.9</td>
<td>26.7</td>
</tr>
<tr>
<td>Capt-Att</td>
<td>42.7</td>
<td>106.4</td>
<td>32.1</td>
</tr>
<tr>
<td>Capt-Dual-Att</td>
<td>43.5</td>
<td>108.5</td>
<td>32.7</td>
</tr>
<tr>
<td>DUDA (Ours)</td>
<td>47.3</td>
<td>112.3</td>
<td>33.9</td>
</tr>
</tbody>
</table>

Table 3: A Detailed breakdown of Change Captioning evaluation on our CLEVR-Change Dataset by change types: Color (C), Texture (T), Add (A), Drop (D), Move (M), and Distractor (DI).

<table>
<thead>
<tr>
<th>Approach</th>
<th>C</th>
<th>T</th>
<th>A</th>
<th>D</th>
<th>M</th>
<th>DI</th>
<th>C</th>
<th>T</th>
<th>A</th>
<th>D</th>
<th>M</th>
<th>DI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capt-Pix-Diff</td>
<td>4.2</td>
<td>16.1</td>
<td>30.1</td>
<td>27.1</td>
<td>18.0</td>
<td>98.2</td>
<td>7.4</td>
<td>16.0</td>
<td>24.4</td>
<td>20.9</td>
<td>18.2</td>
<td>38.9</td>
</tr>
<tr>
<td>Capt-Rep-Diff</td>
<td>44.5</td>
<td>21.9</td>
<td>50.1</td>
<td>49.7</td>
<td>26.5</td>
<td>105.3</td>
<td>19.2</td>
<td>18.2</td>
<td>25.7</td>
<td>23.5</td>
<td>18.9</td>
<td>41.7</td>
</tr>
<tr>
<td>Capt-Att</td>
<td>112.1</td>
<td>75.9</td>
<td>91.5</td>
<td>98.4</td>
<td>49.6</td>
<td>106.6</td>
<td>30.5</td>
<td>25.4</td>
<td>30.2</td>
<td>31.2</td>
<td>22.2</td>
<td>43.2</td>
</tr>
<tr>
<td>Capt-Dual-Att</td>
<td>115.8</td>
<td>82.7</td>
<td>85.7</td>
<td>103.0</td>
<td>52.6</td>
<td>108.9</td>
<td>32.1</td>
<td>26.7</td>
<td>29.5</td>
<td>31.7</td>
<td>22.4</td>
<td>44.0</td>
</tr>
<tr>
<td>DUDA (Ours)</td>
<td>120.4</td>
<td>86.7</td>
<td>108.2</td>
<td>103.4</td>
<td>56.4</td>
<td>110.8</td>
<td>32.8</td>
<td>27.3</td>
<td>33.4</td>
<td>31.4</td>
<td>23.5</td>
<td>45.2</td>
</tr>
</tbody>
</table>

Table 4: Pointing game accuracy results. We report per change-type performance (Color (C), Texture (T), Add (A), Drop (D), Move (M)) and model interplay (DI). The numbers are in %.

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>T</th>
<th>A</th>
<th>D</th>
<th>M</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capt-Att</td>
<td>46.68</td>
<td>57.9</td>
<td>22.8</td>
<td>47.8</td>
<td>17.5</td>
<td>39.37</td>
</tr>
<tr>
<td>Capt-Dual-Att</td>
<td>40.97</td>
<td>46.5</td>
<td>54.3</td>
<td>45.6</td>
<td>19.8</td>
<td>39.35</td>
</tr>
<tr>
<td>DUDA (Ours)</td>
<td><strong>54.52</strong></td>
<td><strong>65.75</strong></td>
<td><strong>48.68</strong></td>
<td><strong>50.06</strong></td>
<td><strong>22.77</strong></td>
<td><strong>48.10</strong></td>
</tr>
</tbody>
</table>
Figure 4: Qualitative results comparing Capt-Att and DUDA. The blue and red attention maps are applied to “before” and “after”, respectively. The blue and red attention maps are the same for Capt-Att whereas in DUDA they are separately generated. The heat map on the lower-right is the visualization of the dynamic attention weights where the rows represent the amount of attention given to each visual feature (e.g. loc bef, diff, loc aft) per word.

Figure 5: Change captioning and localization performance breakdown by viewpoint shift (measured by IoU).

tention learns the spatial correspondence between two images w.r.t. the changed object and the Dynamic Speaker facilitates the learning of the Dual Attention. We now further validate such robustness by analyzing the performance under varying degrees of viewpoint shift. To measure the amount of viewpoint shift for a pair of images, we use the following heuristics: for each object in the scene, excluding the changed object, we compute the IoU of the object’s bounding boxes across the image pair. We assume the more the camera changes its position, the less the bounding boxes will overlap. We compute the mean of these IoUs and sort the test examples based on this (lower IoU means higher difficulty). The performance breakdown in terms of change captioning and localization is shown in Figure 5. Our model outperforms the baselines on both tasks, including the more difficult samples (to the left). We see that both captioning and localization performance degrades for the baselines and our model (although less so) as viewpoint shift increases, indicating that it is an important challenge to be addressed on our dataset.

Figure 6 illustrates two examples with large viewpoint changes, as measured by IoU. The overlaid images show that the scale and location of the objects may change significantly. The left example is a success, where DUDA is able to tell that the object has disappeared. Interestingly, in this case, it rarely attends to the “difference” feature. The right example illustrates a failure, where DUDA predicts that no change has occurred, as a viewpoint shift makes it difficult to relate objects between the two scenes. Overall, we find that most often the semantic changes are confused with the distractors (no change) rather than among themselves, while MOVE suffers from such confusion the most.

5.3. Results on Spot-the-Diff Dataset

We also evaluate our DUDA model on the recent Spot-the-Diff dataset [25] with real images and human-provided descriptions. This dataset features mostly well aligned image pairs from surveillance cameras, with one or more changes between the images (no distractors). We evaluate our model in a single change setting, i.e. we generate a single change description, and use all the available human descriptions as references, as suggested by [25].
Figure 6: Qualitative examples of DUDA. The left is an example in which DUDA successfully localizes the change and generates correct descriptions with proper modulations among “before”, “diff”, and “after” visual features. The right example is a failure case. We observe that significant viewpoint shift leads to incorrect localization of the change, thus confusing the dynamic speaker.

<table>
<thead>
<tr>
<th>Approach</th>
<th>B</th>
<th>C</th>
<th>M</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDLA* [25]</td>
<td>0.081</td>
<td>0.340</td>
<td>0.115</td>
<td>0.283</td>
</tr>
<tr>
<td>DUDA*</td>
<td>0.081</td>
<td>0.325</td>
<td>0.118</td>
<td>0.291</td>
</tr>
</tbody>
</table>

Table 5: We evaluate our approach on the Spot-the-Diff dataset [25]. * We report results averaged over two runs, for DDLA [25], we use the two sets of results reported by the authors. See text for details.

We present our results in Table 5. The DDLA approach of [25] relies on precomputed spatial clusters, obtained using pixel-wise difference between two images, assuming that the images are aligned. For a fair comparison we rely on the same information: we extract visual features from both “before” and “after” images using the spatial clusters. We apply Dual Attention over the extracted features to learn which clusters should be relayed to the Dynamic Speaker. The rest of our approach is unchanged. As can be seen from Table 5, DUDA matches or outperforms DDLA on most metrics. We present qualitative comparison in Figure 7. As can be seen from the examples, our DUDA model can attend to the right cluster and describe changes corresponding to the localized cluster.

Despite the usage of natural images and human descriptions, the Spot-the-Diff dataset is not the definitive test for robust change captioning as it does not consider the presence of distractors. That is, one does not have to establish whether the change occurred as there is always a change between each pair of images, and the images are mostly well-aligned. We advocate for a more practical setting of robust change captioning, where determining whether the change is by itself relevant is an important part of the problem.

6. Conclusion

In this work, we address robust Change Captioning in the general setting that includes distractors. We propose the novel Dual Dynamic Attention Model to jointly localize and describe changes between images. Our dynamic attention scheme is superior to the baselines and its visualization provides an interpretable view on the change caption generation mechanism. Our model is robust to distractors in the sense that it can distinguish relevant scene changes from illumination/viewpoint changes. Our CLEVR-Change Dataset is a new benchmark, where many challenges need to be addressed, e.g. establishing correspondences between the objects in the presence of viewpoint shift, resolving ambiguities and correctly referring to objects in complex scenes, and localizing the changes in the scene amidst viewpoint shifts. Our findings inform us of important challenges in domains like street scenes, e.g. “linking” the moved objects in before/after images, as also noted in [25]. Our results on Spot-the-Diff are complementary to those we have obtained on the larger CLEVR-Change dataset. While Spot-the-Diff is based on real images, there are minimal or no distractor cases in the dataset. This suggests that valuable future work will be to collect real-image datasets with significant semantic and distractor changes.

Acknowledgements

This work was in part supported by the US DoD, Berkeley Deep Drive (BDD), and Berkeley Artificial Intelligence Research (BAIR) Lab.
References


[17] Lionel Gueguen and Raffay Hamid. Large-scale damage detection using satellite imagery. pages 1321–1328, 2015. 1, 2


[24] Rui Huang, Wei Feng, Zezheng Wang, Mingyuan Fan, Liang Wan, and Jizhou Sun. Learning to detect fine-grained change under variant imaging conditions. 2017. 2


[26] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. 2017. 2, 4


[33] Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. 2019. 4

[34] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. 2019. 4


[37] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. 2016. 3, 6

[38] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. pages 7219–7228, 2018. 1


[40] Ruotian Luo and Gregory Shakhnarovich. Comprehension-guided referring expressions. 2017. 1, 2


[45] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. 2018. 1, 2, 3, 6


[48] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. 2018. 5


[56] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. 2015. 5


[60] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. pages 21–29, 2016. 3, 6

[61] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speakerlistener-reinforcer model for referring expressions. 2017. 1, 2
