Attending to Visual Differences for Situated Language Generation in Changing Scenes

Anonymous ACL submission

Abstract

We investigate the problem of generating utterances from pairs of images showing a before and an after state of a change in a visual scene. We present a transformer model with difference attention heads that learns to attend to visual changes in consecutive images via a difference key. We test our approach in instruction generation, change captioning and difference spotting and compare these tasks in terms of their linguistic phenomena and reasoning abilities. Our model outperforms the state-of-the-art for instruction generation on the BLOCKS and difference spotting on the Spot-the-diff dataset and generates accurate referential and compositional spatial expressions. Finally, we identify linguistic phenomena that pose challenges for generation in changing scenes.

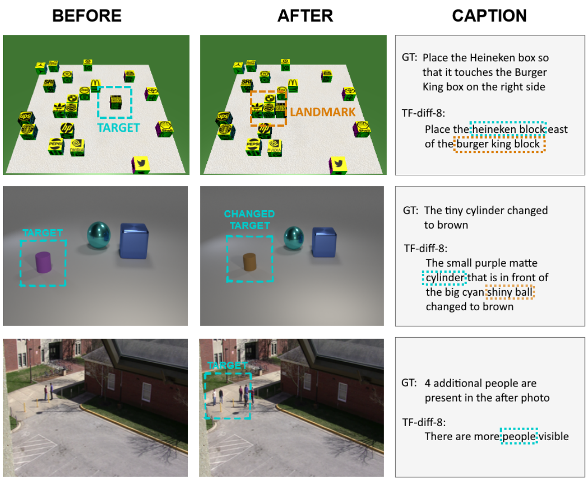


Figure 1: Image-pairs from BLOCKS, CLEVR-Change and Spot-the-diff (top to bottom) with descriptions generated by our best model. The targets and landmarks are manually highlighted for better view.

1 Introduction

Traditionally, work on situated language generation had to rely on symbolic representations of visual environments, cf. (Dale and Reiter, 1995; Chen et al., 2010; Dethlefs and Cuayáhuitl, 2015). Recent work has addressed language generation from images of visual scenes, e.g., in image captioning (Anderson et al., 2018; Cornia et al., 2020), referring expression generation (Yu et al., 2016; Panagiaris et al., 2020) or visual dialogue (Suhr et al., 2019; Agrawal et al., 2015). In other tasks like instruction generation, however, symbolic representations are still used to represent changing scenes and to model reasoning over sequences of states or trajectories in an environment (Fried et al., 2017; Köhn et al., 2020; Schumann and Riezler, 2021), sometimes in combination with images (Fried et al., 2017, 2018).

In this paper, we investigate natural language generation (NLG) in changing scenes from image-only input. Our goal is to detect visual changes and express them in complex referential and compositional language, without the need for elaborate image preprocessing or decomposition as in previous work on change detection in computer vision (Shi et al., 2020; Oluwasanmi et al., 2019a; Gilton et al., 2020). Furthermore, the idea is to model instruction generation without the need for symbolic specification of an action trajectory (Fried et al., 2018), but to learn both reasoning about changes and verbalizing them from images directly. Thus, we present a transformer that generates a verbalization of a change given a pair of images showing a "before state" and an "after state" as can be seen in Figure 1. Our model has multiple *difference attention heads* which learn to relate and attend to relevant regions in the before and after image.

Image pair-based language generation is useful in various tasks that involve changing scenes, such as instruction giving (Rojowiec et al., 2020), difference spotting (Jhamtani and Berg-Kirkpatrick, 2018) or change captioning (Park et al., 2019). Though technically similar, these tasks have been neither modeled in a common framework nor compared in terms of the involved linguistic phenomena and reasoning abilities.

Our contributions are (i) a novel difference attention-based model designed to visually ground complex compositional referential and spatial language in image pairs (Section 3), (ii) a systematic, qualitative comparison of instruction giving, different spotting and change captioning as well as the corresponding visual-linguistic reasoning phenomena (Section 4), (iii) experiments on these three tasks showing that our model achieves similar or superior performance to related state-of-the-art models for change detection from computer vision (CV), see Section 5, according to evaluation with automatic metrics, including metrics that aim at capturing the identified reasoning abilities.

2 Related Work

Instruction Generation is a central task in situated NLG, needed in agents that support humans in carrying out tasks in a shared environment. Previous work on instruction giving in virtual environments has developed planning-based frameworks for verbalising state and action sequences for a human listener, allowing for adaptive generation at different levels of detail (Koller et al., 2010; Dethlefs and Cuayáhuitl, 2015; Köhn et al., 2020). Fried et al. (2017, 2018) extend this line of work and propose a speaker model that generates text based upon visual input and associated symbolic action trajectories, also focussing on pragmatically appropriate, adaptive instructions. Hu et al. (2019) use verbal instructions as representations for action sequences in decision making for high-level planning. Rojowiec et al. (2020), instead, adopt a different perspective and model instruction generation for very local changes in a scene, learning directly from image pairs. Here, the focus is less on pragmatics and more on the semantic and referential accuracy of the instruction, which is difficult to achieve without a symbolic representation. Our work adopts Rojowiec et al.'s set-up, but outperforms their model and compares it to work on change captioning and difference spotting.

Change Detection and Captioning Change detection and its verbalization is an important task in CV, e.g. in captioning surveillance videos (Oh et al., 2011) or remote sensing images (Liu et al., 2018), and builds upon captioning of single images, one of the most well-understood tasks in language & vision. In image captioning, a successful encoding of the visual input that captures an image's content, its objects and their spatial relations has proven to be central (Bernardi et al., 2016; Lu et al., 2017; Anderson et al., 2018; Yao et al., 2018; Yang et al., 2019). A well-known attention mechanism is self-attention (Xu et al., 2015), which is also part of recent image captioning transformers (Herdade et al., 2019; Cornia et al., 2020). For captioning changes, Park et al. (2019)'s recurrent DUDA model exploits differences in latent space. Shi et al. (2020) expand on this by slicing the image into different patches and patch-wise-comparing differences which helps in distinguishing regions where changes occurred from non-changed regions. Oluwasanmi et al. (2019a,b) use a siamese network to encode before and after state, apply a contrastive function on both and then iteratively use softmax attention over the contrastive image in the decoder. While these approaches rely on elaborate methods for decomposing the visual input into regions of relevant semantic features and recurrent neural networks for decoding, we present a relatively simple encoder component as part of a transformer model which is, in contrast to existing work in image captioning, able to encode and attend to differences between a given pair of input images.

Visual Reasoning in L&V is often understood as the task of interpreting complex compositional phenomena like questions, comparisons, spatial expressions, quantification or counting (Suhr et al., 2017, 2019; Johnson et al., 2017; Li et al., 2019; Tan and Bansal, 2019; Shridhar et al., 2020; Li et al., 2020). Similarly to our set-up, NLVR (Suhr et al., 2017) involves determining the truth value of statements about two different images. Also highly related is work on instruction following (Misra et al., 2018; Chen et al., 2019) where the agent needs to resolve instructions to reach a goal state. In our case, the current and the goal state are given and the agent needs to generate a corresponding utterance. Our set-up involves different phenomena of visual reasoning, described in Section 4.

3 Model

We present a transformer model that generates utterances from a pair of images showing a before state and an after state of a change in a visual scene. To achieve this, we implement a difference attention head that computes an attention map for an image based on the difference to its before image (Section 3.1). We use this head to encode visual changes on different levels of granularity (Section 3.2). This encoder is hooked up with a standard transformer (Section 3.3).

3.1 Difference Attention on Image-Pairs

A core element of the standard transformer (Vaswani et al., 2017) is the self-attention head, which computes an attention map over values $\mathbf{V}$ given queries $\mathbf{Q}$ and keys $\mathbf{K}$:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\intercal}{\sqrt{d_k}})\mathbf{V} \quad (1)$$

When processing word sequences, the query, key and value of a self-attention head are given by the embedding of a word. A very simple way to process image pairs alike with this head, is to allocate two self-attention heads $H = 2$: one for the before image embedding v_1 and one for the after image embedding v_2 such that there are as many images as heads with $\mathbf{Q} = \mathbf{K} = \mathbf{V} = v_i$ and defined as:

$$h_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{K}\mathbf{W}_i^{\mathbf{K}}, \mathbf{V}\mathbf{W}_i^{\mathbf{V}}) \quad (2)$$

Now, we propose a difference attention head that exploits an explicit representation of the difference between the before and after state when computing the attention map. In line with Park et al. (2019), we simply substract the before from the after image. As there is no before image for v_1, we obtain two difference attention heads for our image pair: (i) h_1 with $\mathbf{K} = c_1 = 0$, (ii) h_2 with $\mathbf{K} = c_2 = v_2 - v_1$.

In line with Park et al. (2019), we scale the output of the difference with a trainable parameter γ and sum it with the image features for that attention head (weights are omitted for better readability, but applied as in Equation 2):

$$h_i = \gamma \cdot \text{Attention}(v_i, c_i, v_i) + v_i \quad (3)$$

This simple modification of self attention takes the idea of difference images from Park et al. (2019) and implements difference attention heads in a similar way as cross-modal attention in V&L transformers (Tan and Bansal, 2019; Lu et al., 2019).

3.2 Attending to In-between Images

We hypothesize that, to fully leverage the power of difference attention, more heads, i.e. a longer sequence of visual inputs, might be beneficial for grounding and generating utterances. Thus we increase the number of difference attention heads to $H = \{4, 8\}$, where v_H is the after image, and we define a way to compute "in-between image features" for the additional heads:

$$v_t = v_1 + c_t \quad (4)$$

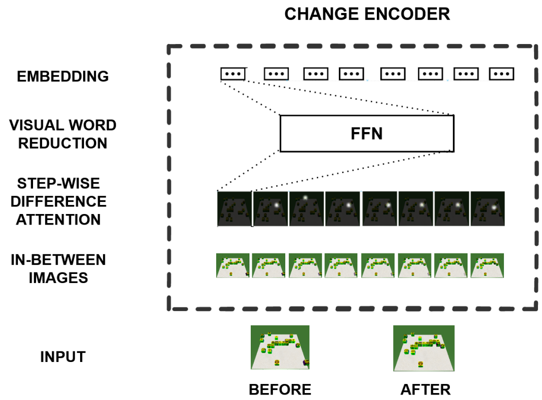


Figure 2: We simulate a trajectory of images with incremental changes given only a before and an after image to apply difference attention with higher granularity.

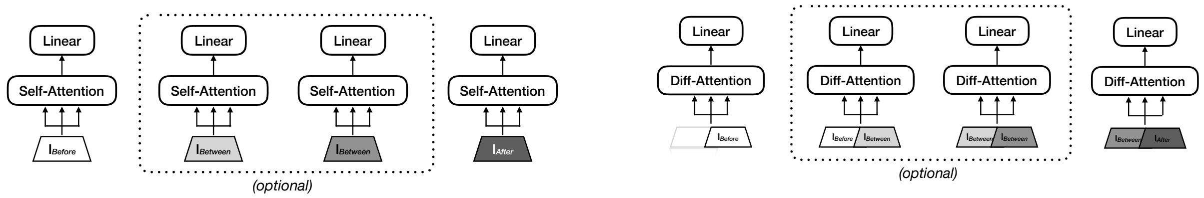
Inutitively, the in-between images represent the trajectory from the before to the after state, as shown in Figure 2. Formally, we define c_t as the weighted difference features, where the weight is the relative position in the trajectory between v_1 and v_H. Thus, each attention head receives image features representing a different degree of the visual change given by $v_H - v_1$:

$$c_i = \frac{i-1}{H-1} \cdot (v_H - v_1) \text{ where } i \in [1, H] \quad (5)$$

Finally, a single-layer feed forward network maps from the high-dimensional visual image space $2048 \times 14 \times 14$ to the reduced visual word space of 512 dimension $\hat{h}_i = r(h_i)$ and a downstream standard transformer receives the stacked sequence of visual words that represent various levels of change:

$$V = [\hat{h}_1; ...; \hat{h}_H] \quad (6)$$

The number of attention heads H is a hyperparameter, which can also be interpreted as a measure of granularity for the simulated visual feature trajectory $\{v_1, ..., v_t, ..., v_H\}$ where later image features contain more and more changes starting from the before image v_1. We report results for 2 and 8 headsto show the effects of a longer trajectories, leaving further experimentation for future work. As baselines, we implement a standard transformer, **TF-self-att**, that computes an attention map for every encoded image of step i simply with self-attention (see Figure 3). These are compared to **TF-diff-att**, the transformer with difference attention. Figure 3 shows how self and difference attention process a sequence of before, in-between and after images.



(a) Self-attention

(b) Difference attention

Figure 3: Sequential self and difference attention

3.3 Overall Architecture

We encode the before and after images with a pre-trained ResNet-101 architecture (He et al., 2016) trained on ImageNet, without any further preprocessing (like e.g. object detection). Our image pair encoder optionally transforms the image pair into a longer sequence containing in-between images. This trajectory is processed by a difference attention layer and then mapped to a sequence of visual words as shown in Figure 2. We apply positional encoding to the visual words generated by the image pair encoder to introduce temporal information into the encoded input. These visual words are processed like embedded word tokens within the 6 layers of the multi-head-attention-based transformer encoder. In the transformer decoder, an embedding layer first maps the words to vectors and then applies masked-self-attention followed by encoder-decoder attention which relates the visual words to words in the caption. In this architecture, difference and self-attention are used consecutively one after the other. In future work, further combinations can be investigated.

The recurrent DUDA model (Park et al., 2019), which is an important baseline in our experiments, uses a different way to compute attention maps based on image differences: first, the difference image is concatenated with the latent before and after image. Second, a self-attention map is computed over each of these and, third, another attention map over the attended concatenated before, after and difference image. Here, intuitively, the different visual inputs are kept separate and the model has to learn when to look at the before, after or difference image. Our approach, in contrast, incorporates differences as a key into the attention head. Intuitively, this corresponds to the idea that the difference image *relates* the after to the before image and that attention maps should capture these relations.

4 Tasks, Environments and Phenomena

We investigate different tasks for generation in changing scenes (Section 4.1). We describe their linguistic differences (Section 4.2 and 4.3), and discuss strengths and weaknesses (Section 4.4).

4.1 Tasks and Environments

Instructions BLOCKS (Bisk et al., 2016) is a dataset of movement instructions for blocks on a simple virtual 3D board (see Figure 1). The image pairs have been generated by down-sizing MNIST images, decorating the resulting blocks with digits or brand logos and randomly move the block's pixels to other positions, one at a time. This sequence in reverse order corresponds to an action sequence for assembling a block configuration that visually represents a number. For each single action, i.e. image pair, Bisk et al. (2016) collected 9 natural language instructions from 3 different crowd-workers. The workers were asked to provide instructions as if they would give them to another person in order to transform the block configuration. While BLOCKS was originally designed for instruction following, Rojowiec et al. (2020) analyze its use for instruction giving.

Differences Spot-the-diff (Jhamtani and Berg-Kirkpatrick, 2018) provides pairs of similar images extracted from real-word surveillance videos. The image pair shows a scene from the same viewpoint in different, but similar states (according to L_2 distance) resulting in very subtle differences that are difficult to spot. Jhamtani and Berg-Kirkpatrick (2018) collected descriptions of these pairs via crowdsourcing and instructed workers to "carefully study the image", "give sufficient time as some difference may not be obvious" and to provide complete English sentences for each difference.

Changes CLEVR-Change (Park et al., 2019) provides synthetic captions for images with changes in a virtual 3D-scene with objects of different shapes and colors. The image pairs are generated in Johnson et al. (2017)'s CLEVR framework and show a change affecting a property of a single object in the scene: (i) color, (ii) texture, (iii) location, (iv) object added, (v) object removed. A template-based generator was used to produce up to 9 different captions of varying length for each pair. Park et al. (2019)'s work is motivated by applications in surveillance and satellite imagery so that they include distractor pairs with non-semantic changes, i.e. change of camera angle or illumination. We do not include this subset in our experiments, in order to avoid introducing to many conceptual differences between our tasks (i.e. BLOCKS and Spot-the-diff contain semantic changes only).

4.2 Reference

Reference to objects in the environment is an important phenomenon in all our tasks, though their referring expressions differ in complexity.

Target object references In all our set-ups, the reference to a target object that changed one of its properties or (dis)-appeared is a key element of the caption. Thus, if an instruction in BLOCKS does not mention the correct target, a potential follower will not be able to execute it in any way. Similarly, in Spot-the-diff and CLEVR-change the meaningfulness of the caption hinges on the mention of the proper target. In BLOCKS, there is one ground-truth target object for each image pair that is generally referred to by its identifying logo, e.g. *the Heineken box* in Figure 1. Thus, references to targets in BLOCKS can be detected in human and generated captions with a simple, rule-based instruction parser (Rojowiec et al., 2020). In Spot-the-diff, there might be several target objects and they are referred to by a more complex vocabulary, e.g. *additional people* in Figure 1. The dataset does not provide a language-external annotation for ground-truth target objects and they cannot be easily detected in an automatic way. In CLEVR-change, expressions referring to targets correspond to the templates of the generator, i.e. they consist of a noun for the shape of the object and optional adjectives referring to the size, color or texture of the object, e.g. *the tiny cylinder* in Figure 1. This template can be automatically detected by a parser reverting the generator.

Landmark object references As the instructions in BLOCKS require detailed descriptions of block configurations, they commonly contain references to landmark objects, e.g. *right of the Burger King block* in Figure 1. In contrast to the target objects, there might be several landmarks produced by different crowd-workers. Generating one of the correct landmarks is important for the success of the instruction, as the BLOCKS environments provides few other means of verbalizing the movement and target location of the target object. A portion of the captions in CLEVR-change also contains landmarks as part of some of the templates of the generator. By qualitative inspection of Spot-the-diff, we establish that landmark objects are mentioned occasionally (e.g. *person behind black suv*, cf. p.3 (Jhamtani and Berg-Kirkpatrick, 2018)), but less systematically as in BLOCKS and CLEVR-change.

4.3 Reasoning

Our set-ups vary further with respect to phenomena related to compositional visual reasoning.

Compositional spatial expressions Many instructions in BLOCKS contain complex, compositional spatial expressions with one or more embedded prepositional and verb phrases, e.g. *place it lined up directly to the right of ...* in Figure 1. Spot-the-diff and CLEVR-change are much less complex in this regard. For instance, the template for location changes in CLEVR-change corresponds to the simple pattern: *object X has changed its location*. Spot-the-diff features occasional, simple spatial expressions, e.g. *people in the middle of the court*, cf. p.4 (Jhamtani and Berg-Kirkpatrick, 2018).

Types of changes BLOCKS instructions feature one type of visual change, i.e. block movement. Here, CLEVR-change is the most complex dataset as captions need to distinguish and refer to 5 different change types. Many Spot-the-diff descriptions refer to the (dis)-appearance of objects, but some also describe movements.

Changing object properties Objects in BLOCKS and Spot-the-diff do not change their internal properties whereas objects in CLEVR-change do change their color or texture (cf. Figure 1), resulting in a complex representation task regarding the identity of objects.

4.4 Summary

The set-ups we investigate in this work are highly similar in terms of the modeling task, i.e. generating an utterance given a pair of images showing similar states of the same scene. At the same time, different visual environments and data collections led to substantial differences in the reasoning abilities that the models will need to account for, see Table 5 in Appendix A.1 for an overview. Generally, BLOCKS and Spot-the-diff exhibit more linguistic complexity than CLEVR-change: BLOCKS instructions have been collected in a dialogue-inspired setting and the resulting utterances are varied, goal-oriented and contain complex spatial expressions. Spot-the-diff utterances are more descriptive and might not naturally occur in situated dialogue, but they still refer to complex real-world scenes and draw on a natural vocabulary. CLEVR-change captions are synthetic and do not constitute natural dialogue data, but they exhibit greater complexity in terms of visual reasoning, i.e. detecting changes of different types, including changes of internal object properties.

5 Experiments

5.1 Data

BLOCKS: We use the MNIST-logo subset with constellations of up to 20 cubes with distinct logos. It is split into $667/95/181$ image pairs for training, validation and testing and $6003/855/1629$ captions respectively (9 per image pair).

Spot-the-Diff: We use the entire dataset of $9524/1634/1404$ image-pairs for training, validation and testing and $17676/3310/2107$ captions respectively. When an image-pair has less than 3 captions, we re-sample from the given ones, so that during training each pair is seen 3 times per epoch.

CLEVR-Change: We use the splits from Park et al. (2019), but only the semantic change subset with $33830/1988/3985$ image-pairs for training, validation and testing and $250415/14651/29654$ captions, i.e. up to 9 captions per image-pair (avg. 7.4 captions). We sample in the same way as above, so that each image-pair is seen 9 times per epoch.

5.2 Training and Hyperparameters

We encode the before and after image separately using a pre-trained ResNet-101 with the last layer cut off which results in image embeddings of size $2048 \times 14 \times 14$ by applying adaptive pooling. The word embedding layer in the transformer decoder is trained from scratch with a size d of 512. We use Adam optimizer with a learning rate of 10^{-4} and a batch size of $8/16$ for training with $8/2$ heads respectively. We also perform early stopping after 5 epochs without improvement on the validation set and apply *Label Smoothing* as proposed by Vaswani et al. (2017). The training on a single NVIDIA Titan X GPU took up to three days for the CLEVR-Change dataset.

For BLOCKS, it turned out to be necessary to fine-tune the image encoder to recognize the small logos distinguishing the single blocks. The training regime on BLOCKS is a two-stage process: the models (DUDA and our transformer models) are first trained with a freezed, pre-trained image encoder, and then trained again by allowing gradients in the image encoder. For Spot-the-diff and CLEVR-Change, we do not fine-tune the image encoder to ensure comparability with previous work.

5.3 Evaluation Metrics

We measure the overlap of generated and human captions with BLEU-4, METEOR, CIDEr and SPICE, using the API of Chen et al. (2015). Furthermore, we assess the models' reasoning abilities on BLOCKS and CLEVR-change, according to the phenomena in Section 4.

For BLOCKS, we rely on Rojowiec et al. (2020)'s parser which detects expressions (phrases) referring to targets and landmarks in ground-truth and generated instructions. Following Rojowiec et al., we compute these word or phrase accuracies: (i) **target**: correctly generated targets, given all generated target phrases (ii) **landmark**: correctly generated landmarks, mentioning one of the landmarks logos from the set of landmarks found in the ground-truth instructions (iii) **spatial**: correctly generated words not contained in target and landmark phrases, as a simple metric for measuring overlap of spatial expressions.

For CLEVR-change, we write a similar parser that detects the template that was used to generate the caption. Based on the parser output, we compute the following accuracies: (i) **type**: portion of captions mentioning the correct change type (i.e. color, texture, add, drop, move) (ii) **target–color**, **target-shape**, **target-material**: portion of correctly generated color/shape/material attributes in target references (iii) **landmark-color**, **landmark-shape**, **landmark-material**: analogous to target accuracies.

5.4 Results

Qualitative samples of generation outputs are shown in Figure 1.

General performance across tasks Our transformer models with difference attention, TF-diff-att-2 and TF-diff-att-8, outperform state-of-the-art models for instruction generation (see BLOCKS results in Table 1) and difference spotting (see Spot-the-diff results in Table 2) in terms of all n-gram overlap metrics. Our version of DUDA trained on BLOCKS improves considerably over the results by Rojowiec et al. (2020), but not over our TF-diff models. On Spot-the-diff, as shown in Table 2, existing systems (mostly developed in the CV community) still obtain relatively low overlap scores. TF-diff-2 and TF-diff-8 improve over the state-of-the-art set by the M-VAM model on Spot-the-diff, with a particularly strong increase of the CIDEr score (0.425 and 0.843 respectively). Table 3 shows that the TF-diff models do not achieve state-of-the-art performance on CLEVR-change, but obtain similar SPICE scores as the DUDA model (see Appendix for other metrics and below for further analysis). In the majority of tasks and settings, transformers with difference attention outperform the standard self attention (TF-self models). This indicates that generation tasks with changing scenes involve complex visual and linguistic reasoning, which cannot be easily achieved with self attention.

In-between images On BLOCKS, TF-diff-8 clearly outperforms TF-diff-2, whereas on Spot-the-diff, TF-diff-2 outranks TF-diff-8. This suggests that difference attention on in-between images is beneficial for visual grounding of complex spatial configurations and landmarks, which are not prominent in Spot-the-Diff. On CLEVR-change, TF-diff-2 outperforms TF-diff-8 on the change type 'ADD' subset, which is in line with the performance of TF-diff-2 on Spot-the-diff (where it is common that new objects are added/appear in the after image). At the same time, TF-diff-8 outperforms TF-diff-2 on 'MOVE' changes in CLEVR-change which is in line with our results on BLOCKS (where objects are moved). Thus, our attention mechanisms behave similarly for similar reasoning abilities across the different tasks.

Reference On BLOCKS, the TF-diff-8 model greatly outperforms the competitive DUDA model in terms of accuracies on target and landmark reference, cf. Table 1. We note that the DUDA model performs better in generating references to targets (59% target accuracy on BLOCKS, and above 90% on CLEVR-change) as compared to landmarks

Model	B	M	C	Target	Landm	Other
LSTM+Att*	0.38	0.28	0.27	0.11	0.28	-
DUDA	0.53	0.37	0.96	0.59	0.42	0.66
TF-self-att-2	0.34	0.28	0.35	0.19	0.26	0.76
TF-self-att-8	0.44	0.32	0.66	0.37	0.45	0.72
TF-diff-att-2	0.55	0.38	1.06	0.73	0.40	0.80
TF-diff-att-8	**0.68**	**0.43**	**1.52**	**0.86**	**0.73**	**0.83**

Table 1: BLOCKS results: B(LEU-4), M(eteor), C(ider) and word accuracies (see Section 5.3), LSTM+Att* as reported in Rojowiec et al. (2020).

Model	B	M	C	S
DUDA*	0.081	0.115	0.34	-
FCC*	0.099	0.129	0.368	-
SDCM*	0.098	0.127	0.363	-
DDLA*	0.085	0.12	0.328	-
M-VAM + RAF*	0.111	0.129	0.425	0.171
TF-self-att-2	0.109	0.135	0.777	0.197
TF-self-att-8	0.110	0.136	0.786	0.191
TF-diff-att-2	**0.117**	**0.137**	**0.843**	**0.205**
TF-diff-att-8	0.113	0.136	0.842	0.202

Table 2: Spot-the-diff results: B(LEU-4), M(eteor), C(IDEr), S(PICE). *Models as reported in Shi et al. (2020)

Model	SPICE Color	Texture	Add	Drop	Move
DUDA*	0.21	0.18	0.22	0.22	0.15
M-VAM + RAF*	**0.30**	**0.30**	**0.32**	**0.33**	**0.30**
TF-self-att-2	0.19	0.17	0.18	0.20	0.18
TF-self-att-8	0.20	0.17	0.15	0.20	0.18
TF-diff-att-2	0.20	0.20	0.24	0.21	0.21
TF-diff-att-8	0.22	0.23	0.23	0.25	0.26

Table 3: CLEVR-change results: SPICE for test sets split up by change types: Color(C), Texture (T), Add (A), Drop (D), Move (M). DUDA is trained on the entire CLEVR-change data, the TF and M-VAM models on semantic changes only. *Models as reported in Shi et al. (2020).

Model	Type	Target S	C	T	Landmark S	C	T
DUDA	**0.79**	**0.95**	**0.99**	**0.88**	**0.38**	0.24	0.24
TF-self-2	0.41	0.64	0.63	0.65	0.29	0.29	0.21
TF-self-8	0.42	0.65	0.61	0.63	0.36	**0.31**	**0.25**
TF-diff-2	0.45	0.70	0.67	0.68	0.34	0.28	0.23
TF-diff-8	0.47	0.74	0.72	0.72	0.32	**0.31**	0.24

Table 4: CLEVR-change: accuracies for change types (type) and word accuracies for S(hape), C(olor), T(exture) in target/landmark references. DUDA is trained on the entire CLEVR-change data, the TF models on semantic changes only.

(42% landmark accuracy on BLOCKS, and below 40% on CLEVR-change). This pattern has, to the best of our knowledge, not been observed in previous work (Park et al., 2019; Shi et al., 2020). On BLOCKS, our TF-diff-2 model clearly improves DUDA's target accuracy (73% acc. for TF-diff-2), but performs similarly on the landmarks (40% acc. for TF-diff-2). The TF-diff-8 model gives further improvement on target objects (86%) and a great improvement on landmarks (73%). This shows that the in-between images combined with difference attention heads allow the transformer model to not only attend to target objects but also to "close-by" landmark objects, i.e. relating the before to the after image. These relations do not seem to be captured well in DUDA's dual attention. This is further illustrated by the example attention maps for TF-diff-att-8 in Figure 5 and DUDA in Figure 6 in Appendix A.2. While the DUDA map is rather fuzzy, the attention of TF-diff-att-8 model is located rather precisely on the target block, its target location and nearby landmarks. Similar tendencies for target and landmarks can be found in CLEVR-change, i.e. DUDA performs much worse on landmarks than on targets. Here, however, our transformers are clearly below DUDA's target accuracy. As we discuss below, this seems to result from the fact that the transformers do not learn certain other visual reasoning abilities on that dataset.

Change types and changing objects The evaluation on CLEVR-change in Table 6 shows an important limitation of our transformers: while DUDA accurately distinguishes between types of changes (e.g. color, add or move changes), all transformers tend to confuse them, e.g. TF-diff-8 achieves 47% and DUDA 79% acc. on change type detection. The confusion matrix in Table 8 (Appendix A.3) shows that the TF-diff-8 model often confuses changes of internal objects properties (color or texture) with moving and (dis)-appearing objects. This also explains why the TF-models perform below state-of-the-art models on this dataset. The example attention maps for TF-diff-att-8 in Figure 4 in Appendix A.2 further illustrates that our transformer does not seem to learn how to exploit the sequential difference attention for reasoning in CLEVR-change. Here, DUDA's dual attention (see Section 3.3) that treats the difference image as a parallel input modality (concatenated with the before and after state) seems to be a more adequate way of representing different visual states.

5.5 Summary and discussion

Our experiments show that instruction generation, change description and difference spotting accommodate different requirements for reasoning and generation in changing scenes. Our transformers achieve state-of-the performance on tasks that focus on linguistically complex, human-like descriptions of visual changes that involve moving or disappearing objects, i.e. instructions in BLOCKS and difference descriptions in Spot-the-diff. More work is needed to extend our approach with more flexible difference attention to be able to capture visual changes that affect internal object properties, i.e. as in CLEVR-change captions. More generally, we believe that analyzing the linguistic phenomena underlying these and other generation tasks and creating datasets that combine them in a systematic way is a highly fruitful direction for future work. Two phenomena that stand out in our experiments are (i) target-landmark configurations, which have received a lot of interest in traditional NLG (Clarke et al., 2013) and are relevant in, e.g., navigation (Schumann and Riezler, 2021) (ii) changing object properties, which might be highly relevant in complex real-world domains like, e.g. cooking (Yang et al., 2016). Another direction for future work is reliable set-ups for human evaluation, a vital topic in current NLG research (Howcroft et al., 2020; Belz et al., 2020). We believe that the tasks investigated here will pose their own challenges as, for instance, the difference between two images can be difficult to spot even for humans.

6 Conclusion

We have investigated language generation in changing scenes. We proposed a simple difference attention head that relates consecutive images in an input trajectory via a difference key. Our method sets a new state-of-the-art on BLOCKS (Bisk et al., 2016) and Spot-the-diff (Jhamtani and Berg-Kirkpatrick, 2018). We have shown that it is important to disentangle reasoning abilities resulting from differences in environments and data collections for change-related generation tasks. We conclude that our approach is able to model situated instruction giving for local changes on controlled visual inputs, while more work is needed to scale it to more realistic inputs and to longer sequences of states that are often looked at in situated interaction with symbolic representations like (Dethlefs and Cuayáhuitl, 2015; Fried et al., 2018; Köhn et al., 2020).

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. International Journal of Computer Vision, 123:4–31.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In Proceedings of the 13th International Conference on Natural Language Generation, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. Journal of Artificial Intelligence Research, 55:409–442.

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. 2016. Natural language communication with robots. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

David L Chen, Joohyun Kim, and Raymond J Mooney. 2010. Training a multilingual sportscaster: Using perceptual context to learn language. Journal of Artificial Intelligence Research, 37:397–435.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 12538–12547.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325.

Alasdair Daniel Francis Clarke, Micha Elsner, and Hannah Rohde. 2013. Where's wally: The influence of visual salience on referring expression generation. Frontiers in psychology, 4:329.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10578–10587.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. Cognitive science, 19(2):233–263.

Nina Dethlefs and Heriberto Cuayáhuitl. 2015. Hierarchical reinforcement learning for situated natural language generation. Natural Language Engineering, 21(3):391–435.

Daniel Fried, Jacob Andreas, and Dan Klein. 2017. Unified pragmatic models for generating and following instructions. arXiv preprint arXiv:1711.04987.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. 2018. Speaker-follower models for vision-and-language navigation. In Advances in Neural Information Processing Systems, pages 3314–3325.

Davis Gilton, R. Luo, R. Willett, and G. Shakhnarovich. 2020. Detection and description of change in visual streams. ArXiv, abs/2003.12633.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. arXiv preprint arXiv:1906.05963.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In Proceedings of the 13th International Conference on Natural Language Generation, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Hengyuan Hu, Denis Yarats, Qucheng Gong, Yuandong Tian, and Mike Lewis. 2019. Hierarchical decision making by generating and following natural language instructions. In Advances in neural information processing systems, pages 10025–10034.

Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. In EMNLP.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2901–2910.

Arne Köhn, Julia Wichlacz, Álvaro Torralba, Daniel Höller, Jörg Hoffmann, and Alexander Koller. 2020. Generating instructions at different levels of abstraction. In Proceedings of the 28th International Conference on Computational Linguistics, pages 2802–2813, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexander Koller, Kristina Striegnitz, Andrew Gargett, Donna Byron, Justine Cassell, Robert Dale, Johanna D Moore, and Jon Oberlander. 2010. Report on the second nlg challenge on generating instructions in virtual environments (give-2). In Proceedings of the 6th international natural language generation conference.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

Zhuowan Li, Quan Hung Tran, Long Mai, Zhe Lin, and A. Yuille. 2020. Context-aware group captioning via self-attention and contrastive features. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3437–3447.

Zhunga Liu, G. Li, G. Mercier, You He, and Q. Pan. 2018. Change detection in heterogenous remote sensing images via homogeneous pixel transformation. IEEE Transactions on Image Processing, 27:1822–1834.

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

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 375–383.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping instructions to actions in 3d environments with visual goal prediction. arXiv preprint arXiv:1809.00786.

S. Oh, A. Hoogs, A. Perera, Naresh P. Cuntoor, Chia-Chih Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, Hyungtae Lee, L. Davis, E. Swears, Xiaoyang Wang, Qiang Ji, K. K. Reddy, M. Shah, Carl Vondrick, H. Pirsiavash, D. Ramanan, Jenny Yuen, A. Torralba, Bi Song, Anesco Fong, A. Roy-Chowdhury, and Mita Desai. 2011. A large-scale benchmark dataset for event recognition in surveillance video. CVPR 2011, pages 3153–3160.

Ariyo Oluwasanmi, Muhammad Umar Aftab, Eatedal Alabdulkreem, Bulbula Kumeda, Edward Y. Baagyere, and Zhiquang Qin. 2019a. Captionnet: Automatic end-to-end siamese difference captioning model with attention. IEEE Access, 7:106773–106783.

Ariyo Oluwasanmi, E. Frimpong, Muhammad Umar Aftab, Edward Y. Baagyere, Zhiquang Qin, and Kifayat Ullah. 2019b. Fully convolutional captionnet: Siamese difference captioning attention model. IEEE Access, 7:175929–175939.

Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2020. Generating unambiguous and diverse referring expressions. Computer Speech & Language.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In Proceedings of the IEEE International Conference on Computer Vision, pages 4624–4633.

Robin Rojowiec, Jana Götze, Philipp Sadler, Henrik Voigt, Sina Zarrieß, and David Schlangen. 2020. From "before" to "after": Generating natural language instructions from image pairs in a simple visual domain. In Proceedings of the 13th International Conference on Natural Language Generation, pages 316–326, Dublin, Ireland. Association for Computational Linguistics.

Raphael Schumann and Stefan Riezler. 2021. Generating landmark navigation instructions from maps as a graph-to-text problem. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 489–502, Online. Association for Computational Linguistics.

Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq R. Joty, and Jianfei Cai. 2020. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. ArXiv, abs/2009.14352.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, R. Mottaghi, Luke Zettlemoyer, and D. Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 10737–10746.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Iris D. Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. ArXiv, abs/1811.00491.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490.

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

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pages 2048–2057.

Shaohua Yang, Qiaozi Gao, Changsong Liu, Caiming Xiong, Song-Chun Zhu, and Joyce Chai. 2016. Grounded semantic role labeling. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 149–159.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 10685–10694.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In Proceedings of the European conference on computer vision (ECCV), pages 684–699.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In European Conference on Computer Vision, pages 69–85. Springer.

A Appendix

A.1 Dataset overview

Table 5 shows a tabular overview of the tasks, environments and datasets used in this work. The Table summarizes the descriptions and discussion in Section 4.

A.2 Attention maps

Figure 4 and 5 show attention maps for the TF-diff-att-8 model on CLEVR-change and BLOCKS. The attention map for BLOCKS suggests that the model was able to precisely locate target and landmark objects, whereas the map on CLEVR-change does not indicates that the model detected a color change. Figure 7 shows an example of a very accurate attention map computed by the TF-diff-att-2 model on Spot-the-diff. Figure 6 shows an attention map of the DUDA model on BLOCKS, for the same scene shown in Figure 5. This example clearly illustrates that DUDA's dual attention mechanism exploits difference images in a very different way than our transformer, i.e. the attention map is much less focused on particular image regions.

A.3 Additional results on CLEVR-change

Table 6 shows CIDEr, METEOR and SPICE scores for our transformer models and three baselines on CLEVR-change. Overall, the transformer models are below the state-of-the-art set by the M-VAM+RAF model from Shi et al. (2020), as discussed in Section 5. Generally we believe that the most informative metrics on CLEVR-change are the accuracies reported in Table 4 as the captions in CLEVR-change are synthetic and use a rather small vocabulary.

Figure 8 shows the confusion matrix for change types: we identified the detected change types in generated captions using the caption parser and compare them to the ground-truth type.

	BLOCKS	Spot-the-diff	CLEVR-change
task	instruction giving	difference spotting	change captioning
language	human	human	synthetic
objects	virtual blocks (logos)	real objects	virtual objects (color, shape, texture)
changes	moves	moves, (dis-)appearance	color, texture, moves, (dis)-appearance
phenomena	logo identification, landmarks, spatial expressions	hardly visible changes, real-world target/landmark objects	landmarks, change types, changing object properties

Table 5: Overview of datasets summarizing Section 4

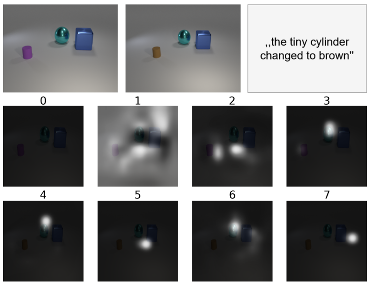


Figure 4: TF-diff-att-8 attention map on CLEVR-Change for the example from Fig. 1

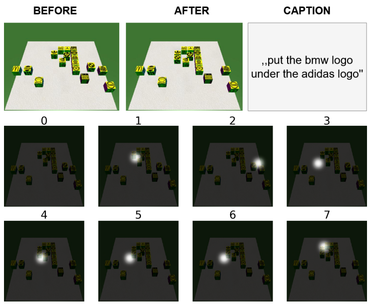


Figure 5: TF-diff-att-8: example caption and attention map on BLOCKS

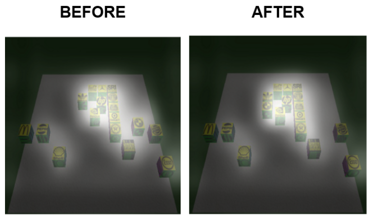


Figure 6: DUDA: example attention map on BLOCKS for the same example as in Figure 5

Model	CIDEr C	CIDEr T	CIDEr A	CIDEr D	CIDEr M	METEOR C	METEOR T	METEOR A	METEOR D	METEOR M	SPICE C	SPICE T	SPICE A	SPICE D	SPICE M
DUDA (with distractors)*	1.20	0.87	1.08	1.03	0.56	0.33	0.27	0.33	0.31	0.24	0.21	0.18	0.22	0.22	0.15
M-VAM + RAF (with distractors)*	1.22	0.98	1.26	1.16	0.82	0.36	0.32	0.38	0.36	0.28	0.28	0.27	0.31	0.32	0.23
M-VAM + RAF (w/o distractors)*	**1.35**	**1.08**	**1.30**	**1.13**	**1.07**	**0.38**	**0.36**	**0.38**	**0.37**	**0.36**	**0.30**	**0.30**	**0.32**	**0.33**	**0.30**
TF-self-att-2	0.69	0.44	0.56	0.47	0.43	0.27	0.25	0.27	0.27	0.26	0.19	0.17	0.18	0.20	0.18
TF-self-att-8	0.77	0.57	0.27	0.60	0.45	0.29	0.27	0.22	0.29	0.26	0.20	0.17	0.15	0.20	0.18
TF-diff-att-2	0.62	0.49	0.77	0.45	0.57	0.29	0.28	0.32	0.28	0.28	0.20	0.20	0.24	0.21	0.21
TF-diff-att-8	0.68	0.58	0.60	0.62	0.80	0.30	0.30	0.29	0.31	0.32	0.22	0.23	0.23	0.25	0.26

Table 6: Detailed breakdown of results on the CLEVR-Change Data set by change types: Color(C), Texture (T), Add (A), Drop (D), Move (M). Our models have only been trained on the semantic change set. *We report the results as provided by the authors in Shi et al. (2020)

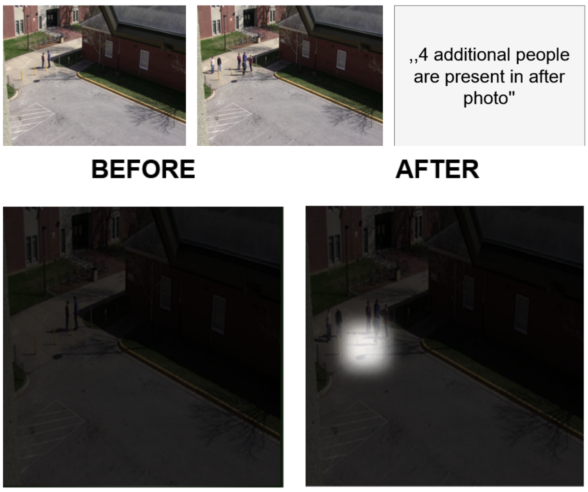


Figure 7: TF-diff-att-2 attention map on Spot-the-diff for the example from Fig. 1

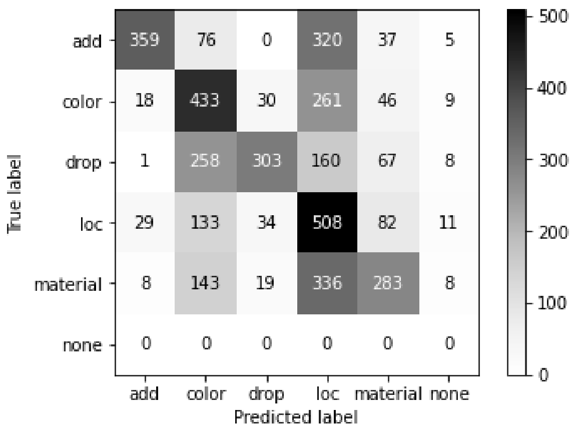


Figure 8: Confusion of change types in TF-diff-att-8 captions for CLEVR-change, change types in ground truth and generated captions are automatically recognized with a rule-based parser