Edit me: A Corpus and a Framework for Understanding Natural Language Image Editing

Ramesh Manuvinakurike*[2], Jacqueline Brixey*[2], Trung Bui[1], Walter Chang[1], Doo Soon Kim[1], Ron Artstein[2], Kallirroi Georgila[2]

[1]Adobe Research, San Jose, USA

[2]USC Institute for Creative Technologies, Los Angeles, USA

{manuvina,brixey}@usc.edu, {bui,wachang,dkim}@adobe.com, {artstein,kgeorgila}@ict.usc.edu

Abstract

This paper introduces the task of interacting with an image editing program through natural language. We present a corpus of image edit requests which were elicited for real world images, and an annotation framework for understanding such natural language instructions and mapping them to actionable computer commands. Finally, we evaluate crowd-sourced annotation as a means of efficiently creating a sizable corpus at a reasonable cost.

Keywords: dialogue, image editing, vision and language

1. Introduction

Photo editing is as old as photography itself. Over the years, darkroom techniques with film and light have given way to digital processing, and software suites such as Adobe Photoshop have made image editing accessible to millions of professionals and hobbyists. But even with the best tools, photo editing requires substantial knowledge, and novices often need to enlist the help of experts. Web sites such as https://www.reddit.com/r/PhotoshopRequest/ and http://zhopped.com/ contain thousands of image edit requests like the following:

- There is a spot on my wedding dress. Can someone please remove it. Please!
- Can you please fix the glare on my dog's eyes? I lost him today and he means the world to me.
- Can you please remove the people in the background? This is the only surviving photo of my mom and I would like to preserve it.
- I love this photo from our trip to Rome. Can someone please remove my ex from this photo? I am the one on the right.

As the examples above show, a natural way for novices to express their editing needs is through ordinary human language. At Adobe Research we aim to develop a software tool that will interpret such natural language image edit requests, and carry them out to the user's satisfaction.

This work presents a step in the direction of understanding human image edit requests: a corpus of such requests, and an annotation scheme for mapping these requests into actionable computer commands. A corpus could be compiled from naturally occurring examples such as the ones cited above, but this would raise concerns about privacy and ownership of the photographs, and the images themselves are unprocessed. Instead, we use a set of publicly available, richly annotated images called the Visual Genome corpus (Krishna et al., 2017). We elicited image edit requests that would pertain to these images, devised a scheme for mapping these requests into actionable commands, annotated requests according to the scheme, and evaluated the reliability of these annotations.

The contributions of this work are threefold. First, a dataset of natural language image edit requests for images with detailed image captions. These captions are particularly useful for the task of language understanding, as many of the requests make reference to objects in the images. Second, a framework for understanding these natural language instructions and mapping them to actionable computer commands. Finally, we provide a crowd-sourcing methodology to offload complex annotation between expert users and novice users and evaluate them. This is particularly useful for creating a sizable corpus.

2. Related Work

Recently, there has been a lot of work on applications that combine vision and language, e.g., understanding and generating image descriptions (Kulkarni et al., 2013), identifying visual reference in the presence of distractors (Paetzel et al., 2015; de Vries et al., 2016), visual question answering (Antol et al., 2015), visual storytelling (Huang et al., 2016), generating questions about an image (Mostafazadeh et al., 2016), and question-answer interactions grounded on information shown in an image (Mostafazadeh et al., 2017). Current image and language corpora typically consist of digital photographs paired with crowd-sourced captions (Lin et al., 2014; Krishna et al., 2017), or in some cases with questions related to those images (Mostafazadeh et al., 2016).

Much of the work above is relevant to the problem at hand. For example, understanding image descriptions is crucial for interpreting the requests quoted above, as all of them contain image descriptions (*my wedding dress; my dog's eyes; the people in the background; my ex*). However, to our knowledge, no work has yet attempted to tackle the specific task of automated image *editing* through natural language. Nor are we aware of any work that even tries to understand what users want to change when editing photos, and how

* Primary authors; work done while at Adobe Research

users talk about making those changes. This is the focus of the present work.

3. Data Collection

Edit requests were collected through crowd-sourcing using Amazon Mechanical Turk (https://mturk.com). We selected a small subset of images from the Visual Genome corpus (Krishna et al., 2017), a richly annotated set of about 108k images sampled from the MS COCO data-set (Lin et al., 2014). To provide enough visual detail for eliciting nuanced edit requests, we only used images with high resolution (1000×700 pixels and above). To elicit language that is similar to naturally occurring requests, we analyzed 200 posts from Reddit and Zhopped, and found that the images in those posts generally fell into eight high-level categories: animals, city scenes, food, nature/landscapes, indoor scenes, people, sports, and vehicles. We then chose images for elicitation that fit these categories. A total of 334 images were used for elicitation.

Each image was given to 5 crowd-source workers (called turkers), and each turker was asked to provide at least 5 unique edits that they would want to see in the image, phrased as requests in natural English (workers were free to provide more edits; Figure 1). The requests entered by the turkers were manually reviewed by the first two authors for quality and variation; incoherent submissions and unrelated requests were excluded. After filtering, we were left with 9101 edit requests with a total of 44727 word tokens (4628 unique word types). An example image with a few annotated requests is shown in Figure 2.

4. The Language of Image Edits

Review of the elicited requests provides insights into how users want to edit images.

Vocabulary. The elicited requests exhibit wide variation in vocabulary, with turkers using different terms to express essentially the same needs. For example, the requests *Make the colors pop*, *Bring out the colors*, and *Change the saturation* all express a desire for more vivid colors. Similarly, the desire for altering the dimensions of an image was expressed using the terms *crop*, *cut out*, and *delete*.

Ambiguity. Some terms are ambiguous between technical and general uses: the word *focus* may appear in a specific optical sense, but in the request *Make the bird the focus of the picture* it is probably used in the sense of general prominence. Technical terminology is also ambiguous, for instance the term *zoom*. A camera's zoom changes the focal length of the lens, and thus the angle of view and picture frame; whereas the zoom feature in a graphical user interface typically changes the view of a document. Thus, the interpretation of a request like *zoom in on the man* depends on context: as a standalone request, it probably represents the intention of changing the picture frame; but in interactive dialogue or as part of a multi-instruction sequence, it could indicate a request to change the view of the image in the editing interface.

Structure. Many of the edit requests contain a verb in the imperative, often at the beginning, such as in *Make the picture brighter*; sometimes the verb appears in a conjunct

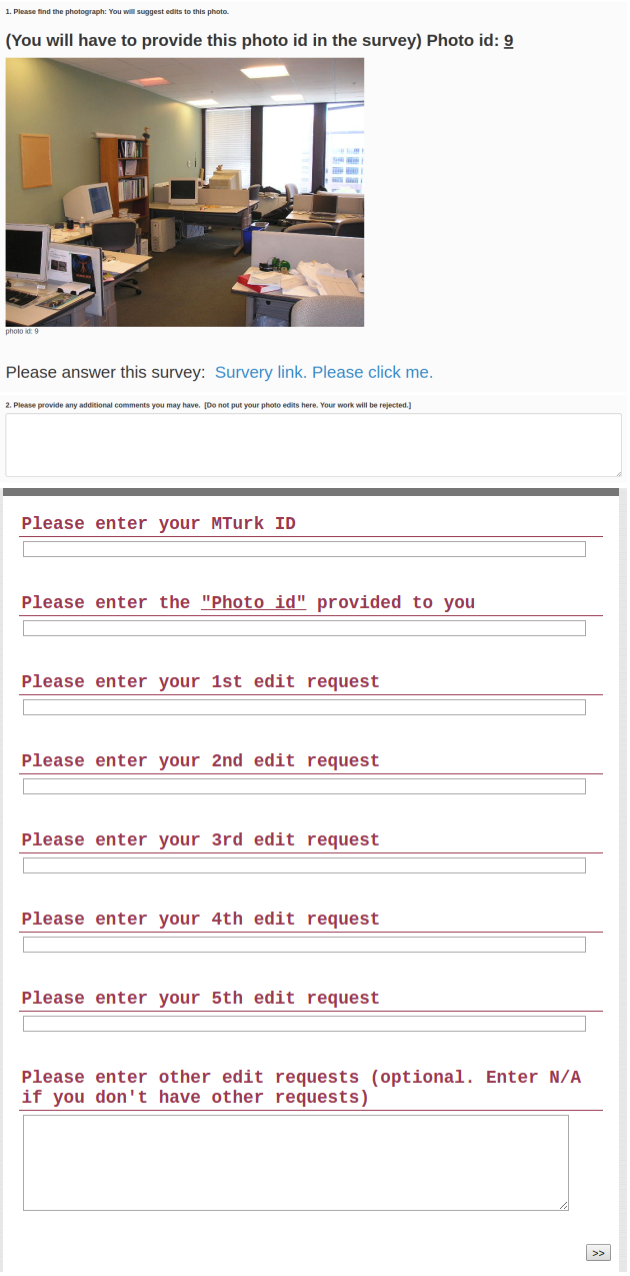

Figure 1: The interface shown to the turkers with the image for which they provide the editing commands. The turkers need to provide at least 5 unique edits.

clause, such as *The tree is distracting so remove it*. In other instances, however, the request takes the form of a comment, with the desire remaining implicit. For example, *The image is very blurry* suggests a goal of making the image less blurry, but does not directly state this intention nor how to achieve the goal.

Domain knowledge. Some variation in the language of edit requests can be attributed to the turkers' expertise. Novice turkers often use broad and high-level language, while those familiar with image editing may provide very specific instructions tailored for an editing tool. For example, a desire for better color balancing was expressed by a novice turker as *I would like to see more character and*

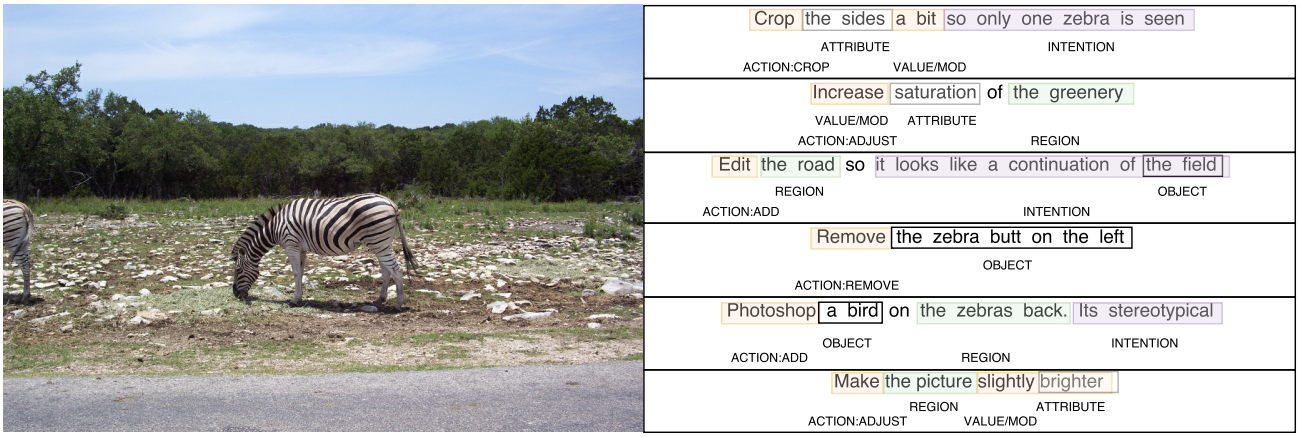

Figure 2: Example image from the corpus, with a few Image Edit Requests and their annotations.

color to the cobblestone sidewalk. It is lovely. An expert turker expressed a similar desire as *Adjust the brightness on the white tool to avoid making it look plain white.*

5. Annotation Framework

The ultimate goal of this research is to develop a tool that can carry out image edit instructions given in natural language; that is, a tool that can interpret these instructions in terms of editing functionality available in software like Photoshop, in a manner similar to how human expert photo editors interpret such requests today. To support this goal, we have devised an annotation framework that is an intermediary form between natural language and Photoshop commands. The actions in the intermediary form are fairly close to what is available in Photoshop, so carrying out these instructions is primarily a matter of interpreting the various properties and attributes – for example, which region of the photo is referred to by *my wedding dress* or *my dog's eyes*. This is by no means an easy task, but not one that concerns us here; we focus on identifying the actions, properties and attributes in the natural language requests.

The annotation framework is structured in three levels. In the first, an utterance is determined to be either an Image Edit Request (IER) or a comment. IERs receive two further levels of annotation, namely actions and entities: each IER is composed of one action and zero or more entities (an utterance which expresses multiple actions is segmented into separate IER). The terminology for the entities is borrowed from the work by Williams et al. (2015) which uses the entities to indicate the higher level intents. Comments are utterances that do not have an action.

To clarify the distinction between IERs and comments, we define an IER as an *actionable* item that can be interpreted up to some degree of certainty, albeit incompletely. A comment is an utterance that pertains to the image and may well include a request, but does not contain an action that could be completed by an image editing program given a particular image. For example, the utterance *This photo should have been taken with a Nikon camera* would be a comment as it is impossible to fulfill this request in an image editing program. Comments do not have any additional annotations.

Adjust (44.89%) Increase saturation a bit on the elephants.
Delete (13.70%) Remove the jacket hanging from the girl's side.
Crop (6.89%) Crop the photo to eliminate the space to the left and right of the elephants.
Add (6.85%) Insert a ball hitting the tennis racket.
Replace (2.47%) Please change the pamphlet she is holding into a dictionary.
Apply (1.44%) Add a Gaussian blur to the background.
Zoom (0.87%) Zoom in on the man.
Rotate (0.71%) The photo looks tilted. Rotate it clockwise so the lines are straight.
Transform (0.62%) Flip the photo horizontally.
Move (0.60%) Move the white framed picture to the blue wall.
Clone (0.33%) Use a cloning tool to blend grass to cover any patches of dirt on the ground.
Select (0.19%) Select the white dog.
Swap (0.14%) Please perform a face swap using the man in the yellow shirt and the man in the blue/black polo.
Undo (0.02%) If possible uncrop photo to allow more space to frame, rather than cut off the bike.
Merge (0.02%) Blend the grey smudges so they are the same color as the rest of the dirt.
Redo (0.01%) Redo all white traffic lines in street.
Other (0.01%) Resize photo to show large elephant and trainer.
Scroll (0.00%) No example in the corpus.

Figure 3: Action types, with frequency and examples.

IERs are annotated with at most one action, and its related entities, if any. An utterance that expresses multiple actions, such as, *Crop the left side of the photo and increase the saturation*, is marked as two IERs to accommodate the two separate actions. The action is usually an action verb which either explicitly or implicitly provides a mapping of a word or a phrase to a vocabulary that must be interpretable by most of the popular image editing programs. The frame-

Attribute Properties of the image to adjust, such as saturation, hue, or brightness.

Object An item to be inserted or deleted.

Region Location within the image where an action is being applied, such as top, entire image, or on a requested subject already in the photo.

Modifier/value Degree or direction of the change such as increase/decrease, modifiers of degree (examples: *a little*, *a lot*, *all*), directions (*to the left*), or numerical values (*25%*).

Intention User's reason or end goal for the change.

Figure 4: Types and descriptions of entities.

work supports 18 possible actions (Figure 3). The most common action represented in the data set was *adjust*, for such utterances as: *Make the image brighter*, *Increase the saturation*, and *Decrease the shadows*. Some actions are extremely rare in our corpus: this is because the framework was designed to also allow for interactive dialogue with an image editor. The framework therefore contains actions like *undo*, *redo*, *select*, *merge*, and *scroll*, which rarely come up in one-shot IERs of the type elicited here (such actions do show up when expert turkers give a complex, multi-stage request, for example: *Free select the sky, following building edge and around halo of the sun then increase contrast to reduce glare*).

The action provides first level of understanding of an IER. However, it is not sufficient to have the action alone if the user has provided additional details in an utterance. Actions support a list of five entities that complete the interpretation of an IER (Figure 4). Entities mark information about how the action is applied as an edit, such as detailing where a crop should occur or by how much the saturation level should be increased. Our framework supports the flexibility of an utterance having zero entities as well as an IER with multiple entities of the same type.

The various entity types are given in Figure 4. ATTRIBUTE holds information about what property of the image to adjust, and MODIFIER/VALUE provides information about the degree or direction of the change. For example, the IER *Increase the saturation* is annotated with the Action *adjust*, Attribute *saturation*, and Modifier *up*. We make a distinction between OBJECT, which is inserted or deleted, and REGION, which is the area where the action is to be applied. For example, in the IER *Add a dog*, the word *dog* is labeled as an Object as it is an entity to insert, but in *Brighten up the wave*, the word *wave* is regarded as a Region rather than an Object, as the person is interested in adjusting its brightness (and thus we have Action *adjust*, Attribute *brightness*, Region *wave*, and Modifier *up*). Finally, we included the entity INTENTION as users often provide information about their objective for performing the change. These intentions are by themselves not actionable but provide additional information: for example, the utterance *Paint the rocks unnatural but interesting colors like purple, green, yellow, and red to make the effect surreal* expresses a user's desire to make a *surreal* looking image, and is therefore annotated with Intention *to make the effect surreal*. A unique feature of this framework is that the same word can have multiple labels or one can be a sub-set of another. In the example *Increase the saturation*, the word *increase* is labeled both as an *adjust* action as well as a Modifier entity.

We thus propose this intermediary language scheme as a means to address the variability in vocabulary, structure, and ambiguity in IERs. To our knowledge, no other such published annotation scheme exists, and no one-to-one mapping of edit requests to executable actions in an image editing program permits for the described flexibility and range in natural language image edit utterances.

Feature	Krippendorff's alpha		
IER vs. comment	0.28	0.53	0.35
Action type	0.74	0.62	0.59
Attribute	0.47	0.41	0.38
Object	0.51	0.27	0.47
Region	0.55	0.35	0.43
Modifier/value	0.31	0.04	0.07
Intention	0.51	0.67	0.52

Table 1: Inter-rater reliability for 3 groups of annotators.

6. Analysis

To validate the annotation scheme we conducted an inter-rater reliability study on a sample of 600 utterances. Nine annotators received training, feedback on a set of 25 utterances, and support during the annotation process. The annotators were divided into groups of three, and each group annotated a different set of 200 utterances. Reliability was measured separately on actions and entities using Krippendorff's alpha (Table 1). For action type we used the nominal distance metric; IER versus comment was treated as a binary feature, and so were the five entity types, marking either presence or absence of that entity in a particular utterance. The highest agreement is reached on the action types; agreement on entities is somewhat lower but still well above chance. For some entities (value in particular) agreement borders on chance level, suggesting that annotation of entities in general and value in particular need to be better defined.

Crop and *add* actions were the most agreed upon actions between annotators. Requests to alter features of people in a photo presented the majority of discrepancies. For example, the utterance *You could have everyone smiling* was annotated as *add*, *replace*, and as a comment. Utterances with the phrases *clean up* and *edit* also presented differing annotations of *adjust*, *delete*, and *other*. Finally, utterances without an imperative verb were frequently annotated differently. The utterance, *The photo is too bright* was interpreted by annotators as a comment or as an IER with an *adjust* action. Future versions of the annotation manual will attempt to clarify these issues in order to ensure more consistent annotation.

For annotating at scale we used crowd-sourcing: 6000 elicited utterances were annotated, with only one turker annotating a given utterance. Only actions were marked by this group. Workers received an instructional video which was mandatory to watch before annotation. To determine

Annotators	Krippendorff's alpha		
3 trained	0.74	0.62	0.59
3 trained + 1 crowd	0.47	0.25	0.40

Table 2: Inter-rater reliability on identifying action type, for three groups of trained annotators and the same groups with crowd annotators.

the reliability of annotations completed by turkers, reliability between turkers and trained annotators was calculated on the 600 utterances completed by the trained annotators (Table 2). Reliability between turkers and trained annotators was much lower than between trained annotators, suggesting substantial differences between the groups; these may be due to training (turkers frequently selected the action *other*), or to the population from which the annotators were drawn.

7. Conclusion

This paper introduced a data-set for the domain of image editing using natural language. This is a currently unexplored task that combines language and vision. Our corpus comprises more than 9000 IERs collected via crowdsourcing. We built an annotation scheme for understanding such natural language image editing instructions and mapping them to actionable computer commands. Finally, we evaluated crowd-sourced annotation as a means of efficiently creating a sizable corpus at a reasonable cost. In future work, the corpus will be used for learning models that can automatically detect actions and entities, as well as the sequences of these components in an IER.

8. Acknowledgments

We would like to thank the many Mturkers who contributed image edits and annotations for this work. We would also like to thank Robert Dates, Takeshi Onishi, Arman Cohan, and Quan Tran for their annotations. The first author was also supported by a generous gift of Adobe Systems Incorporated to USC/ICT. The second author thanks the National GEM Consortium for fellowship funding and internship placement at Adobe. The last two authors were supported by the U.S. Army; statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

9. Bibliographical References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual Question Answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, Santiago, Chile.

de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., and Courville, A. (2016). GuessWhat?! visual object discovery through multi-modal dialogue. *arXiv preprint arXiv:1611.08481*.

Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., and Mitchell, M. (2016). Visual storytelling. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT)*, pages 1233–1239, San Diego, California, USA.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., and Fei-Fei, L. (2017). Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73.

Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. (2013). Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, Zurich, Switzerland.

Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., and Vanderwende, L. (2016). Generating natural questions about an image. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1802–1813, Berlin, Germany.

Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G., and Vanderwende, L. (2017). Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, Taipei, Taiwan.

Paetzel, M., Manuvinakurike, R., and DeVault, D. (2015). "So, which one is it?" The effect of alternative incremental architectures in a high-performance game-playing agent. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 77–86, Prague, Czech Republic.

Williams, J. D., Kamal, E., Ashour, M., Amr, H., Miller, J., and Zweig, G. (2015). Fast and easy language understanding for dialog systems with Microsoft Language Understanding Intelligent Service (LUIS). In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 159–161, Prague, Czech Republic.