Investigating the Viability of Semantic Compression Techniques Relying on Image-to-Text Transformations

Adit Chintamaneni, Rini Khandelwal, Kayla Le, Sitara Mitragotri, Jessica Kang

Mentors: Lara Arikan, Tsachy Weissman

Abstract

Data compression is a crucial technique for reducing the storage and transmission costs of data. As the amount of data that is consumed and produced continues to expand, it is essential to explore more efficient compression methodologies. The concept of semantics offers an interesting new approach to compression, enabled by recently developed technology. Concisely, we sought to discover whether the most important features of an image could be compressed into text, and if this text could be reconstructed by a decompressor into a new image with a high level of semantic closeness to the original image. The dataset of images that were compressed is composed of five common image categories: single person, group of people, single object, group of objects, and landscape. Each image was compressed through the following pipeline: image-to-text conversion, text compression and file size determination, file decompression and text recovery, and text-to-image conversion. This pipeline enables any image to be compressed into a few dozen bytes. When examining image-to-text compressors, we experimented with both human and artificial intelligence (AI) powered procedures. We selected the text-to-image model DALL-E 2 as our decompressor. We released multiple surveys to assess structural fidelity and semantic closeness between original images and reconstructed images. We also included compressed JPEGs and WebPs to benchmark performance. Human and AI reconstructions received lower structural fidelity scores than WebP and JPEG images. Individually, images reconstructed from human captions were perceived to have higher structural fidelity and semantic closeness to the original images than AI captions did. Participants' textual descriptions, of both human and AI reconstructions, had high semantic fidelity scores to their descriptions of the original images. This demonstrates that the proposed pipeline is a viable semantic compression mechanism.

Background

Images account for a large portion of all existing digital data. Conventional lossy image compression algorithms, such as discrete cosine transform, eliminate redundant data while preserving essential image features. In recent years, researchers have developed

semantic-assisted compression techniques in which important semantic features of an image are identified and preserved by the compression algorithm [1].

Modern advancements in both image-to-text and text-to-image transformations allow for the generation of text with high semantic fidelity to the original image and vice-versa.
Earlier this year, Salesforce released BLIP-2, a multimodal language model that can generate text descriptions of images [2]. In April of 2022, OpenAI introduced DALL E-2, a leading generative model that converts text descriptions into images [3]. We aim to investigate the viability of a semantic compression pipeline based on such transformations.

Methods

We identified five image categories that are semantically distinct: “Single Person”, “Group of People”, “Inanimate Object”, “Multiple Inanimate Objects”, and “Landscape”.

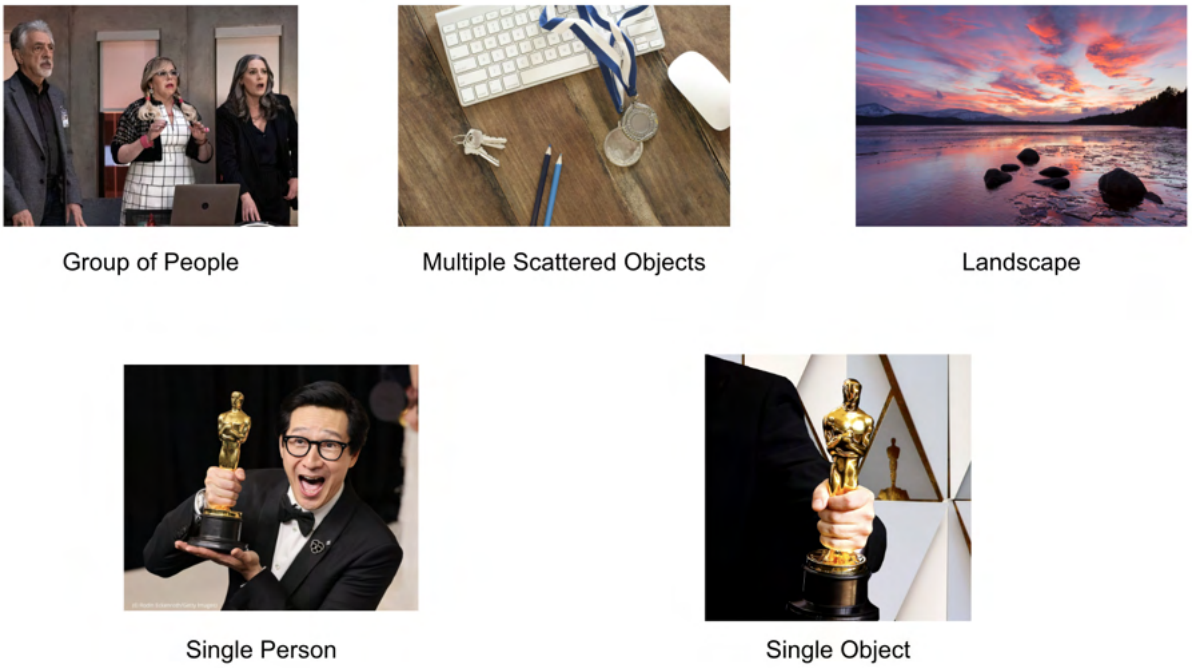


Figure 1: one set of original images.

We assembled 5 sets of 5 images (25 total images), where each set contained all of the identified image categories.

Human Compression

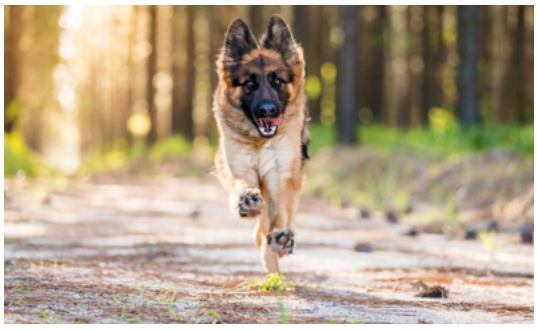
Which features of an image constitute its meaning? We sought to answer this initial question while developing our methodology for human-based Image-To-Text transformations. Through polling, referencing other works [4], and intuition, we found that the most important, universal features of an image include:

- major foreground objects
- all background objects
- colors, forms, and shapes of those objects
- dispositions of those objects
- the relationships between those objects, including actions and geometries of positions
- temporal context
- patterns in the image; repeating features of those objects

We compiled these features into a tight syntax for human captions to follow:

<major foreground objects>, <colors, forms, and shapes of those objects>, <dispositions of those objects>, <relationships between those objects, including actions and geometries of position>, <immediate context>, <temporal context>, <background context>

We manually captioned all 25 original images using this syntax.



a young, fluffy german shepherd with its tongue hanging out prancing merrily on a clear road in the afternoon with a forest in the distance.

Syntax Component	Words
major foreground objects	"german shepherd"
colors, forms, and shapes of those objects	"a young, fluffy"
dispositions of those objects	"prancing merrily"
relationships between those objects	"on"
immediate context	" on a clear road"
temporal context	"in the afternoon"
background context	"with a forest in the distance"

Figure 2: an image captioned using our tight syntax.

AI Compression

Current artificial intelligence-driven Image-To-Text transformation tools do not have the capability to capture the semantic schema of an image as well as humans can through our syntax. However, they offer a speed advantage, making them worthwhile to explore as a

potential component in our compression pipeline. We employed the Salesforce BLIP model to caption all of the collected images. By now, we had 25 human-generated captions and 25 AI-generated captions.

```
[ ] # Place to store captions
    captions = []

    # Loop over every image in your folder
    for filename in os.listdir():
        # This part taken with little modification from github
        raw_image = Image.open(filename).convert("RGB")
        # preprocess the image
        # vis_processors stores image transforms for "train" and "eval" (validation / testing / inference)
        image = vis_processors["eval"](raw_image).unsqueeze(0).to(device)
        # generate caption
        caption = model.generate({"image": image})
        # ['a large fountain spewing water into the air']
        print(filename + " was captioned :", caption)
        captions += caption

Multiple Scattered Objects.png was captioned : ['a desk with a cell phone, notebook and glasses']
Landscape.png was captioned : ['a mountain is reflected in the still water of a lake']
Single Person.png was captioned : ['a woman wearing a hat and smiling for the camera']
Group of People.png was captioned : ['a man and a woman standing next to each other']
Inanimate Object.png was captioned : ['a black dice sitting on top of a table']
```

Figure 3: captions generated by BLIP for an image set.

Decompression

We used DALL-E to reconstruct images from 25 human-generated captions. We refer to these as "Human Reconstructions". We followed the same procedure to reconstruct images from the 25 AI-generated captions, referring to these as "AI Reconstructions". We now had 75 images in total: 25 original images, 25 Human Reconstructions, and 25 AI Reconstructions.

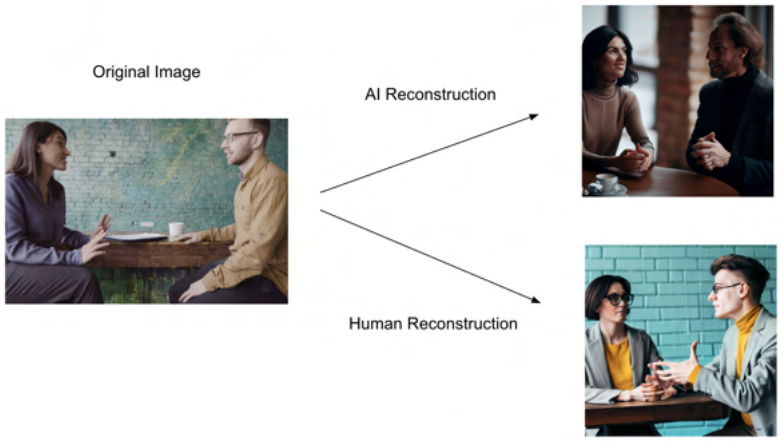


Figure 4: a visualization of the reconstructions of one of the original images.

Benchmarking Reconstructions

We benchmarked our reconstructed images against the images compressed via JPEG and WebP compression. We used a Jupyter Notebook to compress each of our 25 original images as JPEGs and WebPs at two qualities: 1 and 25. Here, quality = N, where N is a whole number between 1, the lowest quality, and 100, the highest quality. This provided us with 100 more

images. We also compressed each caption into a gzip file using the gzip compression algorithm [6]. The compressed human-generated captions had a mean file size of 154.8 bytes. The compressed AI-generated captions had a mean file size of 63.15 bytes.

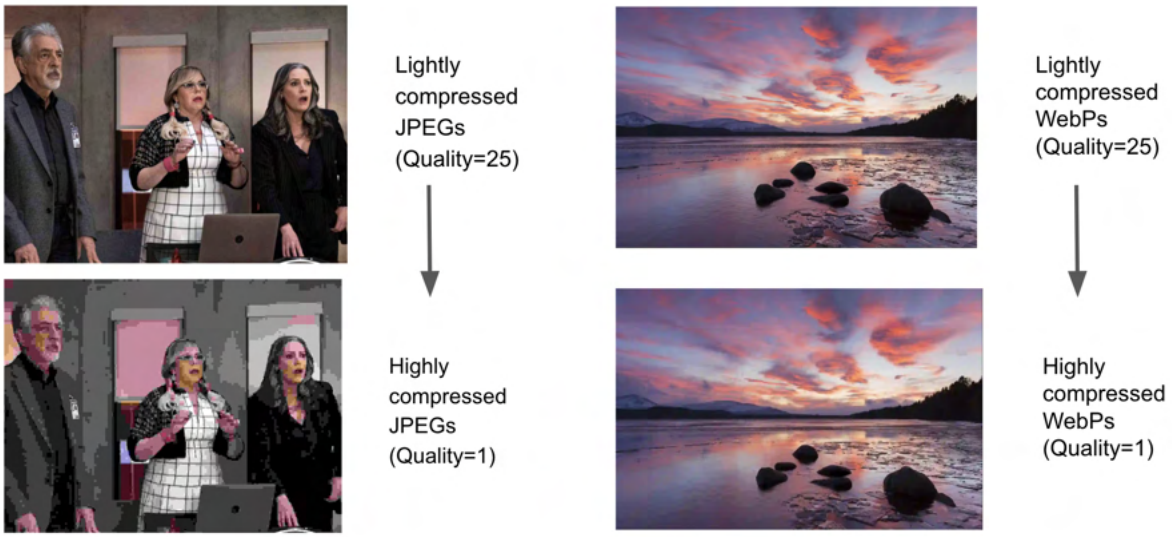


Figure 5: a visualization of the JPEG and WebP compressed files at Quality = 25 and Quality = 1 for two images from Figure 1.

By this stage, each original image had a corresponding AI Reconstruction, Human Reconstruction, highly compressed JPEG, highly compressed WebP, lightly compressed JPEG, and lightly compressed WebP. We refer to this as an image group. We issued three surveys for each of the 25 image groups.

- An **original_image survey** to capture what information people received from the original images
    - Questions: What object(s), element(s), or person(s) stand out to you most in this image? How do the “most important” object(s) you named above relate to each other? What three adjectives best describe this image? Please describe the image in one sentence.
- A **reconstructed_image survey** to record what information people received from the reconstructed images
    - human reconstructed_image section uses reconstructed human images
    - AI reconstructed_image section uses reconstructed AI images
    - Questions: same as original_image survey
- A **comparative_image survey** to compare the semantic closeness between original images and reconstructed images; this also compares original images with their corresponding JPEGs and WebPs
    - human comparative_image section compares original and human reconstructed images

- AI comparative_image section compares original and AI reconstructed images
- Questions (for each comparison): How similar are these images in terms of their effect on you, and what they mean to you? How similar are these images in terms of their content and appearance?

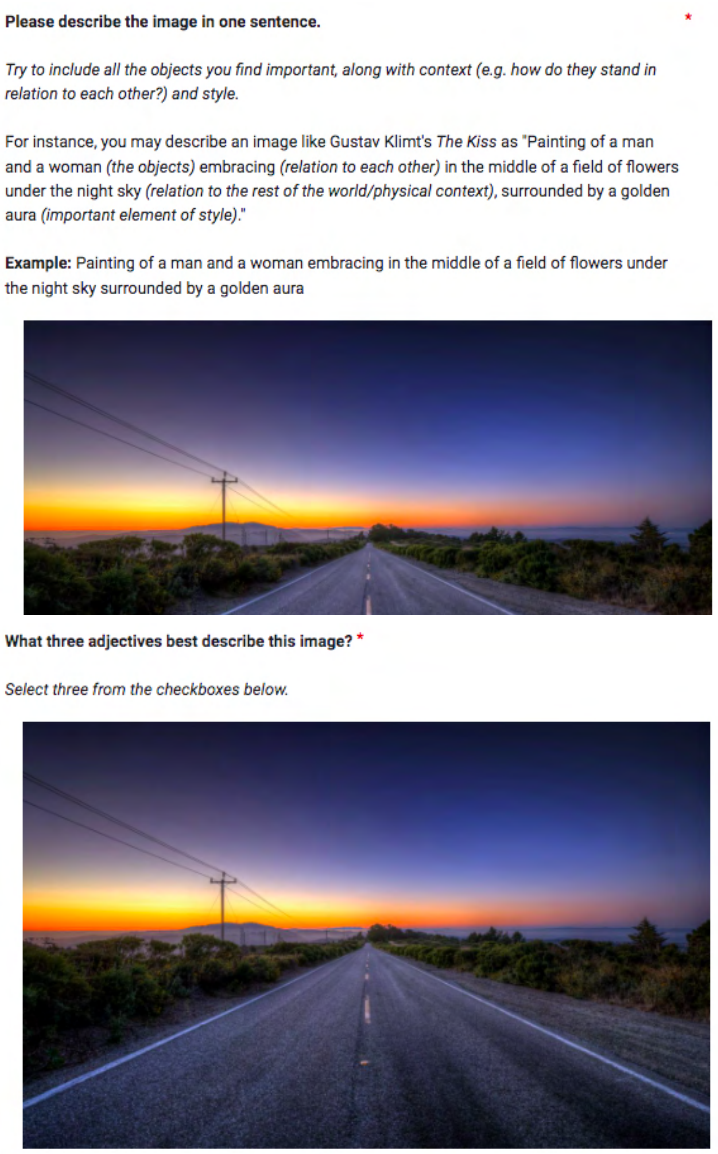


Figure 6: examples of qualitative survey questions from our original_image form and reconstructed_image forms. Questions remained constant, but the original_image forms contained only the original images while the reconstructed_image forms contained reconstructed AI/human images.

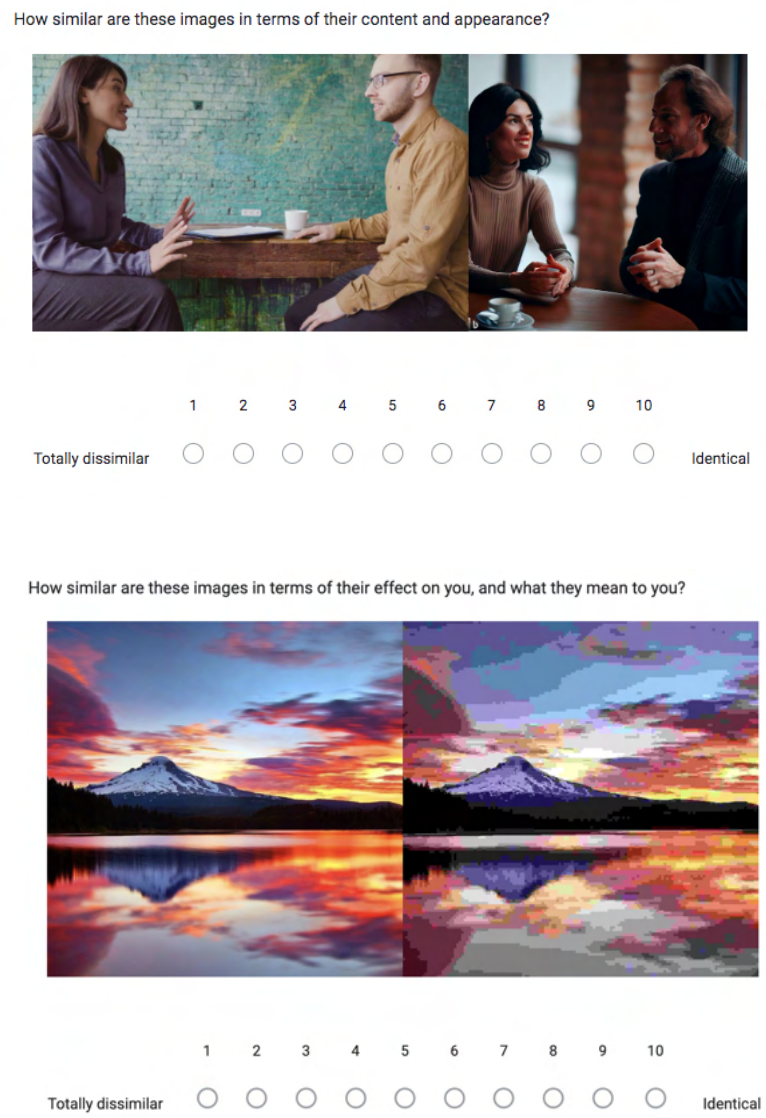


Figure 7: examples of quantitative survey questions from our comparative_image forms. The top question compares an original image (left) to its corresponding AI reconstructed image (right). The bottom question compares an original image (left) to its corresponding highly compressed JPEG (right). Both questions were asked for each image pair (ex. Original: AI Reconstructed, Original: Human Reconstructed, etc), for a total of 10 questions per section.

We collected 125 survey responses from people of various backgrounds. We split the data into two categories: quantitative and qualitative. Here, quantitative data consisted of all the comparative_image survey responses, and qualitative data consisted of all the original_image and reconstructed_image survey responses.

Analyzing Quantitative Data

Initially, the question “how similar are these images in terms of their effect on you, and what they mean to you” was meant to determine semantic closeness between images, and the question “how similar are these images in terms of their content and appearance” was mean to determine structural fidelity between images. However, after further polling, we found that most respondents interpreted both questions in the latter context. Thus, we rendered responses to the former question invalid. When analyzing this data, we examined the median ratings because the data was skewed, and used the mean ratings to further distinguish data with identical medians (see Results).

Analyzing Qualitative Data

The collected qualitative data was a better indicator of semantic closeness because the questions provided insights into the important semantic schema perceived by survey respondents. We focused on the following questions: “Please describe this image in one sentence.”, “What three adjectives best describe this image?”, and “What object(s), element(s) or person(s) stand out to you most in this image?” Based on these questions, we created six CSV datasets: AI Description vs. Original Description, Human Description vs. Original Description, AI Adjectives vs. Original Adjectives, Human Adjectives vs. Original Adjectives, AI Objects vs. Original Objects, and Human Objects vs. Original Objects. We used the similarity function in spaCy’s en_core_web_lg model to evaluate semantic closeness between descriptions, adjectives, and notable elements of original and reconstructed images. The function yields an output ranging from 0, indicating no semantic closeness, to 1, indicating that the texts are identical.

```
import spacy
import pandas as pd
import numpy as np
!python -m spacy download en_core_web_lg
```

```
[2] nlp = spacy.load("en_core_web_lg")
```

```
[3] df = pd.read_csv("AI vs Original.csv")
    descriptionSimilarities = []

    for index, row in df.iterrows():
      AI_Image_Description = row["reconstructed_AI"]
      Original_Image_Description = row["original"]
      currSimilarity = nlp(AI_Image_Description).similarity(nlp(Original_Image_Description))
      descriptionSimilarities.append(currSimilarity)

    avgSimilarity = np.mean(descriptionSimilarities)
    medianSimilarity = np.median(descriptionSimilarities)
    print(avgSimilarity)
    print(medianSimilarity)
```

```
0.8449301324790744
0.9014027986474126
```

Figure 8: calculating the text similarity between reconstructed AI image descriptions and original image descriptions using a Jupyter Notebook.

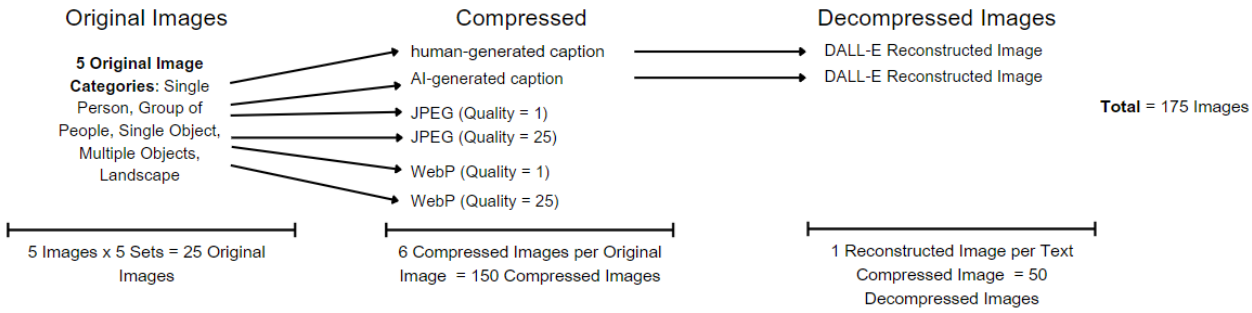


Figure 9: a diagram of our compression pipeline, as described throughout the Methods section.

Results

Quantitative

Respondents to the comparative AI survey, we found that people perceived the greatest similarity between the original images and lightly compressed JPEGs (median similarity rating of 10, mean of 9.7). The least similarity was found between the original images and the AI images (median similarity rating of 6, mean of 5.9). In the Comparative Human survey, people perceived the greatest similarity between original images and lightly compressed WebPs (median similarity rating of 10, mean of 10) and the least similarity between original images and human reconstructed images (median of 6.5, mean of 5.8).

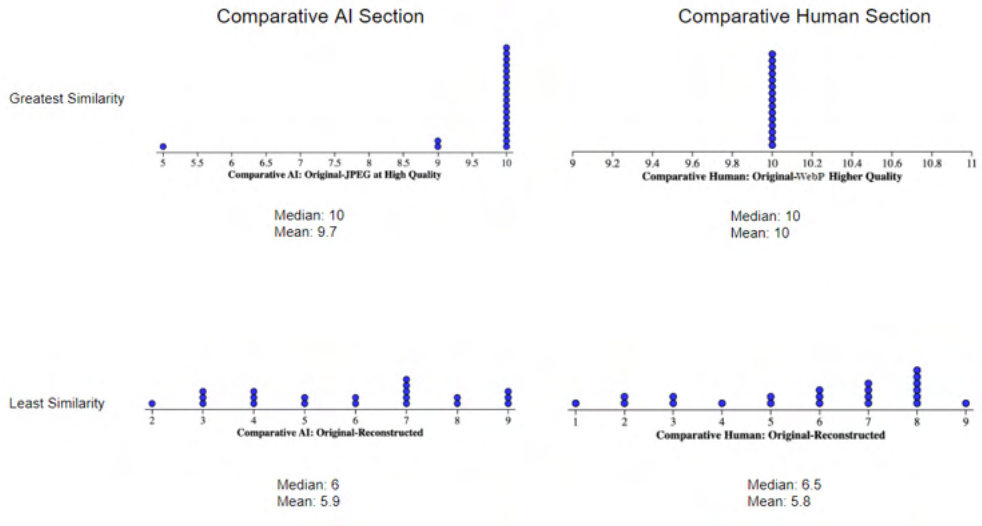


Figure 10: dot plots of structural fidelity between compressed images and original images from comparative form results (scaled down by a factor).

We analyzed quantitative data using direct comparisons of absolute values. Thus, if the absolute value of a number was greater than the absolute value of another number, the difference was considered to be significant. We did not use statistical significance tests to draw conclusions because the general human population was the determiner of similarities between images, so the probability of a response being a chance event can be ignored. Thus, every difference, as described above, is considered a significant difference on its own.

Qualitative

Respondents' descriptions of human reconstructed images and descriptions of original images had a mean text similarity of 0.87, while respondents' descriptions of AI reconstructed images and descriptions of the same original images had a mean text similarity of 0.84. Additionally, the three adjectives chosen to describe human reconstructed images and those chosen to describe original images had a mean text similarity of 0.98, while the three adjectives chosen to describe AI reconstructed images and those chosen to describe the same original images had a mean text similarity of 0.95. Lastly, answers to "What object(s), element(s) or person(s) stand out to you most in this image?" for human reconstructed images and corresponding original images had a mean text similarity of 0.80, while answers to the same question for AI reconstructed images and corresponding original images had a mean text similarity of 0.78. We provided a syntax for respondents to utilize when answering the first question, and a list of adjectives for respondents to select from when answering the second question. We did not design such structural rules for the third question; this may explain the lower similarity scores.

Conclusions

The quantitative results demonstrate that, on average, the human and AI reconstructions were less similar to the original image in terms of content and appearance (structural fidelity) than the WebP and JPEG compressed files. When paired against each other, human captions produced images with slightly greater structural fidelity and semantic closeness to the original than AI captions, as shown by both the median similarity scores and qualitative results. Furthermore, the qualitative results indicate that human textual descriptions of both types of reconstructions had high semantic fidelity with the human textual descriptions of the original images. Thus, we consider our pipeline, based on both human and AI captioning, to be a viable semantic compression mechanism. Indeed, for many images, semantic schema is more important than pixel-wise fidelity–the proposed pipeline can be integrated into storage and sharing of such images at unprecedented compression ratios.

Future Directions

Before conducting further surveys, some adjustments to our methods are advisable. Increased variation between image categories would test our pipeline's ability to capture semantic schema

across a wide range of image types. Within an image category, a greater variation between images would provide better data–for example, many of the “multiple scattered images” were images of typical desk objects. Furthermore, offering a clearer definition of semantics to survey respondents would enable them to answer quantitative questions about semantic closeness with greater accuracy. Although DALL-E is among the best Text-To-Image models, others are worth exploring. Stable Diffusion’s ControlNet, for example, can incorporate both image descriptions to preserve meaning and various maps to preserve structural fidelity. Although the compression ratios of this method are lower than those from our pipeline, they are still substantially better than those of existing algorithms.

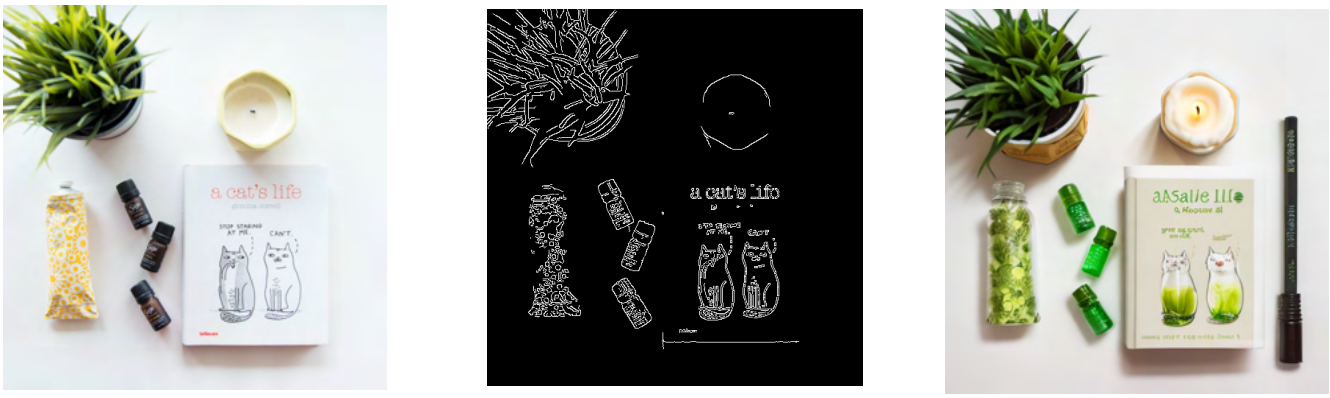


Figure 11: an example of a compression pipeline using ControlNet. Here, the original image (left) was captioned “a top view of a plant in a pot, a hexagonal candle, a tube of cream, bottles of essential oils, and a book titled "A Cat's Life" are arranged in a group, with the plant on the top left, the candle on the top right, the tube of cream on the bottom left, the bottles of essential oils in the middle, and the book on the bottom right, all on a well-lit, plain white surface.” An example of an edge map generated using cv2 is shown (middle). Together, both elements give the output shown on the right.

References

[1] Akbari, M., Liang, J., & Jingning H. (2019, Apr 18): DSSLIC: Deep Semantic Segmentation-based Layered Image Compression. arXiv.Org. https://arxiv.org/abs/1806.03348

[2] Li, J., Li, D., Xiong, C., & Hoi, S. (2022, Feb 15): BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv.Org. https://arxiv.org/abs/2201.12086

[3] Petsiuk, V., Siemenn A., Surbehera, S., ..., & Drori, I (2022, Nov 22): Human Evaluation of Text-to-Image Models on a Multi-Task Benchmark. arXiv.Org. https://arxiv.org/abs/2211.12112

[4] Perkins School for the Blind. (2023, Jul). How To Write Alt Text and Image Descriptions for Visually Impaired. perkins.Org.
https://www.perkins.org/resource/how-write-alt-text-and-image-descriptions-visually-impaired/

[5] Mahtab V., Pimpale G., Aldama J., & Truong P. (2019, Aug). Human-Based Image Compression; Using a Deterministic Computer Algorithm to Reconstruct Pre-Segmented Images. theinformaticists.com.
https://theinformaticists.com/2019/08/29/human-based-image-compression-using-a-deterministic-computer-algorithm-to-reconstruct-pre-segmented-images/