Live Photos: Mitigating the Impacts of Low-Quality Images in VQA

Lauren Olson
University of Delaware
lolson@udel.edu

Chandra Kambhamettu
University of Delaware
chandrak@udel.edu

Kathleen McCoy
University of Delaware
mccoy@udel.edu

Abstract

People with visual impairments use Visual Question Answering (VQA) systems to obtain information about their environment by submitting an image and question and receiving an answer. In this paper, we propose using live photos to capture additional information and reduce the impacts of image quality issues on accuracy of VQA predictions. We test live photos using a neural network trained on VizWiz data and show an improved ability to answer questions based on the live photo frames. Further, we present a method using entropy to extract the frame with the best prediction from the full set of live photo frames. Finally, we propose future directions for live photos as part of VQA systems.

1. Introduction

The objective of Visual Question Answering (VQA) is, given an image and a question, to determine an answer to the question based on the image's information. This task requires a combination of computer vision to process the image, natural language processing to analyze the text of the questions and answers, and machine learning to utilize the different aspects of the data to learn to answer questions based on images. VQA can then be applied to automate providing information for people with visual impairments.

Gurari et al. collected data from people with visual impairments to establish the VizWiz dataset and detailed the additional complexities that this data presents for machine learning systems [1]. Follow-up studies on VizWiz data by Chiu et al. showed that the ability to answer questions was impacted by quality issues such as blur, low lighting, rotation, and framing [2]. Prior research has proposed panorama stitching [3], omnidirectional imaging [4], and voice guidance in framing images [5] as systems to aid people with visual impairments in taking quality photos.

We introduce a new method using live photos for VQA to mitigate quality issues and improve answer accuracy. A live photo consists of a video file generated by the iPhone when the shutter button is pressed and a corresponding 'best' frame selected by the iPhone. The video file includes data from a couple of seconds before and after the picture was taken [6]. Live photo's ability to capture data over time provides more information about a scene and enables the photographer to move and make adjustments while taking the photo. The additional information provided by live photos can compensate for quality issues occurring over a short time. VQA systems may then examine the entire live photo to determine answers to questions and extract a single frame with the best answer.

Using live photos for VQA is distinct from existing image VQA tasks because spatio-temporal data in the live photos allows for analysis of multiple frames when low image quality exists in a single frame. Compared with uses of image sets [7], GIFS [8], and video for VQA [9, 10, 11, 12], our live photos are taken solely for the purpose of answering a question. As each live photo corresponds to a question that can be answered from a single image, we are able to generate answers across all frames of the live photo and use entropy to localize a single frame that provides the best answer.

Tests with live photos on the neural network architecture provided with the VizWiz data [1] show that there is more often at least one frame where the model can correctly predict the answer to a question. This paper's main contributions are as follows: (1) We propose live photos as a means to address quality issues and better answer VQA questions. (2) We show that a correct answer can be generated more often from looking at all live photo frames. (3) We propose entropy as a method to extract the frame with the correct answer from the full set of live photo frames.

2. Method

Our proposed method involves the following steps: i) collecting live photos data, ii) analyzing the full set of live photo frames to identify if at least one frame can answer the question correctly, and iii) using entropy to extract the frame most likely to contain the correct answer to the question.

This paper introduces a novel set of data collected using Live Photo on the iPhone 11. Live photos were selected as they use the same mechanism as capturing a photo but collect multiple frames of data over one to three seconds. This allows for the analysis of information over time. However, the live photos differ from a full-length video in that the goal is to capture a single moment rather than a sequence of events. Using live photos, we provide context around a scene and dilute the impacts of momentary quality issues that affect a single frame or a few frames. The set of live photos was taken in an indoor, household environment that mimics the environment present in VizWiz. The live photos use camera motion to generate different perspectives

(a) Frame Selected by iPhone Compared with Live Photo

Image Formats Being Compared	Percent of Datapoints
A Live Frame Better Than iPhone Frame	12.9%
iPhone Frame Better Than Any Live Frame	0%
Both Equal Accuracy	87.1%

(b) Frame Selected by Entropy Compared with Live Photo

Image Formats Being Compared	Percent of Datapoints
A Live Frame Better Than Entropy Frame	3.2%
Entropy Frame Better Than Any Live Frame	0%
Both Equal Accuracy	96.8%

(c) Frame Selected by Entropy Compared with Frame Selected by iPhone

Image Formats Being Compared	Percent of Datapoints
Entropy Frame Better Than iPhone Frame	9.7%
iPhone Frame Better Than Entropy Frame	0%
Both Equal Accuracy	90.3%

Table 1: Results on the data collected: the frame selected by iPhone, all the live photo frames, and the frame selected by entropy. Comparisons are based on if the answer is right or wrong for each frame and if at least one frame from the live photo was predicted correctly.

and capture quality issues such as blur and obstruction of the camera. This allows us to analyze how multiple frames of data may provide additional information for low-quality images. The full set of images analyzed contains 30 live photos with 80 to 90 frames captured per live photo.

To test the impact of live photos on a neural network's ability to perform visual question answering, each live photo is broken into a series of still frames. For each frame, the question "What is it?" is applied. This question is selected as it is the most commonly asked question in the VizWiz dataset [1]. The object in the scene is known before the live photo is taken and so this is used as the ground truth to generate a set of answers. The live photo frames are then passed into the neural network architecture provided with and trained on VizWiz data. Additionally, the iPhone selects a frame as the default image to show from the live photo. These frames are also passed through the neural network to provide a baseline of performance on a single image. Results are then generated by comparing answers on all live photo frames to the answer generated for the frame selected by iPhone.

Knowing that our question can theoretically be answered from only a single image, we aim to identify the single frame in the live photo that can best answer the question. For each frame of the live photo, entropy is used to generate a numeric value for the information contained in that image. A higher entropy corresponds to more information and a lower value corresponds to less information. Quality issues such as blur or obstruction over a large portion of the image reduce the overall amount of information. Therefore, the live photo frame with the highest entropy value is chosen as the candidate best frame for VQA.

3. Results

To generate results, each live photo is considered a single data point with a corresponding still iPhone frame and a corresponding frame selected by entropy. When looking at the answers predicted for the live photo, the answer is considered correct if it is predicted correctly in at least one frame. This allows results to then be directly compared with either a correct or incorrect prediction for the individual images. When comparing methods, accuracy is considered equal if both methods correctly predicted the answer or incorrectly predicted the answer.

The results presented in Table 1 show that live photos improve upon the baseline frame selected by the iPhone in 12.9 percent of the data points, with the other 87.1 percent of data points having answers predicted equally well in both formats. In addition, when entropy frame predictions were compared to live photos, the number of equally good predictions increased to 96.8 percent, with only 3.2 percent of live photos predicting answers better than the entropy frame. This suggests that using a high entropy value as a filter to select an image for VQA can nearly match performance when using multiple frames of the same scene. Further, the comparison of the default iPhone frame with the frame selected by entropy shows that in 9.7 percent of cases, an answer for the entropy frame was better able to be predicted, and it was never the case that an answer for the iPhone frame was better able to be predicted. Our results validate that entropy provides a method for selecting a single best frame to use in VQA.

4. Conclusions and Future Work

Using the additional frames of data from live photos was confirmed to provide supportive information, allowing the system to still make accurate predictions for some frames if a quality issue occurs in others. Future work includes extending this small study to collect data from people with visual impairments to develop more sophisticated uses of live photos to better answer questions from people with visual impairments.

References

[1] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham, “Vizwiz grand challenge: Answering visual questions from blind people,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari, “Assessing image quality issues for real-world problems,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* June 2020.

[3] Walter S. Lasecki, Yu Zhong, and Jeffrey P. Bigham, “Increasing the bandwidth of crowdsourced visual question answering to better support blind users,” in *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, New York, NY, USA, 2014, ASSETS ’14, p. 263–264, Association for Computing Machinery.

[4] Masakazu Iwamura, Naoki Hirabayashi, Zheng Cheng, Kazunori Minatani, and Koichi Kise, “Visphoto: Photography for people with visual impairment as postproduction of omni-directional camera image,” in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2020, CHI EA ’20, p. 1–9, Association for Computing Machinery.

[5] Nathan Davis, Bo Xie, and Danna Gurari, “Quality of images showing medication packaging from individuals with vision impairments: Implications for the design of visual question answering applications,” in *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, pp. e251, 2020.

[6] Apple Support, “Take and edit live photos,” Available at https://support.apple.com/en-us/HT207310, Jul 2020.

[7] Ankan Bansal, Yuting Zhang, and Rama Chellappa, “Visual question answering on image sets,” in *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., Cham, 2020, pp. 51–67, Springer International Publishing.

[8] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, “Tgif-qa: Toward spatio-temporal reasoning in visual question answering,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1359–1367.

[9] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg, “TVQA: Localized, compositional video question answering,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, Oct.-Nov. 2018, pp. 1369–1379, Association for Computational Linguistics.

[10] Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea, “LifeQA: A real-life dataset for video question answering,” in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4352–4358, European Language Resources Association.

[11] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler, “Movieqa: Understanding stories in movies through question-answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] H. L. Tan, M. C. Leong, Q. Xu, L. Li, F. Fang, Y. Cheng, N. Gauthier, Y. Sun, and J. H. Lim, “Task-oriented multi-modal question answering for collaborative applications,” in *2020 IEEE International Conference on Image Processing (ICIP),* 2020, pp. 1426–1430.