Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 139–143
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Automatic Myanmar Image Captioning using CNN and LSTM-Based Language Model

San Pa Pa Aung†, Win Pa Pa†, Tin Lay Nwe‡
†Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar
‡Visual Intelligence Department, Institute for Infocomm Research, Singapore
{sanpapaaung, winpapa}@ucsy.edu.mm, tlnma@i2r.a-star.edu.sg

Abstract

An image captioning system involves modules on computer vision as well as natural language processing. Computer vision module is for detecting salient objects or extracting features of images and Natural Language Processing (NLP) module is for generating correct syntactic and semantic image captions. Although many image caption datasets such as Flickr8k, Flickr30k and MSCOCO are publicly available, most of the datasets are captioned in English language. There is no image caption corpus for Myanmar language. Myanmar image caption corpus is manually built as part of the Flickr8k dataset in this current work. Furthermore, a generative merge model based on Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) is applied especially for Myanmar image captioning. Next, two conventional feature extraction models Visual Geometry Group (VGG) OxfordNet 16-layer and 19-layer are compared. The performance of this system is evaluated on Myanmar image caption corpus using BLEU scores and 10-fold cross validation.

Keywords: Convolutional Neural Network, Long-Short Term Memory, Visual Geometry Group.

1. Introduction

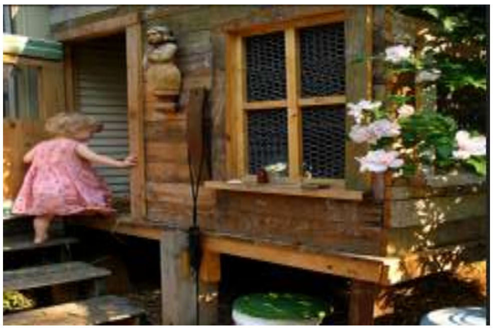
An image consists of several information such as the objects, attributes, scenes and activities. Humans are capable of generating captions for images with much less difficulty. However, automatic caption generation for a given image is a very challenging task for machine (Yang et al., 2018). Automatic image caption generation involves two tasks: 1) recognizing and understanding significant objects in an image and 2) describing the proper relationship between these objects. To perform these two tasks, image captioning uses a combination of two sub-networks, CNN for salient object detection in images and LSTM for understanding relationship objects and decoding into sentences (Shiru et al., 2017).

With the availability of extremely large numbers of images in internet nowadays, image captioning becomes more and more popular for retrieving images by Google search engines or newspaper companies (Huda et al., 2018). In addition, image captioning is useful for description of images for visually impaired persons, teaching concepts for children and social media network like Facebook and Twitter can directly generate captions from images (Zakir et al., 2018).

Myanmar language is morphologically complex and scarcity of annotated resources than English. Therefore, it is necessary to build a corpus which is large enough to get the accurate caption for Myanmar automatic image captioning. Example of an image and five different Myanmar captions can be seen at Figure 1.

In this paper, we used the combination of two sub-network: deep Convolutional Neural Network for image feature extraction and Long Short Term Memory for sentences generations. These two sub-networks communicate with each other in a merge layer to predict the next word of the sentences and then generate the caption for the specific image (Huda et al., 2018).

This paper is organized as follows: the related work is discussed in Section 2. Methodology is proposed in Section 3. In Section 4, experiments details and evaluation results are explained. Finally, the concluding remarks and future work are summarized in Section 5.



(1) ပန်းရောင် အင်္ကျီ နဲ့ ကလေးငယ် က အိမ် ထဲကို ဝင် နေတယ်.
(2) မိန်းကလေးငယ် က သစ်သား အိမ် ထဲကို ဝင် နေတယ်.
(3) ကလေးငယ် က အိမ် ထဲကို ဝင် နေတယ်.
(4) ကလေးငယ် က အိမ် ပေါ်ကို လှေကားထစ် မှ တက် နေတယ်.
(5) ပန်းရောင် အင်္ကျီ နဲ့ ကလေးငယ် က သစ်သားအိမ် ထဲကို ဝင် နေတယ်.

Figure 1 : Example of an image and its Myanmar descriptions.

2. Related Work

The restriction of image caption corpora for morphological complex language rather than English is an issue to get the accurate results.

The image caption generation is mainly split in retrievable-based approaches and constructive-based approaches. The first category is used in the earlier attempts to solve image captioning which has the problem as a retrieval task. A database is constructed based on image features extraction and caption generation for given images and then the most appropriate sentence is extracted (Jacob et al., 2015). This approach is not effective to describe novel captions and the caption generation is restricted to the features size of the images and the database size. Therefore, retrieval-based approach is not appropriate for today’s demand.

Recently, constructive-based approaches become popular due to recent progress in automatic image caption generation and neural machine translation. A constructed-based approach gradually constructs a novel caption for each image (Chetan and Vaishli, 2018 ; Yajurv et al., 2019) . The authors (Parth et al., 2017) used this approach that can be further divided into two phrases as deep convolutional neural network for encoding image attributes and Long Short Term Memory network for decoding to generate a syntactically correct caption.

The authors (Huda et al., 2018) implemented automatic image captioning in Arabic by using Deep Learning Technique. MSCOCO and Flickr8k dataset are used and Arabic image captions corpus is built using a professional English-Arabic translator and Google translator.

In this paper, we used constructive-based approaches and Myanmar images caption corpus is built so that the generated image captions are more accurate and relevant with each other. Furthermore, two different feature extraction models are compared in this paper.

3. Methodology

Figure 2 shows the Architecture of CNN-LSTM-based image captioning system. The architecture involves two main modules. The first one is image understanding module using CNN and the second one is text understanding module using LSTM. Each module is described in details in the following subsections.

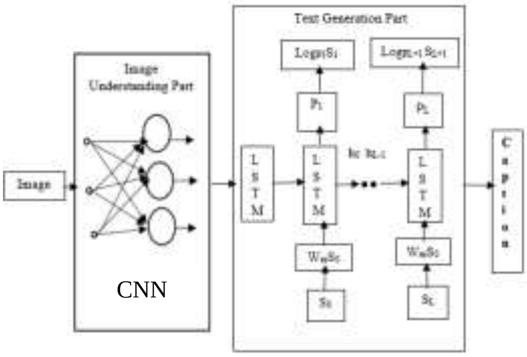


Figure 2 : Architecture of CNN-LSTM-based Image Captioning.

3.1 Convolutional Neural Network(CNN)

For image caption generation task, CNN is widely used because it has solved successfully for image annotation problems with high accuracy (Aditya et al., 2019). We have trained and tested two different models for feature extraction of images datasets. The two models have different capabilities in extracting features of images and the input image size of both models are 224× 224× 3 and the convolutional feature size of VGG is 4096.

VGG16: is a pre-trained model on ImageNet dataset based on Visual Geometry Group (VGG) OxfordNet 16-layer CNN (Rahul and Aayush, 2018 ; Lakshminarasimhan et al., 2018). The VGG16 neural network is used for image classification. Output of VGG16 is probability of individual classes that the classification system has to classify. We remove the last layer of the VGG16 and use the output from second last layer as feature parameters for each image. We extract 4096 parameters for each image, which are further processed by a Dense layer to produce a 256 element representation of an image (Micah et al., 2013).

VGG19: We also used a fully convolutional neural network based on Visual Geometry Group (VGG) OxfordNet 19-layer to extract features of each image. VGG16 and VGG19 networks have the total number of weight layers 16 and 19 respectively. VGG19 has 3 more convolutional layers than VGG16.

3.2 Long-Short Term Memory (LSTM)

LSTM can maintain information in memory for long periods of time and retrieve sequential information through time (Yang et al., 2018). The text understanding part produces words or phrases based on the word embedding vector of previous part. The language generation model is trained to predict each word in the caption after it has seen both image and all previous words. For any given sentence in Myanmar corpus we add two extra symbols for start word and stop word which designates the start and end of the sentence. Whenever stop word is found it halts generating caption and it denotes end of the sentence.

Sequence Processor is a word embedding layer to handle the input text and then followed by a Long Short-Term Memory (LSTM) recurrent neural network layer (Shiru et al., 2017). The proposed model is defined by the input sequences length (21 words) which are fed into an Embedding layer and then uses a mask to ignore padded values and followed by an LSTM layer with 256 memory units (Parth et al., 2017).

Both input models produced a 256 element vector and used regularization of 50% dropout to reduce over fitting during the training. In decoding, the model combined the vectors from both input models by using an addition operation and then fed to a Dense 256 neuron layer to make a softmax prediction over the whole output vocabulary for the next word in the sentence.

Loss function for both models are evaluated as,

$$L\ (I, S) = - \sum_{t=1}^{N} \log p_t(S_t) \qquad (1)$$

Where I is input image and S is generated sentence, N is the length of generated caption. p_t and S_t are probability and predict word at time t respectively. During the training process we have tried to reduce this loss function.

4. Experiments

4.1 Myanmar Image Captions Corpus Construction

The Flickr8k[1] dataset (Khumaisu et al., 2018 ; Micah et al., 2013) is applied in the first Myanmar Image Captioning task. It contains 8092 images and five annotated English captions for each image. Due to the limited time, we selected only 3k images of the Flickr8k dataset with five annotated Myanmar captions for each image. We constructed Myanmar image captions corpus in two different ways: 1) Automatic translation from English descriptions and 2) Direct image descriptions with Myanmar language.

[1] https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_Dataset.zip

4.1.1 Translation from English to Myanmar Captions without Images

Firstly, we translated the English image description of the Flickr8k dataset to Myanmar sentences without the image itself by using English to Myanmar Machine Translation. Attention based Neural Machine Translation model from English to Myanmar language (Yi et al., 2019), trained on UCSY Corpus that has 220k English Myanmar Parallel sentence, is applied in this stage. Due to the domain of the training data is general and influenced by News and conversations, the translation accuracy on 3k images of Flickr8k dataset is 13.93 multi-BLEU. Although the translation accuracy is low, the translated sentences help to reduce manual captioning time.

4.1.2 Direct Construction Myanmar Captions from Images

In this stage, we manually checked and corrected the translation of Myanmar captions by looking at the image and creating sentence descriptions correspond to the pictures. We have written our own natural language expressions based on our perception of the image without utilizing English descriptions. The total Myanmar captions for 3k images are 15,000 sentences with a vocabulary size of 3,138. The length of longest sentence is 21 words. The experiment was set as 2500 images for training, 300 images for validation and 200 images for testing.

4.2 Experiments Details

We conducted experiments to observe the different components of the image captioning system, and we evaluated the experiment results. The two different models are trained on K80 GPU machine using Keras API library with TensorFlow backend that are used for creating and training deep neural networks. The large amount of training data are given, the models fit the 10 epoch. After the 4th epoch both models stabilized and save the loss for each fold. The smallest value of loss on the training dataset is 2.097 and the validation loss on the development dataset is 2.513 in 10 folds cross validation setting using VGG16 with LSTM. And, the smallest loss on the training dataset is 2.114 and the validation loss on the development dataset is 2.513 when we used VGG19 with LSTM. As we can see the smallest validation loss for both models with 10 folds cross validation settings are the same. Figure 3 and 4 show the variation of training and validation loss using two different models.

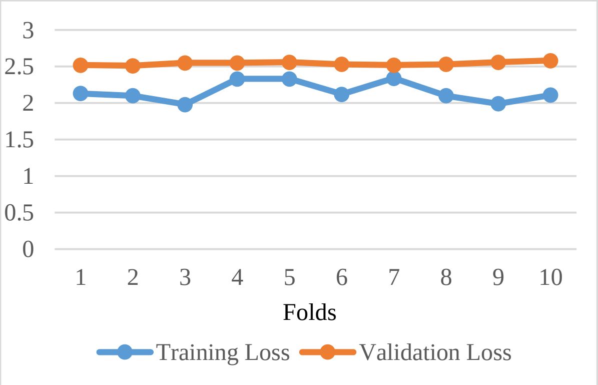


Figure 3 : Variation of training and validation loss in 10 folds using VGG16 with LSTM.

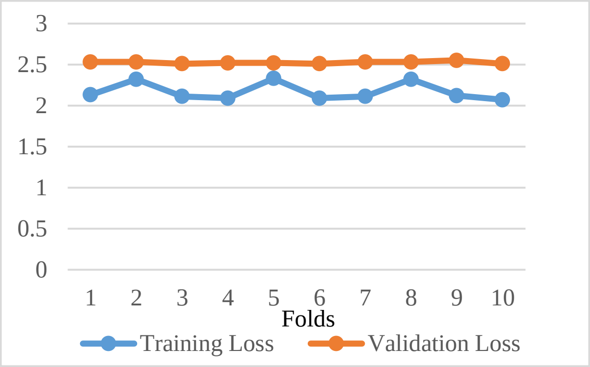


Figure 4 : Variation of training and validation loss in 10 folds using VGG19 with LSTM.

4.3 Evaluation Metric

BLEU (Bilingual Evaluation Understudy) is a metric that is used to compute the quality of machine translated texts (Zakir et al., 2018). The generated captions for each model are evaluated using BLEU to get the quality of machine translated texts (Shiru et al., 2017). BLEU score values range from 0 to 1 and higher values indicate the best score between the reference caption and machine generated captions. BLEU evaluates the modified precision of n-grams (Parth et al., 2017). In our experiment, BLEU scores are calculated as in equation 2:

$$\text{BLEU}=\min\left(1,\frac{output\_length}{reference\_length}\right)\left(\prod_{i=1}^{4} precision_i\right)^{1/4} \quad (2)$$

Where output_length is the output caption length and reference_length is the reference caption length.

4.4 10-Fold Cross Validation

This paper used 10-fold cross validation to compute predictive models by partitioning the original dataset into a training dataset to train the model and a test dataset to evaluate performance. In 10-fold cross validation, the original dataset is randomly partitioned into 10 equal subsets. Among these 10 subsets, one set is used as the validation data for testing the model, and the rest of 9 sets are used for training data. The cross-validation process repeated 10 times (the fold), with each of the subsets used exactly once for the validation data. We compute the average accuracy over all the folds to produce a single estimation. In table 1 and table 2, we show the BLEU scores of each fold with different testing datasets. Figure 5 shows the comparison of VGG16 with LSTM and VGG19 with LSTM average BLEU scores.

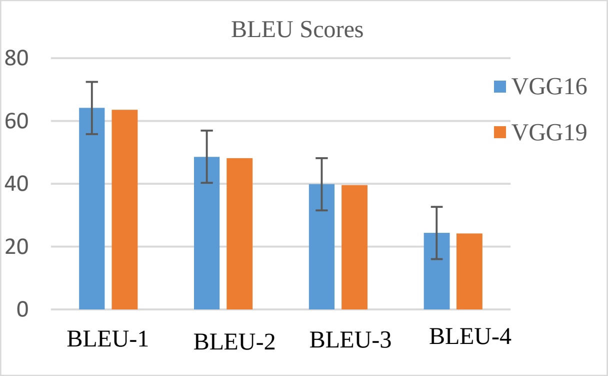


Figure 5 : Comparison of VGG16 and VGG19.

Training Times	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Fold 1	61.5	47.6	40.3	25.9
Fold 2	62.4	46.4	37.1	22.2
Fold 3	64.8	49.3	40.5	23.8
Fold 4	65.6	49.8	41.2	26.0
Fold 5	66.2	50.8	41.6	25.2
Fold 6	64.3	48.7	40.5	25.6
Fold 7	62.5	46.4	36.9	21.4
Fold 8	63.1	46.8	37.9	22.8
Fold 9	64.9	50.2	42.4	26.8
Fold 10	66.1	49.8	40.2	24.1
Total	641.4	485.8	398.6	243.8
Average	**64.14**	**48.58**	**39.86**	**24.38**

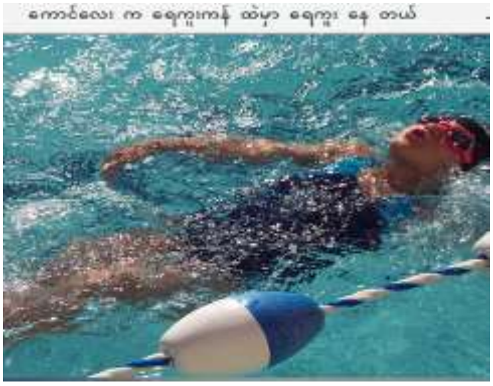
Table 1 : 10-Fold Cross Validation for VGG16

Training Times	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Fold 1	65.1	50.6	42.7	28.0
Fold 2	60.6	44.7	36.4	20.9
Fold 3	58.8	44.1	36.3	21.2
Fold 4	65.6	49.3	39.2	22.7
Fold 5	67.0	51.1	41.5	24.9
Fold 6	65.9	50.0	41.0	25.4
Fold 7	54.6	40.7	34.2	21.3
Fold 8	65.5	49.2	40.1	24.6
Fold 9	65.4	51.4	43.6	28.2
Fold 10	66.6	50.1	40.5	24.6
Total	635.1	481.2	395.5	241.8
Average	**63.51**	**48.12**	**39.55**	**24.18**

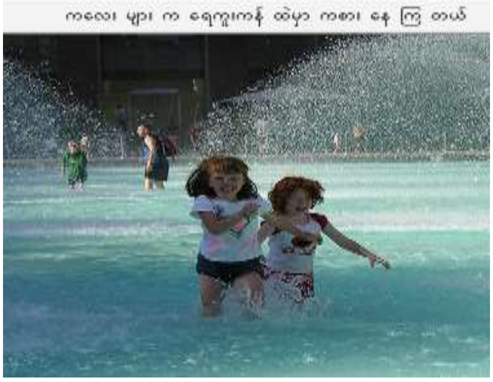
Table2 : 10-Fold Cross Validation for VGG19

4.5 Experiments Results

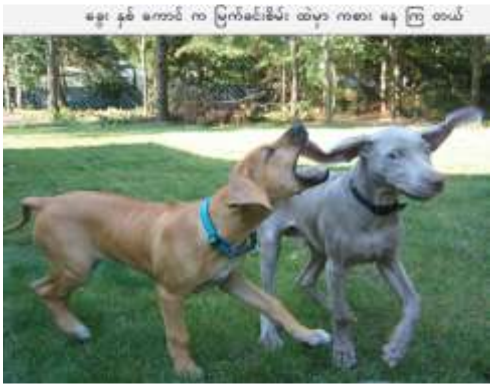
The captions generated from VGG16 and VGG19 are approximately similar and do not provide any qualitative difference. Therefore, in this section, we mainly focused on the results generated from VGG16 with LSTM. In figure 6(a), the model accurately generated the major features in the image such as “ကောင်လေး က ရေကူးကန် ထဲမှာ ရေကူး နေတယ်” (“The boy is swimming in the swimming pool”) and the relationship between these features of image also describes accurately. In figure 6(b), the generated caption: “ကလေး များ က ရေကူးကန် ထဲမှာ ကစား နေ ကြ တယ်” (“Children are playing in the swimming pool”) and in figure 6(c), the generated caption: “ခွေး နှစ် ကောင် က မြက်ခင်းစိမ်း ထဲမှာ ကစား နေ ကြ တယ်” (“Two dogs are playing in the green grass”). If we look at figure 6(b) and 6(c), the significant features of the images are captured accurately and grammatically correct. Nonetheless, we can see at figure 6(d) for random image, the model captures the major feature which is လူ တစ်ယောက် (a person) and ထိုင် နေတယ် (sitting) but fails to depict the minor features and incorrectly captures like နံရံ (wall) it is actually ခုံတန်းရှည် (bench). Finally, it is the limitations of our model and we would like to highlight the necessity for future work regarding the model. We are confident that larger datasets can be used to resolve these issues, and our models can accurately generate the relationship between images and its captions even for random images. All of the figures 6(a), 6(b), 6(c) and 6(d) are captioned automatically with Myanmar Language without any human interference.



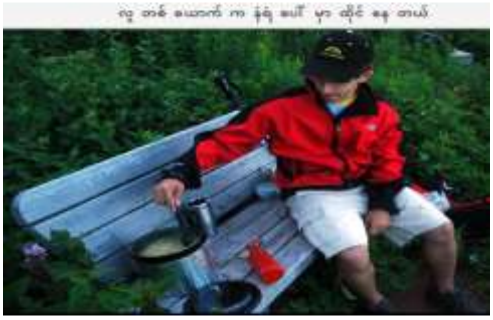
(a) In English: The boy is swimming in the swimming pool



(b) In English: Children are playing in the swimming pool



(c) In English: Two dogs are playing in the green grass



(d) In English: A person is sitting on the wall

Figure 6 : Example of Myanmar image captioning results (a,b,c,d).

5. Conclusion

We created the first corpus of image captioning for Myanmar language, and manually checked and built the descriptions in detail to match captions and images. Convolutional Neural Network based on Visual Geometry Group (VGG) OxfordNet CNN and single hidden layer LSTM model were applied for Myanmar automatic image caption generation in this work. The experimental results showed that applying CNN and LSTM based image captioning trained on our corpus can give acceptable performance.

This tiny corpus will help building large corpora for Myanmar Image Captioning. Moreover, the other image feature extraction models of CNN will be applied for future research.

6. Bibliographical References

Aditya, A. N., Anditya, A. and Suyanto, (2019). “Generating Image Description on Indonesian Language using Convolutional Neural Network and Gated Recurrent Unit”, 7th International Conference on Information and Communication Technology (ICoICT).

Chetan, A. and Vaishli, J. (2018). “Image Caption Generation using Deep Learning Technique”, Fourth International Conference on Computing Communication Control andd Automation (ICCUBEA).

Huda A. Al-muzaini, Tasniem N. and Hafida B. (2018) “Automatic Arabic Image Captioning using RNN-LSTM-Based Language Model and CNN”, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 9, No.6.

Jacob, D., Saurabh, G. and Ross, G. (2015). “Exploring Nearest Neighbor Approaches for Image Captioning”, arXiv: 1505.04467.

Khumaisu, N., Johanes, E., Sakriani, S., Mirna, A. and Satoshi, N. (2018). “Corpus Construction and Semantic Analysis of Indonesian Image Description”, The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages, Gurugram, India.

Lakshminarasimhan, S. , Dinesh, S. and Amutha, A. (2018). “Image Captioning - A Deep Learning Approach”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13.

Micah, H., Peter, Y. and Julia, H. (2013) “Framing image description as a ranking task: Data, models and evaluation metrics”, Journal of Artificial Intelligence Research, Vol. 47, pp. 853-899, May.

Parth, S., Vishvajit, B. and Supriya, P. (2017). “Image Captioning using Deep Neural Architectures”, International Conference on Innovations in information Embedded and Communication Systems (ICIIECS).

Rahul, S. and Aayush, S. (2018). “Image Captioning using Deep Neural networks”.

Shiru, Q., Yuling, X and Songtao, D. (2017). “Visual Attention Based on Long-Short Term Memory Model for Image Caption Generation”, 29th Chinese Control and Decision Conference (CCDC).

Sreela, S. R. and Sumam, M. I. (2018). “AIDGenS: An Automatic Image Description System using Residual Neural Network”, International Conference on Data Science and Engineering (ICDSE).

Shuang, L., Liang, B. and Yanming, (2018). "Reference Based on Adaptive Attention Mechanism for Image Captioning", 2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM).

Xinwei, H., Yang, Y. and Baoguang S. (2018). “VD-SAN: Visual-Densely Semantic Attention Network for Image Caption Generation”, Neurocomputing.

Yajurv, B., Aman, B., Deepanshu, R. and Himanshu, M. (2019). “Image Captioning using Google’s Inception-resnetv2 and Recurrent Neural Network”,IEEE.

Yang, F., Jungang, X., Yingfei,S. and Ben, H. (2018). “Long-term Recurrent Merge Network Model for Image Captioning”, IEEE 30th International Conference on Tools with Artificial Intelligence.

Yi, M. S. S., Win, P. P. and Khin, M. S. (2019). “UCSYNLP-Lab Machine Translation Systems for WAT 2019”, Proceedings of the 6th Workshop on Asian Translation.

Zakir, H., Ferdous S., and Mohd F. S. (2018). “A Comprehensive Survey of Deep Learning for Image Captioning”, ACM Computing Surveys.