Image Captioning System

Gokul Raj G[1], Hasan Farook N[2], Karthikeyan P[3], Rajaramanan V[4]
[1, 2, 3] Dept of Computer science and Engineering
[4]Professor,Dept of Computer science and Engineering
[1, 2, 3, 4] Adhiyamaan college of Engineering, Hosur.

Abstract- *Image captioning has been one of the major problem statements over the past few years in the machine learning world. The major problem of the image captioning system is that the object to be identified is unknown i.e. there may be many objects in the picture and the caption may concentrate on any of the objects. This perspective differs from user to user. There is a need for a system that would learn from the user's caption. Moreover, the images that users upload cannot be used for training since this would come as a privacy issue. Hence, an efficient image captioning system which would learn without sending the photos out of the devices is necessary. A general image captioning system that would identify the objects in the picture and generate a suitable meaningful caption for that. Personalisation of those captions by learning from the captions given by the specific user. Improving privacy by devising a method to learn from images by not sending them out of the device.*

Keywords- Image Captioning, Federated Learning, Machine Learning and Deep Learning.

I. INTRODUCTION

The major problem with image captioning is the availability of dataset(s). One possible solution is to train on user images which they would upload in social media along with captions. However, it would be a privacy intrusion for the user. Hence, we have to train on user data without their privacy getting disturbed.

This can be solved by using Federated learning, a distributed machine learning algorithm where the models would be trained on the user device. In addition, different people concentrate on different parts of the image. Hence, this federated learning system can also be used to personalize the generated caption according to user needs.

This project aims at using federated learning to improve the performance of image captioning systems by training on the user images without sending it anywhere outside the device. In addition, it also aims to provide personalized caption suggestions to the user. There are three objectives to accomplish:

A general image captioning system that would identify the objects in the picture and generate a suitable meaningful caption for that. This model would be present in both the central server and the user device. Personalization of those captions by learning from the captions given by the specific user. Improving privacy by devising a method to learn from images by not sending them out of the device.

This system would provide a federated architecture to the image captioning problem which may be re-used for other image processing problems.

This system can be used to train image description systems based on user data. In later stages, the same model can be used for other purposes by using transfer learning.

The image description systems can be used by blind people to know what is in front of them.

The federated learning architecture can be used to provide personalized captions which would increase user engagement. In this project, we propose a fast personalization system using Federated Reinforcement Learning for image captioning systems. This is achieved by the fusion of two research papers – one for fast personalization and another one for generating image descriptions

This project provides privacy by not sending the user image data out of the user device. The captions are personalized according to the user.

As with every personalization algorithm, it suffers from a cold start. Hence, a few instances of image-caption data is initially needed to provide personalized captions. This project can be developed further by using the trained model in tasks like image querying, object identification, etc. The project suffers from the threat of utilizing the resources of the user. In order to train the model on user devices, their computing resources should be utilized which is not possible all the time.

II. RELATED WORKS

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects

computer vision and natural language processing. Earlier methods first generate annotations (i.e., nouns and adjectives) from images (Sermanet et al., 2013; Russakovsky et al., 2015), then generate a sentence from the annotations (Gupta and Mannem,). Donahue et al. (Donahue et al.,) developed a recurrent convolutional architecture suitable for large-scale visual learning, and demonstrated the value of the models on three different tasks: video recognition, image description and video description.

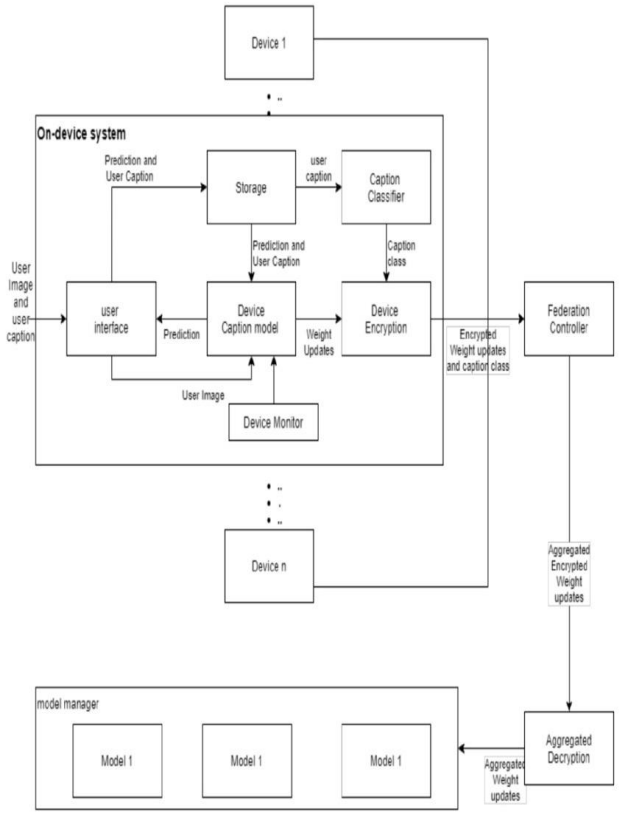
In these models, long-term dependencies are incorporated into the network state updates and are end-to-end trainable. The limitation is the difficulty of understanding the intermediate result. The LRCN method is further developed to text generation from videos (Venugopalan et al.,). Instead of one architecture for three tasks in LRCN, Vinyals et al. (Vinyals et al.,) proposed a neural image caption (NIC) model only for the image caption generation. Combining the GoogLeNet and single layer of LSTM, this model is trained to maximize the likelihood of the target description sentence given the training images. The performance of the model is evaluated qualitatively and quantitatively. This method was ranked first in the MS COCO Captioning Challenge (2015) in which the result was judged by humans. Comparing LRCN with NIC, we find three differences that may indicate the performance differences. First, NIC uses GoogLeNet while LRCN uses VGGNet. Second, NIC inputs visual feature only into the first unit of LSTM while LRCN inputs the visual feature into every LSTM unit. Third, NIC has simpler RNN architecture (single layer LSTM) than LRCN (two factored LSTM layers). We verified that the mathematical models of LRCN and NIC are exactly the same for image captioning. The performance difference lies in the implementation and LRCN has to trade off between simplicity and generality, as it is designed for three different tasks. Instead of end-to-end learning, Fang et al. (Fang et al.,) presented a visual concepts based method. First, they used multiple instance learning to train visual detectors of words that commonly occur in captions such as nouns, verbs, and adjectives. Then, they trained a language model with a set of over 400,000 image descriptions to capture the statistics of word usage. Finally, they re-ranked caption candidates using sentence-level features and a deep multi-modal similarity model. Their captions have equal or better quality 34% of the time than those written by human beings. The limitation of the method is that it has more human controlled parameters which make the system less re-producible. We believe the web application captionbot (Microsoft,) is based on this method.Then, the traffic light controller presented a green light in the direction of an emergency vehicle until It exited the intersection. An RFID- and GPS-based automatic lane clearance protocol for ambulances was proposed in [9]. The objective of this protocol was to minimise the travel times of ambulances by clearing lanes prior to an ambulance reaching an intersection. In [10], a cellular automata model was established for intersections to analyse the different characteristics of vehicles in two different environments (i.e., a non-vehicle networking environment and Internet of Vehicles (IoV) environment). This model considers the speed effects of leading vehicles, influence of brake lights, and many other rules to accurately reflect the operation of traffic flows at an intersection. A comparison of traffic parameters, such as vehicle speed, traffic flow, and average travel time, was conducted via numerical Simulations for the two environments. The results revealed that in an IoV environment, a vehicle's queue length is shorter, congestion dissipates faster, and traffic runs more smoothly.

III. ARCHITECTURE OF THE PROPOSED SYSTEM

Figure, which shows the overall architecture diagram, indicates both the phases of the system. The first one is the On-device system which is indicative of the federated model's instance present across multiple devices (devices 1...n). This system accepts an image and suggests words as and when the user types the caption. This suggestion is obtained from the device caption model to which the image is sent initially.

The device monitor manages and regulates the functioning of the on device model and is responsible for deciding when the weight updates are to be sent to the centralised server. Unlike conventional systems where the image is sent to the server, here, the weight updates are encrypted and sent from each individual device to the server along with the caption class.

The second phase (centralised server) begins with the decryption of the aggregated encrypted weight updates. Data from the individual device can't be obtained separately once it has been aggregated with weight updates from other similar devices. Post decryption the centralized model is updated using the obtained aggregated weight updates. The new model is federated and is sent to the individual devices at a suitable time.



FEDERATION USING PYSYFT:

The PySyft library is used to federate the trained model across various devices. The PySyft library helps to create worker nodes which are listed through particular ports and accept requests from other devices as well. By allocating workers in the client devices it is possible to make that listen for models and updates. The devices will constantly collect the images and captions that are made on the device. Then, it will train the model received from the central device. Then, it will train on the data collected in the device locally.

After that, the weights are sent back to the central model. The same process is repeated for multiple devices. Then, the model is sent back to the central device. There, the weights are aggregated and are fed into the central model. Here, the devices will also send a personalization id which is used to identify the class of the user. This id is used to determine which central model needs to be fed with the incoming data.

In PyTorch, the model is trained using a loss function and an optimizer. A loss function accepts the predicted value of the model and the target value as its parameter in order to calculate the loss of the particular iteration. After this, the backward() function of the loss variable is called in order to calculate the gradients of the trainable parameters of the model. Then, the step() function of the optimizer updates the weights. This process is repeated for each iteration of every epoch.

IV. CONCLUSION

A general image captioning system was constructed using transfer learning from InceptionV3 and GloVe word embedding for image and text input respectively, which would identify objects from the images and generate meaningful caption for that. The model is such that it takes image and current sequence as the input data to generate the next word. A federated architecture for the same is proposed, in order to train the model locally in the user devices. Also, these captions are personalized by learning from the captions given by the specific user. The system functions by having the working federated model across all devices, from which, encrypted aggregated data was obtained at the centralised server. The centralised model gets updated frequently with the data, but the user's data doesn't leave their devices, only the weight updates are sent. Privacy is improved by learning from images by not sending them out of the device.

REFERENCES

[1] Satanjeev Banerjee and Alon Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments", In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72, Ann Arbor, Michigan, June 2005, Association for Computational Linguistics.

[2] D. Cao, S. Chang, Z. Lin, G. Liu, and D. Sun, "Understanding distributed poisoning attack in federated learning", In 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), pp. 233–239, 2019.

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil, "Universal sentence encoder", CoRR, vol. abs/1803.11175, 2018.

[4] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions", In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[5] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith, "Federated learning: Challenges, methods, and future directions", 2019.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space", 2013.

[7] C. Nadiger, A. Kumar, and S. Abdelhak, "Federated reinforcement learning for fast personalization", In 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pp. 123–127, 2019.

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning, "Glove: Global vectors for word representation.", In EMNLP, volume 14, pp. 1532–1543, 2014.

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, "Rethinking the inception architecture for computer vision", CoRR, vol. abs/1512.00567, 2015.

[10] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensusbased image description evaluation", In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575, 2015.

[11] Xiaoling Xia, Cui Xu, and Bing Nan, "Inception-v3 for flower classification", In 2017 2nd International Conference on Image, Vision and Computing (ICIVC), pp. 783–787, 2017.

[12] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong, "Federated machine learning: Concept and applications", ACM Trans. Intell. Syst. Technol., vol. 10, num. 2, January 2019.

[13] J. Zhang, J. Chen, D. Wu, B. Chen, and S. Yu, "Poisoning attack in federated learning using generative adversarial nets", In 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (Trust-Com/BigDataSE), pp. 374–380, 2019.