DES2/AIS2-2
Invited

Challenges of Integrating Vision and Language

Yoshitaka Ushiku[1,2]

contact@yoshitakaushiku.net
1OMRON SINIC X Corp., Nagase Hongo Bldg. 3F, 5-24-5 Hongo, Bunkyo, Tokyo, Japan
2Ridge-i Inc., Otemachi Bldg. 438, 1-6-1 Otemachi, Chiyoda, Tokyo, Japan
Keywords: Deep Learning, Vision and Language, Computer Vision, Natural Language Processing, Encoder-Decoder

ABSTRACT

The benefits of deep learning are not limited to advanced recognition and generation of data in different modalities, such as images, acoustic signals. As a result of the fact that they are now implemented using commoditized tools based on deep learning, it has become possible to import approaches to understanding other modal data quickly. As a result of the fact that they are now implemented using commoditized tools based on deep learning, it has become possible to import approaches to understanding other modal data quickly. Thus, research on multimodal machine learning that integrates and converts multiple modalities has been dramatically advanced. Thus, research on multimodal machine learning that integrates and transforms multiple modalities has been significantly advanced. This presentation will introduce such multimodal machine learning, mainly focusing on the fusion of images. It has been called Vision and Language, where many researchers from multiple fields such as machine learning, computer vision, and natural language processing enjoy challenging problems. This paper will introduce major tasks in Vision and Language.

1 Introduction

In 2021, it will be exactly ten years since the first international competition in which deep learning achieved a significant improvement in accuracy in speech recognition. In the following year, deep learning achieved a similarly dramatic improvement in image recognition. Two years later, in 2014, deep learning achieved the same accuracy in machine translation as highly complex previous systems.

These problems have been studied in various fields that support computer science, such as speech signal processing, computer vision, and natural language processing. In particular, since around 2000, data-driven solution methods that apply statistical machine learning methods to data sets with annotations have been attracting attention in each of these areas.

The mainstream methods for each task at that time were a multi-stage combination of data pre-processing, feature design, and post-processing developed independently for each study. However, basic machine learning methods and some ideas were shared. In other words, if a researcher in natural language processing, for example, wanted to start a new research project that included images, the cost of learning the techniques of the computer vision field and incorporating them into their research was a significant barrier.

In this context, research on deep learning, as mentioned earlier, has caused a sensation in various fields. Speech recognition was rewritten by multi-layer perceptron (MLP), image recognition by convolutional neural network (CNN), and machine translation by recurrent neural network (RNN). These MLP/CNN/RNN modules soon became uniformly used in their respective fields. For example, in speech and video recognition, MLPs, CNNs, and RNNs are used in a unified way. CNNs and RNNs are used for speech and video recognition time series, while CNNs are used for series understanding in machine translation. Combined with the fact that deep learning does not require manually designed features, as is already well known, it makes the barrier for integrating different tasks in each field much more effortless.

As a result, research to understand multiple modalities simultaneously, which had started gradually before the era of deep learning, was accelerated discontinuously by the rocket booster of deep learning. For example, the pioneering work on image caption generation [1] was born in 2010, and a pipeline training on pairs of images and associated captions [2] was proposed in 2011. With the introduction of deep learning in image recognition [3] in 2012 and machine translation [4] in 2014, the adoption of deep learning immediately progressed. In June 2015, at CVPR (Conference on Computer Vision and Pattern Recognition), the premier international conference in the field of computer vision, image caption generation based on the approach of understanding images with CNNs and generating sentences with RNNs was proposed simultaneously by universities and companies around the world.

In this way, deep learning on multimodal data is now widely conducted. The input can be any combination of various modalities such as speech, images, and natural language. The output can be any combination of these modalities by integrating and understanding the content. This paper aims to provide an overview of such research bridging computer vision and natural language processing, referred to as Vision and Language.

2 Vision and Language

The success of CNNs, RNNs, and more recently,

Transformer representation learning in image and natural language processing is mainly due to the ability to collect large amounts of images and text as datasets. It can be said that the data of these modalities are benefiting from such a large number of models to learn networks with huge parameters with a small inductive bias. Many texts related to images and videos are posted, especially on web services such as YouTube, Twitter, Instagram, and TikTok. The paired data of images or videos and natural language is multimodal data easily collected in large quantities.

In this section, I will discuss cross-modal understanding that connects such images and natural language. The combination of these two modalities has a particular name, Vision and Language. For several years, it has been a constant presence at international conferences on computer vision and natural language processing.

2.1 Caption Generation

Image captioning is the task of generating text that indicates the content of an image, and like other natural language processing research, it is often targeted at English. Image captioning has a long history in the field of vision and language. Several efforts in this area have been gradually increasing since around 2010, before the popularity of deep learning.

The first paper that tackled this problem [1] used Conditional Random Field (CRF) to estimate the three labels (triplet) of an image: "object," "action," and "scene," and then retrieved sentences with similar triplets from a set of captions with triplet added. In other words, the system does not generate new sentences, but it searches for sentences that match the contents of existing data. It is necessary to add a triplet to them. Later, a study [2] was proposed that aimed to generate image captions by taking image recognition and sentence generation modules from the fields of computer vision and natural language processing, respectively, and generating new sentences from only image and caption pair data.

Then, as mentioned in the previous section, caption generation by deep learning became a hot topic at CVPR 2015. Several papers proposed caption generation with the same pipeline simultaneously. The method by Vinyals et al. of Google [5] combines Google's CNN image recognition model, Inception, with an LSTM machine translation model. Despite its simple structure, the generated captions show significant improvement in both accuracy and fluency.

2.2 Visual Question Answering

Visual question answering (VQA) is the task of answering questions about a given image. In VQA, the information for answering a question is contained in the input image. VQA is a study of multimodal understanding, where the multimodal input of image and question text is used to classify the candidate answers.

This research project was the first application proposed in user interfaces [6]. It was intended to be used by visually impaired people to take pictures of something they were having trouble understanding while traveling and enter the questions they wanted to be answered. And in this paper, the questions were responded to manually by crowdsourcing people connected to the app. The visual question answering system presented here can be seen as an attempt to automate this process.

Visual question answering has been one of the most widely addressed problems in vision and language. A benchmark paper [7] was published in 2015. A new dataset was constructed by crowdsourcing questions that could not be answered without looking at the images and then crowdsourcing ten responses to each question from various people. The authors then proposed a protocol that evaluates a question as correct if three or more people answer it in unison. They also reported experimental results of a simple baseline in which images and question sentences are converted into features by an encoder, concatenated, and then classified among a set of existing answers. This well-designed data set, evaluation method, and easy-to-understand baseline are the factors that brought a large number of new entrants.

2.3 Multimodal Machine Translation

Multimodal machine translation uses a given image and its caption to translate the caption into another language. Conceptually, it is expected to improve the translation accuracy by using images to resolve ambiguities in the input text.

Hitschler et al.'s method [8] generates multiple candidate translation sentences from the input image and caption pairs by a well-known machine translation method on the captions only. The method then searches for the image-caption pairs of the target language to find multiple pairs consisting of input images and translation candidate sentences and updates (re-ranks) the score of the translation candidate sentences based on the search results to improve accuracy.

Multimodal machine translation is one of the typical examples of improving the accuracy of a problem that can only be performed in some modalities by increasing the number of modalities.

2.4 Image generation from text

Image generation from text is the inverse problem of image caption generation, which is generating an image showing the same content from a caption indicating the content of the image. It is a task that can be interpreted as the study of cross-modal understanding of text-to-image conversion. As a result of the widespread use of generative models such as generative adversarial network (GAN) and variational autoencoder (VAE), the challenges of this complex task are increasing.

Mansimov et al. [9] proposed a pipeline in which the input caption is encoded in a bidirectional RNN and then updated multiple times to make the image clearer in an RNN-based image decoder. Reed et al. [10] introduced GANs to this problem for the first time. In the paper of StackGAN [11], a 64-pixel-square image generated using almost the same approach as Reed et al. are re-input into another GAN with the input caption, resulting in a 256-pixel-square image. This approach of connecting GANs to generate high-resolution images will be continued.

The breakthrough in this problem was the approach by Transformer, and OpenAI published a blog post [12] on January 5, 2021, about DALL-E, a method that can generate much more diverse and natural images/illustrations from captions than ever before. It shocked the community by showing an example of how extraordinarily varied and realistic images and illustrations can be generated from captions.

2.5 Visual Dialog

Visual dialog is the addition of images and video to the study of dialogue, which has been done only with language. In a 2017 paper [13] titled Visual Dialog, a conversation model is learned by collecting data on dialogue related to images in natural language. It can be interpreted as a study of continuous cross-modal understanding. It generates natural language and other information as the next dialogue act from the history of dialogue in natural language and related visual information.

Dialogue generation is a research topic that has been addressed in artificial intelligence for a long time. There are many studies in a visual dialogue that deal with dialogue as such an artificial intelligence task. We ask two pairs of people to have a conversation using an image and collect their natural language data in a typical setup. For example, in a dialogue where one person is looking at an image and the other cannot see it, the content is conveyed verbally through a Q&A-style dialogue [13]. In a game where two people are looking at an image, one person has to guess the area the other is looking at in particular through a Q&A-style dialogue [14]. There are some settings for real-world applications; [15] deals with conversations between a store clerk and a customer, and [16] deals with ones between a telephone navigator and a tourist in a city.

In a broader sense, there is also research on cross-modal understanding by agents and robots. VLN (vision and language navigation) task [17] is one of the most famous of these, in which an agent or robot in a virtual environment relies on given verbal instructions and current visual information to reach a destination. In the sense that the next action, such as rotation or movement, is output as a dialog act based on the current visual information and the first verbal instruction. VLN is also research related to visual dialog.

3 Conclusion

This paper overviewed the research topic of connecting natural language and images as multimodal understanding research. It reviewed the research in the field of Vision & Language, which connects images and natural language. The author hopes that this paper will inspire more researchers to become interested in multimodal understanding.

References

[1] Ali Farhadi et al. Every Picture Tells a Story: Generating Sentences from Images. European Conference on Computer Vision. 2010, pp. 15-29.

[2] Yoshitaka Ushiku et al. Understanding images with natural sentences. ACM International Conference on Multimedia. 2011, pp. 679-682.

[3] Alex Kriuzhevsky et al. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems. 2012, pp. 1097-1105.

[4] Ilya Sutskever et al. Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems. 2014, pp. 3104-3112.

[5] Oriol Vinyals et al. Show and Tell: A Neural Image Caption Generator. IEEE Conference on Computer Vision and Pattern Recognition. 2015, pp. 3156-3164.

[6] Jeffrey P. Bigham et al. VizWiz: Nearly Real-time Answers to Visual Questions. ACM Symposium on User Interface Software and Technology. 2010, pp. 333-342.

[7] Stanislaw Antol et al. VQA: Visual Question Answering. IEEE International Conference on Computer Vision. 2015, pp. 2425-2433.

[8] Julian Hitschler et al. Multimodal Pivots for Image Caption Translation. Meeting of the Association for Computational Linguistics. 2016, pp. 2399-2409.

[9] Elman Mansimov et al. Generating Images from Captions with Attention. International Conference on Learning Representations. 2016.

[10] Scott Reed et al. Generative Adversarial Text to Image Synthesis. International Conference on Machine Learning. 2016, pp. 1060-1069.

[11] Han Zhang et al. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. IEEE International Conference on Computer Vision. 2017, pp. 5907-5915.

[12] Aditya Ramesh et al. DALL·E: Creating Images from Text. https://openai.com/blog/dall-e/.

[13] Abhishek Das et al. Visual Dialog. IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 326-335.

[14] Harm de Vries et al. GuessWhat?! Visual Object Discovery Through Multi-Modal Dialogue. IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 5503-5512.

[15] Amrita Saha et al. Towards Building Large Scale Multimodal Domain-Aware Conversation Systems. AAAI Conference on Artificial Intelligence. 2018, pp. 696-704.

[16] Harm de Vries et al. Talk the Walk: Navigating New York City through Grounded Dialogue. arXiv. 2018, 1807.03367.

[17] Peter Anderson et al. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 3674-3683.