別紙４－１（課程博士（英文））

Date of Submission (month day, year) : July 07th, 2021

Department of Computer Science and Engineering	Student ID Number	189303		Supervisors	Yoshiteru Ishida Jun Miura
Applicant's name	Rizal Setya Perdana				

Abstract (Doctor)

Title of Thesis	Contextualization of Multimodal Sequence Models on Image to Text in Story Generation

Approx. 800 words

A major achievement in artificial intelligence (AI), particularly in visual recognition, is developing a machine capable of understanding a complex visual scene. Beyond understanding the visual object, the other research subfield in AI, natural language generation systems, attempt to develop machines capable of describing objects with human language. Understanding dynamic visual scenes then describing them in words is easy for humans, but this task is difficult for machines. Research work in generating sentence descriptions from visual representations, e.g., image captioning, is continuously improving as computing technology, social networking platforms, and algorithms continue to evolve. In the real-life implementation, image to text system helps a visually disabled person understand images and possibly needed by a computer device with usability assistance. Although the image captioning algorithm has been shown to achieve success, it is limited to describing only a single image with a literal object description.

The current image captioning system cannot process a sequence of images directly with the output of multiple cohesive sentences. Shifting from a single image, static moment, and generating no-context-description toward a sequence of images that depict the dynamic event with an output cohesive narrative is challenging. Visual storytelling is an image-to-text task that comes with the more complicated scenario describing an image sequence into story style sentences. It utilizes the advance of computer vision capabilities in recognizing the complex visual object as a human-like inferencing ability in terms of structure and subjectivity. Previous approaches in visual storytelling systems generate less-than-accurate narrative sentences compared with the human-generated story. Several drawbacks were accused of describing literal objects only, monotonous sentences, and low-lexical variance so that the generated story suffers from being less coherent. Furthermore, less-context stories led to inaccurate information delivery due to the absence of the visual object validation mechanism and the minimum number of learning resources for the language model. Based on the aforementioned problems and limitations, we consider improving the output quality, i.e., generating a contextualized narrative story.

This dissertation develops techniques and models focusing on generating a narrative text based on visual sequences close to a human-generated story. To accomplish the achievements of the mentioned objectives, we break down the proposed approach into three sub-works. First, we design the experiment to build the image-text feature pair representation as a singular data point. This approach is named multimodal instance-based transfer learning tested in an image captioning task. As an initial part of the whole architecture, a simple setting is considered by representing the non-sequential feature with no coherent output expectation objective. Second, the absence of non-visual concept words in the generated sentence story, an important component in composing a narrative, was discussed. A non-visual concept is a word entity that accompanies the literal object that the visual object detection algorithm cannot recognize. We investigate the correlation of image-text pairs to generate new feature representation with the underlying concept of canonical correlation analysis in figuring out the drawback. Third, the lack of context in the previous approach's outputs leads to delivering an erroneous message and unwanted context. Thus, we attempt to improve the architecture with context-awareness by supplementing new features and incorporating external language resources on the decoding language generation stage.

The question of how to represent image-text pairs into new data representation was an early problem statement in this dissertation. To express a pair of image-text data into a single data point, simple concatenation of image feature and word representation vectors was not applicable due to the difference of data distribution between modalities. A single data point in the form of a vector should represent both modalities within a pair of image-text data. We employ a binary hashing mechanism that generates a mapping between original spaces into a Hamming space structured as binary codes. The new mapping representation was applied to transfer learning among the image captioning datasets to confirm the effectiveness. This method is particularly effective for single pair image-sentence comparison only; a new question was raised about how if the pairs are sequences of images paired with multiple sentences.

We formulate the investigation by exploiting the multimodal pairs' correlation to map the sequential pattern feature of image and text pairs for the visual storytelling task. To extract the sequential pattern from the array of images, we attempt to maximize the cross-modal correlation by extending the canonical correlation analysis, thus suitable for the sequential setting. The proposed end-to-end architecture particularly aims to encode the non-visual concept from the visual representation. As its objective is only to maximize the likelihood of input with the target sequence, it lacks the semantic correlation and results in the low-context output.

Considering the aforementioned limitation, we improve the architecture by extending the encoding and language generation decoding process to generate a coherent, object-focused, and contextualized sentence story. To overcome the inaccurate context, we incorporate the visual object detection feature to validate the literal object during the encoding process. From the language decoding, we utilize the pre-trained language generation model to contextualize the sequence generated language. In this study, all sequential pattern learning employs the self-attention mechanism that excels compared with the other approach.

Experimental results on the VIST dataset demonstrate the effectiveness of our proposed contextualized language generation and multimodal representation over the baseline based on automatic metrics.