PAPER

Vision-Text Time Series Correlation for Visual-to-Language Story Generation

Rizal Setya PERDANA^{†,††a)}, Nonmember and Yoshiteru ISHIDA^{†b)}, Member

SUMMARY Automatic generation of textual stories from visual data representation, known as visual storytelling, is a recent advancement in the problem of images-to-text. Instead of using a single image as input, visual storytelling processes a sequential array of images into coherent sentences. A story contains non-visual concepts as well as descriptions of literal object(s). While previous approaches have applied external knowledge, our approach was to regard the non-visual concept as the semantic correlation between visual modality and textual modality. This paper, therefore, presents new features representation based on a canonical correlation analysis between two modalities. Attention mechanism are adopted as the underlying architecture of the image-to-text problem, rather than standard encoder-decoder models. Canonical Correlation Attention Mechanism (CAAM), the proposed end-to-end architecture, extracts time series correlation by maximizing the cross-modal correlation. Extensive experiments on VIST dataset (<http://visionandlanguage.net/VIST/dataset.html>) were conducted to demonstrate the effectiveness of the architecture in terms of automatic metrics, with additional experiments show the impact of modality fusion strategy.

key words: visual storytelling, correlation analysis, attention mechanism

1. Introduction

The tremendous advancement of text generation from visual representation (image captioning), has brought a slightly more complicated task, known as visual storytelling. Figure 1 presents how visual storytelling translates a sequential set of images into a cohesive narrative story through time [1]. The enormous number of images uploaded to social network services (SNS) triggers the generation of sets of images known as photo albums that motivate the research development of visual storytelling. Such an application potentially can help visually impaired person to grasp information about the images in an album automatically instead of needing manually labeling. Besides describing the literal object from images, the generated story is also composed of non-visual concepts, which require creativity and subjectivity to interpret.

Previous studies have had a similar focus, designing deep neural network architecture [2]–[4] to gain higher stan-

Fig. 1 Visual storytelling task translates visual input to text story output. There are five pairs of images sequence with sentence stories each. (a) Maximize the correlation for two exact matches of the image (triangle-shape) and the text (circle-shape) which no correlation among them. (b) Not only maximize the correlation between image and text, but the proposed model also maximizes the correlation between each pair in a story (rectangle-shape). This will extract the non-visual concept as shown in the red printed text of Story Output.

dard evaluation metrics results of vision-to-stories. Most of the architecture relies on modification of the encoder-decoder, such as the attention mechanism [5] that is widely used in machine translation, image captioning, video captioning, and other tasks. Current areas still unexplored are non-visual concepts (information from the image that not literally appears as a visual object) those used to compose stories. As shown in Fig. 1, the red printed word in **Story Output** such as ‘beach’, ‘Christmas’, ‘picked’, and ‘great time’ are the example of non-visual concept words composed a story. Previous study VSCMR [6] by Li et al. categorized non-visual concepts that are crucial in generating stories from visual representations including events, actions, attributes, and entities. They also applied association rule mining to explore cross-modal rules, looking at joint representation between visual features and an embedded vector of text as a filter on attention layer.

Knowledgeable Storyteller [7] focuses on common-sense reasons in addressing the generation of imaginary concepts. Similar to [7], in order to generate a description of the novel (or non-visual) objects which not present in the paired image-sentence dataset, *Deep Compositional Captioner* (DCC) [8] leveraging the large visual object recognition datasets and external text corpora. Previous works, such as [6], [7], achieved significant results by considering the non-visual concept. As it turns out, they both have

Manuscript received June 17, 2020.

Manuscript revised October 23, 2020.

Manuscript publicized March 8, 2021.

[†]The authors are with Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi-shi, 441–8580 Japan.

^{††}The author is with Department of Information System, Faculty of Computer Science, Universitas Brawijaya, 65145, Indonesia.

a) E-mail: rizal.setya.perdana.to@tut.jp

b) E-mail: ishida@cs.tut.ac.jp

DOI: 10.1587/transinf.2020EDP7131

strong ties to an external knowledge base instead of end-to-end deep architecture. Thus the story generated is dependent on the external sources that cause a lack of flexibility and independently as learning representation from data.

In order to address the aforementioned limitation, cross-modal correlation will be included as the semantic association factor instead of depending on external knowledge. By using cross-modal correlation learning, the two different modalities of visual storytelling data representations will be mapped into a common space by maximizing the correlation. The new correlation feature vector works for the attention layer as the guidance in order to the decoding process in generating stories. Semantic relation extraction between images and text in [9] utilizes canonical-correlation analysis (CCA), a classic method for this problem. In contrast, deep canonical-correlation analysis (DCCA) [10] has non-linearity feature transformation through the deep neural network. This research introduces end-to-end architecture: correlation analysis attention mechanism (CAAM). It consists of three main modules includes a modality encoder module, cross-modal correlation learning with attention module, and decoder story module. A summary of the contributions of this paper is as follows:

- We propose an end-to-end architecture of visual story generation by introducing an attention mechanism with time series correlation as the guidance for the decoding process to reveal the composition of non-visual concept.
- As the visual storytelling data is arranged in order timeline, the study introduces the first correlation learning applied to sequential data.
- The results of experiments and an evaluation of the model for visual storytelling dataset (VIST) are presented using standard automatic metrics in confirmation of the outperform results of the proposed architecture.

2. Related Work

As one of the image-to-text active research tasks, visual storytelling can be separated into two groups of research focus. First is focus on content understanding by [4], [5], [9]–[11] that attempted to discover the latent context of the sequence of images and generation by the decoder. Most of them focused on how to extract important features on the sequential images and decode them into a text story. Second, the focus on sentence story generation by [12]–[16] explores the story generation procedure from standard features. This group emphasizes the quality of the story generation process which has many differences compared to literal description generation. Our research focused on extracting the latent representation of multimodal data (sequential images and text story) rather than using external knowledge source to develop a near human story generation model.

Related to this research in learning new data representation, multivariate statistical method called CCA (Canon-

ical Correlation Analysis), introduced by [17], allows one to obtain a latent representation of multimodal data. The current improvement on CCA is deep canonical correlation analysis (DCCA) [18] learned that complex nonlinear transformations of two different modalities resulted in a new feature representation, with high linearly correlated. One of the implementations of DCCA is [19] where multimodal data is combined for emotion recognition. In the method presented in this paper, the DCCA was used to maximize the correlation between image and text story to generate new representation.

Later, to respond the limitation of the encoder-decoder model in compressing long sequence data into a fixed length vector, [20] proposed attention mechanism to overcome this issue. GLAC Net [5] the first attention-based visual storytelling research that effectively focused only on the important part of visual representation. This paper incorporated the new features representation based on time series canonical correlation to produce the context vector.

3. Background

Deep Canonical Correlation Analysis (DCCA), proposed by [19] attempted to discover new feature representation by fusing two modalities for emotion recognition. It attempted to discover the relation of two different feature modalities, such as text and sound, based on the traditional canonical correlation analysis (CCA) proposed by [17]. Let $X \in \mathbb{R}^{n_1 \times m}$ and $Y \in \mathbb{R}^{n_2 \times m}$ are two sets of vectors with the same number of vectors m . CCA attempts to learn new vectors representation $A \in \mathbb{R}^{n_1 \times r}$ and $B \in \mathbb{R}^{n_2 \times r}$ as linear transformation that satisfying an objective to maximize the correlation between $A^T X$ and $B^T Y$. Given X and Y are the two sets vectors, the covariances are S_{11} and S_{22} with the cross-covariance as S_{12} . The CCA is to learn to optimize the objective as follows:

$$\begin{aligned} A^*, B^* &= \arg \max_{A, B} \text{corr}(A^T X, B^T Y) \\ &= \arg \max_{A, B} \frac{A^T S_{12} B}{\sqrt{A^T S_{11} A} \sqrt{B^T S_{22} B}}. \end{aligned} \quad (1)$$

The solution of the objective in Eq. (1) is already done, but it possible to be solved in other ways as suggested in [21]. The suggestion define U, S, V^T be the Singular Value Decomposition (SVD) of the matrix $Z = S_{11}^{-\frac{1}{2}} S_{12} S_{22}^{-\frac{1}{2}}$ where S_{11}, S_{22} is the covariance of of two vector set X and Y , while S_{12} is the cross-covariance. By these definition, we can state the total maximum canonical, A^* , and B^* as follows:

$$\begin{aligned} A^* &= S_{11}^{-\frac{1}{2}} U \\ B^* &= S_{22}^{-\frac{1}{2}} V \\ \text{corr}(A^{*T} X, B^{*T} Y) &= \sqrt{\text{trace}(Z^T Z)}. \end{aligned} \quad (2)$$

As mentioned before, CCA has a limitation, only addressing linear transformation instead of nonlinear transformation. DCCA attempt to learns non-linear transformations

between two independent neural networks. The two separate neural networks denoted as p and q ; DCCA has the objective to optimize θ_p and θ_q as the parameters of these networks. The canonical correlation of the two networks p and q with the two random variable input X and Y is denoted as $F_X = p(X; \theta_1)$ and $F_Y = q(Y; \theta_2)$ can be maximized by generating the two linear transformations M^*, N^* . The DCCA is to learn to optimize the objective as follows:

$$\begin{aligned} \theta_p^*, \theta_q^* &= \arg \max_{\theta_p, \theta_q} \text{CCA}(F_X, F_Y) \\ &= \arg \max_{\theta_p, \theta_q} \text{corr}(M^{*T} F_X, N^{*T} F_Y). \end{aligned} \quad (3)$$

The network parameters in calculating the canonical correlation need to be updated based on the loss function in a back-propagating manner. Defined F_X and F_Y as the result of the two canonical correlation from two random variables, R_{11} and R_{22} are the covariances with the cross-covariance R_{12} . Refers to the suggestion in Eq. (2), for non-linear CCA can be restated as U, R, V^T , which are the single value decomposition of the matrix E so that the value of θ_p^*, θ_q^* stated as follows:

$$\begin{aligned} E &= R_{12} R_{11}^{-\frac{1}{2}} R_{22}^{-\frac{1}{2}} \\ \theta_p^* &= R_{11}^{-\frac{1}{2}} U \\ \theta_q^* &= R_{22}^{-\frac{1}{2}} V \end{aligned} \quad (4)$$

The loss function of the DCCA for updating the weight F_X and F_Y can be defined as follows:

$$\text{CCA Loss} = -\sqrt{\text{trace}(E^T, E)}. \quad (5)$$

The negative sign added for the loss function to inverse the objective i.e maximizing the correlation. Both of the two network parameters θ_p and θ_q from networks $p(X; \theta_p)$ and $q(Y; \theta_q)$ are optimized to minimize the loss value or maximizing the total of canonical correlation.

4. Proposed Model

4.1 Overview

Different from standard image to text problem, visual storytelling has an objective to translate the array of ordered images D_V into a coherent text story D_S , which contains both literal objects and non-visual concepts. The sequential array of images $D_{V_i} = \{v_1, \dots, v_t\}$ represents the timeline of events in the photo story and generated text story $D_{S_i} = \{s_1, \dots, s_t\}$. A sentence s_t that composing a story D_{S_i} is representing the sequence of words w with length varies between each sentence. For each story in VIST dataset, we learn from 5 images and generate 5 sentences with $t = 5$ from each to build a visual storytelling model. Our proposed model first attempts to build a new feature representation based on the canonical correlation analysis between two modalities including visual and textual, as explained in Sect. 4.2. Note: later, the

Fig. 2 The architecture overview of feature fusion based on non-linear canonical correlation analysis. The left side is the feature learning from visual modality and the right side is the textual modality learning representation.

new feature representation will be used to build the context vector of attention mechanism described in Sect. 4.3.

4.2 Visual-Textual Modality Correlation

Exploring the relation between visual modality and text modality on automatic visual storytelling remains an unexplored research area. Existing non-linear transformation of deep neural networks in canonical correlation analysis (DCCA) is not able to handle sequential data such as a sequence of images and words in a sentence. Previous research has been conducted in exploring the correlation of single modality time-series data such as [22], [23]. In this research, we apply a combination of RNN based deep neural network and canonical correlation analysis (CCA), as shown in Fig. 2.

4.2.1 Long Short Term Memory Cell

To handle time-series data both images sequence and sentence of stories, Long Short Term Memory (LSTM) is used to learning patterns across time. The learning process of the LSTM network evaluate each variables in Eq. (6) for every epoch.

$$\begin{aligned} f_t &= \sigma(W_f[c_{t-1}, h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i[c_{t-1}, h_{t-1}, x_t] + b_i) \\ g_t &= \tanh(W_g[h_t, x_t] + b_g) \\ c_t &= f_t \times c_{t-1} + i_t \times g_t \\ o_t &= \sigma(W_o[c_{t-1}, h_{t-1}, x_t] + b_o) \\ h_t &= o_t \times \tanh(c_t) \end{aligned} \quad (6)$$

The LSTM cell requires the following input $[c_{t-1}, h_{t-1}, x_t]$ which annotate cell state from the previous step, the output from the previous step, and the input at the current time step respectively. All of these inputs pass through upon the two sigmoids to obtain the input gate i_t and forget gate f_t

Fig. 3 The proposed framework for time-series canonical correlation analysis for visual storytelling. This proposed approach generates a combined time-series multimodal feature based on non-linear correlation analysis.

separately. The raw update as the new candidate, g_t , updates the cell state c_t obtained from passing through the parameters into the hyperbolic tangent function $\tanh$. Raw output o_t obtained from passing the input parameters through the sigmoid. Finally, the output h_t defined by the element-wise multiplication of raw output o_t with the cell state within $\tanh$ function.

Two independent LSTMs composed in forward and reverse direction in reading the sequence each other known as bidirectional LSTM (Bi-LSTM). Bi-LSTM was introduced by [24] to improve the efficiency and capacity in learning the sequences of data. The learning process of Bi-LSTM networks as shown in Eq. (7) optimize the $LSTM_f$ and $LSTM_b$ are the forward and backward of LSTMs respectively.

$$\begin{aligned} h_t^f &= LSTM_f(x_t, h_{t-1}^f) \\ h_t^b &= LSTM_b(x_t, h_{t-1}^b) \\ h_t &= W_f h_{t-1}^f + W_b h_{t-1}^b + b \end{aligned} \quad (7)$$

4.2.2 Time-Series CCA on Visual Storytelling

Based on previous research related to multimodal feature fusion [25], [26], the use of the outer-product through different modality can learn the combination effectively. In our proposed approach in visual storytelling as shown in Fig. 3, the outer-product between visual and textual sequences are applied in implementing time-series CCA. Let $D_v \in \mathbb{R}^{d_v \times l_v \times N}$ be the visual inputs from image sequence, and $D_s \in \mathbb{R}^{d_s \times l_s \times N}$ as the textual inputs from text embedding sequence with N number of data. Both D_v and D_s are the input for the network with the same length of sequence $l_v = l_s = 5$ for

each data in VIST dataset. A pre-trained networks ResNet, stack of convolutional neural network layers, used to extract the spatial feature of an image with the output denoted as $D_{v1} \in \mathbb{R}^{d_{v1} \times l_v \times N}$. In order to obtain the temporal features from an image sequence, the LSTM network applied in five sequences of extracted spatial features, with the output a final hidden state denoted as $D_{v2} \in \mathbb{R}^{d_{v2}}$.

Next, the textual modality is comprising of two-level data representation, including sentence-level and story-level representation. In the VIST dataset, a story-level representation is consisting of five sentences each. For each sentence, the embedding representation of a sentence-level fed into the LSTM network, then the final output of the LSTM $D_{s1} \in \mathbb{R}^{d_{s1} \times l_s \times N}$ will fed into the Bi-LSTM network to obtain the story-level representation. The output of story-level representation learning $D_{s2} \in \mathbb{R}^{d_{s2} \times N}$ obtained from the final output of the Bi-LSTM layers. Both visual sequence and textual story features then fed to the Bi-LSTM network to obtain the final output of Bi-LSTM $D_{v2} \in \mathbb{R}^{d_{v2} \times N}$ for visual modality and $D_{s2} \in \mathbb{R}^{d_{s2} \times N}$ for textual modality.

The extracted features from both visual and textual modality then proceed to learn the fully connected layers that optimized the network parameters based on the canonical correlation analysis loss function (CCA Loss) as described in Eq. (5). The final representation from the combination of both visual and textual modality $D_{vs} \in \mathbb{R}^{d_{v2} \times d_{s2} \times N}$ is the outer-product of the final output vectors of fully connected layers as shown in Eq. (8). Furthermore, the final combination will be used as the context vector in the attention mechanism in the next stage.

$$D_{vs} = D_{v2} \otimes D_{s2}, D_{vs} \in \mathbb{R}^{d_{v2} \times d_{s2} \times N} \quad (8)$$

4.3 Correlation for Attention Mechanism

The underlying architecture of the proposed method was the seq2seq method with an attention mechanism [20], which overcame the problem of standard encoder-decoder in memorizing the long data sequence. In this research, we propose to take advantage of the new feature representation with the modality fusion for context vector generation based on the canonical correlation as shown in Eq. (8). In a standard encoder-decoder architecture, the encoder transforms the input into a single fixed-length vector representation, followed by the decoder that translates the fixed-length vector into the outputs. Instead of translating to a single, fixed-length vector representation, the attention model works to find out the context vector from each step of the input as the guide for the decoder. Later, the decoder not only considers the encoder's output but also involves feature combinations that include the semantic correlation between visual and textual modality, as described in Sect. 4.2.2. Detailed processes of the attention mechanism are described as follows:

4.3.1 Encoding the Input

This process is similar to the encoding of the seq2seq approach, which feeds each input sequence $D_v \in \mathbb{R}^{d_v \times l_v}$ into the RNN based network. Instead of using only one final hidden state output from all sequences, the attention mechanism takes on every hidden state output $D_{v1} \in \mathbb{R}^{d_{v1} \times l_v}$ generated from each visual input sequence encoding process as described in Sect. 4.2.2.

4.3.2 Calculating the Alignment Score

In standard encoder-decoder architecture, after the result of the encoding process is obtained, it then continues to decode the final encoder output to produce the final output. An additional step, as the essence of the attention mechanism, calculating the alignment score aims to measure the alignment between each step of all encoder output with the previous decoder hidden state. This research proposed to add an extra component in calculating the alignment score by considering a matrix $D_{vs} \in \mathbb{R}^{d_{v2} \times d_{s2}}$ as the multimodal features combination which highly correlated. The main objective of involving the feature combination is to provide the decoder guidance related to the unexplored semantic correlation between visual and textual modality. Derived from [20], the alignment score $score_i$ measured between the previous decoder hidden state H_d to each step of encoder hidden state D_{v1} that aims to quantify the degree of attention. The proposed approach learn by adding the three parts: feature combination matrix D_{vs} , the decoder hidden state H_d , and the encoder hidden state D_{v1} . The trainable weight of calculating the alignment score including $W_d, W_{D_{v1}}, W_{D_{vs}}, W_c$ that represents decoder weight, encoder weight, feature combination matrix weight, and the combined weight consecutively.

$$\begin{aligned} enc_i &= fc(W_{D_{v1}} \cdot D_{v1_i}), i = 0 \dots l_v \\ dec_j &= fc(W_d \cdot H_d), j = 0 \dots l_v \\ cmb &= fc(W_{D_{vs}} \cdot D_{vs}) \\ score_{ij} &= W_c \cdot \tanh(enc_i + dec_j + cmb) \end{aligned} \quad (9)$$

In training process, encoder output from each step, decoder hidden state, and feature combination matrix are feeds through the fully connected layer individually with the output enc_i, dec, cmb as shown in Eq. (9). The three training weights will be joined together by additive operators before passed through a hyperbolic tangent activation function. In particular, for each encoder output will be added with the decoder hidden state. The last, to obtain the alignment score, the output combination from the hyperbolic tangent is multiplied by the combination vector weight W_c .

4.3.3 Calculating the Attention Weight

The next procedure, after all of the alignment scores measured is obtaining the attention weight α_{ij} for the encoder output vector. For each encoder vector output will have the degree of attention ranging between 0 to 1. This procedure applies the softmax function for all encoder vector so that the sum up of all weight is 1.

$$\alpha_{ij} = \frac{\exp(score_{ij})}{\sum_{i=0}^{l_v} \exp(score_{ij})} \quad (10)$$

4.3.4 Generating the Context Vector

In order to generate a mapping between the encoder output with the decoder, the attention mechanism creates a shortcut between the context vector and all of the input. After the previous step finishes generating the attention weight α_{ij} , the context vector c_i produced by the element-wise multiplication between all encoder outputs D_{v1_j} with the attention weight α_{ij} .

$$c_i = \sum_{j=1}^{l_v} \alpha_{ij} D_{v1_j} \quad (11)$$

As the context vector is the derivation from the encoder output and the softmax of attention weight, an input element tends to has a value close to 1, which has strong attention rather than the input element close to 0 that has weak attention.

4.3.5 Decoding the Result

If only the context vector has ready, the decoding process can be executed by concatenating the previous decoder output x_{t-1} with the generated context vector c_i . The resulting concatenation as input then passes through the decoder LSTM networks together with the previously hidden state h_{t-1} to obtain a new hidden state h_t . The iteration will be repeated from the 4.3.2 until facing the conditions of the decoder generates stop token or exceeds the maximum length.

Fig. 4 The overall architecture design. (a) Images feature extraction and story embedding representation. (b) Time Series Canonical correlation analysis to calculate new feature fusion representation. (c) Image feature sequence story extraction. (d) Attention mechanism based on new feature fusion. (e) LSTM decoder to generate sentence story. (f) The output of sentence story.

At the final process, to obtain the next word prediction, the new hidden state h_t passes through a linear layer as the classifier to get the probability score s_t .

$$\begin{aligned} h_t &= \text{LSTM}(c_i + x_{t-1}, h_{t-1}) \\ s_t &= \text{softmax}(fc(W_s \cdot h_t)) \end{aligned} \quad (12)$$

5. Experiments

5.1 The Dataset

We performed comprehensive experiments on the Visual Storytelling Dataset (VIST) dataset [1], also called the Sequential Image Narrative Dataset (SIND) v.2. The VIST comprises descriptions of images-in-isolation (DII) and stories of images-in-sequence (SIS), which have different annotation characteristics when describing the photos. For the visual storytelling problem, the SIS text annotation was used. Each story consisted of five images, followed with five sentences that described the sequence of images. The VIST dataset contains 209,651 images in total split into 167,528 for training, 21,048 for validation, and 21,075 for testing purposes. Since each story contains five images each, this dataset arranges in 50,200 stories in total and splits up as 40,155 stories, 4,990 stories, and 5,055 stories for training, validation, and testing, respectively.

5.2 Settings

Text Preprocessing. Sentences tokenization performed to transform the sentences into words or tokens that used for further purposes i.e., transform the text into numerical values. By counting the presence from each token, only token

that adequate the minimum threshold of presence number will be preserved as the vocabulary, other than that, will be filtered out. In this research, the minimum threshold of token presence is 10 refers to [27]. The vocabulary is a custom data structure that comprises token-index pairs, later this structure used as the index dictionary of the word embedding purpose. The extra tag-like token also appended such as <pad>, <start>, <end>, and <unk> which has special purpose of each. The aim of the text pre-processing is to reduce the training complexity due to the high variance of the infrequent tokens.

Token Distribution Analysis. To investigate the token frequency distribution, Fig. 5 shown the word frequency distribution across the story's sequences. Word frequency distribution is not including the stop word, number, and punctuation. Based on Fig. 5, the appearance of some words is uniquely describing the sequence of the story. For instance, the first sequence story, the presence of words 'location', and 'today' indicating the opening of a story. The closing words of the story which appears in the last sentence, such as 'end' and 'night' can indicate that some words have their specific purpose. In addition, in order to investigate the frequency distribution of non-visual concepts in the story generation, Fig. 6 presents the heat map of two kinds of word groups as well as transition and adjective words. To produce this heat map, first, the investigation of considering transition and adjective word as the foremost part of the non-visual concept was decided. This figure examines both the quality and quantity of the presence of non-visual concept words with or without applying the CAAM architecture. Token distribution analysis is aimed to analyze the initial condition of the word distribution in general. From the result, the pattern can be presumed that both the general word and

Fig. 5 Word frequency distribution for each sequence in the story. The ‘Sequence 1’ is the first sentence regards as the opening of the story that followed by Sequence 2, 3, 4, and 5. The number of frequency indicating the uniqueness of words appears in a particular sequence.

Fig. 6 Top-40 of the non-visual concept comprises two groups of words, such as transition and adjective words. The two heat map briefly represents the density of frequency distribution on each sequence of the story for both of non-visual concept words groups.

the transition word are clearly segmented based on the sequence order as shown in Fig. 5 and 6.

Image Preprocessing. Visual pre-processing performed by re-sizing all images in 256-pixel squares in order to reduce the computational cost in tensor processing. For feature extraction, the pre-trained model ResNet-152 introduced by [28] was used by removing the last two layers. Before the features were extracted, the entire image was cropped at random to 224 pixels and normalized by both

averages and the standard deviation value of the Image Net dataset. The use of a pre-trained model for feature extraction aimed at computational and time efficiency rather than training from scratch with a limited number of datasets.

Hyperparameter Configuration. In the experiment, the training process used 64 data instances for a mini-batch that reshuffled at every epoch. This research applies two different loss functions and two different optimization functions for particular purposes. The loss function for the time-

series canonical correlation analysis following the CCA loss function from Eq. (5), while the story generation networks use the standard Cross-Entropy loss function. Adam optimizer [29] was used in story generation networks to optimize and update the network weights with the initial learning rate parameter set at $1e-3$ and weight decay set at $1e-5$. Different from the story generation networks, by following the DCCA proposed by [19], the time-series CCA networks apply the RMSprop optimizer with the initial learning rate parameter set at $1e-3$ and weight decay set to $1e-5$ as the regularization parameter of the network. The proposed architecture implemented in PyTorch [30], an open source Python deep learning library, in GPU [31] computer that optimizes for tensor computation to learn the model architecture.

Overall Settings. Refers to the overall architecture, settings experiment divide into two main parts. First, the networks associated with the time-series canonical correlation analysis, which produces the feature combination between visual and textual modality. Second, the networks associated with the story generation based on the attention mechanism which involving the result from the first networks. Some parts share for both two networks mentioned before, such as the visual feature extractor and the textual embedding networks. In this research, the visual features vector with the output vector length 1024 and the word embedding vector with length 256 use in learning both time-series canonical correlation and generating the output stories.

5.3 Evaluation Metrics

Some automatic evaluation metrics were applied for natural language generation tasks to quantify the effectiveness of the proposed model. METEOR [32], a standard evaluation originally for a natural language machine translation task that quantifies the quality of the generated text was used. This metric proposed to overcome the problem of the BLEU [33] score, which lacks specific matches. We also applied CIDEr-D [34] to measure the similarity between the generated text and the human-generated story which did not use precision-based metrics. We also applied ROUGE-L [35], which focuses on recall oriented for generated text summary. For the evaluation, this research implement the code from this repository[†].

6. Results and Analysis

6.1 Experiment Results

In this research, there are two primary investigations conducted, such as the effect of multimodal correlation and the story generation quality.

Visualization Analysis. To examine the effectiveness of the new feature generation based on canonical correlation, Fig. 7 presenting the t-SNE visualization [39] of high

dimensional multimodal features that split into several scenarios. The t-SNE algorithm reduces the data dimension to visualize the high dimensional data adequately. Based on Fig. 7 (a), data points which have the same color represent the same sequence among stories, it appears that the majority of data points from the same story tend to have a close position. Besides, most of the same color data points scattered randomly, but concurrently they group with other colors data points in arranging a story. In Fig. 7 (b), the same color data point representing that the features are from the same story. Data points in Fig. 7 (a) and 7 (b) can be analyzed that the similarity of the intra visual-feature of a story tends to have higher similarity compared with the inter visual-feature (another story), which means the images from the same story tend to have similar looks. Figure 7 (c) and 7 (d) are the t-SNE representation for the textual modality. Figure 7 (c) has a similar pattern compared to the Fig. 7 (a), while the Fig. 7 (d) has scattered data points randomly among the same story in comparison with the Fig. 7 (b). Data points in Fig. 7 (d) do not present a clear pattern as in Fig. 7 (b). This condition can be analyzed that some groups of words have a particular tendency to present in a specific sequence, as presented in Fig. 6. In other words, the dominating character of the sequence-specific word affects this condition. Figure 7 (e) and 7 (f) are presents the feature fusion of the visual and textual features. Figure 7 (e) shows the fusion of the original features, both visual and textual, that look randomly scattered with no obvious pattern that can be observed. After applying the time-series canonical correlation analysis, the fusion of the two modalities as the new features can be observed in Fig. 7 (f). The semantic pattern group of data plots is clearly present where the data point from the same sequence and same modality forming a cluster as the consequence of time-series correlation.

Observing Automatic Metrics. Table 1 presents a comparison of the automatic evaluation results of the baseline to the proposed model. **Neural Image Caption (NIC)** [36] inspired by seq-to-seq machine translation that encodes the visual representation into a fix-length vector and decodes it into a variable-length of text. NIC trained to maximize the likelihood of input image with the target sequence of words. The adaptation was performed due to the difference in the input form that originally singles into an array of images to fulfill the required visual storytelling task. **Visual Attention** [37] allows the salient features of visual representation dynamically used in the language generation. Instead of using the static visual representation, this approach focuses on the important object and omitting unimportant objects. Similar to [36], adaptation performed to receive an array of images rather than the single one. **Global Local Attention Cascading (GLAC)** [5] faces the problem of generating an image-specific sentence that covers the overall image representation in sequence. Two-level of attention comprised of overall global encoding and local features of an image. These attentions implement via hard connections from the output encoder onto the sentence generator. **Hierarchical Aligned Cross-modal Attention (HACA)** [38]

[†]https://github.com/lichengunc/vist_eval

Fig. 7 Visualization comparison by t-SNE of high dimensional multimodal features of VIST dataset divided as follows: (a) visual features group by sequences, (b) visual feature group by stories, (c) textual features group by sequences, (d) textual features group by stories, (e) the combination of original features, (f) the new representation of correlation dataset from the proposed model.

Table 1 The automatic evaluation (METEOR, CIDEr, ROUGE-L, BLEU) comparison of the proposed architecture with the baselines.

Method	METEOR	CIDEr	ROUGE-L	BLEU 1	BLEU 2	BLEU 3	BLEU 4
NIC [36] (a)	27.60	1.60	21.80	29.20	14.00	7.00	3.60
NIC [36] (b)	29.30	3.60	23.10	33.41	17.70	8.90	4.60
Visual Attention [37]	30.41	3.40	24.28	34.89	18.87	9.32	4.82
GLAC [5]	28.90	2.60	22.80	32.80	17.20	8.60	4.40
HACA [38]	30.00	2.00	23.70	33.80	18.00	9.10	4.40
Knowledgeable VIST [7]	30.89	3.12	23.32	30.41	16.98	9.12	4.80
Proposed approach	31.23	3.30	24.72	33.32	18.93	9.60	4.98

fuses both local and global temporal dynamics of different modalities. HACA addressing the fusion problem of the multimodal domain to learn temporal features from multiple modalities. By using cross-modal attention through the temporal structure, this approach discovers the beneficial to learn and align both global and local temporal transitions of multiple modalities. **Knowledgeable Storyteller** [7] attempts to overcome the problem in the absence of the non-visual concept by incorporating external knowledge. By using a knowledge graph, the non-visual or imaginary concept can effectively integrate with semantic-relevance that enhance the coherence of the generated text.

Effect of Time-series Canonical Correlation. Perfor-

mance comparison to the other image to text task or known as image captioning, is conducted by re-implementing the research from [36] in two scenarios. First, before the training performed, the visual features are concatenated with the textual features and apply the encoder-decoder approach. The second scenario is flipping the concatenated step into the last after the decoding process. Based on the result, the latter scenario has outperformed the concatenation before training. Another analysis based on the result comparison, the video captioning [38] approach outperformed the image captioning [37] approach in learning the data model for visual storytelling. Based on this result, there is an indication that the visual storytelling task is more close to video cap-

Table 2 The automatic evaluation of ablation study. Full model meaning that the proposed time-series CCA with attention mechanism is applied, “w/o time-series CCA” meaning that the decoding process performed without adding features combination from time-series CCA, “w/o CCA (with feature concatenation)” meaning that the decoding process performed with features combination by features concatenation only, “w/o Attention decoder” meaning that the decoding process performed without attention mechanism.

Models	METEOR	CIDEr	ROUGE-L	BLEU 1	BLEU 2	BLEU 3	BLEU 4
Full model	31.23	3.30	24.72	33.32	18.93	9.60	4.98
w/o time-series CCA	23.43	1.82	22.91	30.54	15.23	6.71	3.80
w/o CCA (with feature concatenation)	27.30	2.50	23.20	34.98	17.20	8.40	4.20
w/o Attention decoder	24.67	2.43	21.34	28.12	13.73	5.58	3.92

Fig. 8 Average correlation coefficient comparison of visual-textual features over a different number of dimensions.

tioning than image captioning regarding the data structure similarity. Later, to compare the performance from standard attention of image to text, both research in [37] and [5] are implemented. Our proposed approach is outperform compared to the attention approach by adding the feature combination from canonical correlation. The last, compared to the research in the knowledgeable storyteller [7] that involve the external knowledge, our proposed approach outperforms except in BLEU 1 score.

Correlation Coefficient Observation. In this analysis, we present a comparison of the correlation strategy for the time-series multimodal data. Figure 8 is adopted from [18] in showing the average of the correlation coefficient of the visual and textual modality from the original features compared to several correlation-based fusion features. The various number of dimension size (feature-length) used as the input to determine the correlation coefficient applied to the original dataset, CCA, DCCA, and the proposed approach time-series CCA. The result shows that the proposed approach has the highest average score compared to the other, and correlating the original data is the lowest average score. As the visual storytelling data is arranged in a sequence

Fig. 9 The image sequence sample of visual storytelling dataset.

Ground truth #1 our kids have many different interests. one likes to play football. #2 Another love matching games. #3 one likes to play on the computer. #4 the little one loves to cuddle #5 but all of them love to eat fresh berry pie.

Proposed approach #1 first the kids were having a great time at the party. #2 they were all dressed up for the occasion. #3 there were many people playing games. #4 afterward some of them were very tired. #5 finally, we had a lot of food.

Fig. 10 Story output comparison with extra non-visual concept word of Fig. 9.

manner, it can be analyzed that the result of the proposed method is averagely outperformed as the consequence of considering the time-series aspect in determining the correlation in order to fusion the multimodal features. The significant score of correlation indicates that the semantic fusion between modality in a time-series setting gives an impact on the story generation.

Qualitative Analysis. Figure 9 and 10 is the qualitative comparison of the generated story output compared with the ground truth on VIST dataset. By this sample output, the proposed architecture able to generate a text story with the additional non-visual words, as shown in the underlined words. The presence of the non-visual word can be regarded as the result of extra guidance on the attention mechanism by combined multimodal features. From the result can be analyzed that the obtained time-series multimodal correlation can perform significantly in the entire sequence. This

is shown in Fig. 10, from five sentences the model can generalize sequence-specific words such as ‘first’, ‘great time’, ‘afterward’, ‘tired’, ‘finally’ in the proper sequence.

6.2 Ablation Study

To examine and gain insight into the effect of the building block on its overall performance, we divide the ablation study into several parts.

Time-series CCA ablation. The time-series CCA ablation performed to explore the effect of combining multimodal features in visual storytelling tasks. This investigation divide into two parts, first the training process perform without fully additional feature combination, and the second the training perform with additional features concatenating without CCA. The result is shown in Table 2 can be analyzed that the absence of time-series CCA gives the lowest score compared with the feature combination by concatenation. This implies that in multimodal manners, the feature fusion has an impact on guiding the decoder. On the other hand, using the features fusion by only concatenation of the textual and visual feature is ineffective due to the difference in the probability distribution.

Attention ablation. The attention mechanism ablation performed to investigate the effectiveness of the guided decoder by the attention mechanism. Table 2 shown the performance of evaluation metrics if the attention is absent it produces a large gap with the full modal. As the drawback of the standard seq-to-seq model instead of the attention model is the difficulty in memorizing long sequences to generate text from a long vector representation, this ablation results answer that kind of problem.

7. Conclusion and Future Directions

This study incorporated a time-series canonical correlation feature representation to improve visual storytelling. The new feature representation is the result of learning how to maximize the visual-textual pairs used by the attention mechanism. The attention mechanism can guide the decoder to generate a coherent story. The experiment showed that the proposed model outperformed some automatic evaluation metrics that arise with the use of more non-visual concept words. In the future, instead of automatically quantitative measurement, the text generation task like visual storytelling also need automatic qualitative measures such as the degree of coherence.

Acknowledgements

This research was supported by Takahashi Industrial and Economics Foundation (<http://takahashi-f.or.jp/>).

References

- [1] T.H.K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C.L. Zitnick, D.

- Parikh, L. Vanderwende, M. Galley, and M. Mitchell, “Visual storytelling,” *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, California, pp.1233–1239, Association for Computational Linguistics, June 2016.
- [2] J. Tang, J. Wang, Z. Li, J. Fu, and T. Mei, “Show, reward, and tell: Adversarial visual story generation,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol.15, no.2s, July 2019.
- [3] J. Wang, J. Fu, J. Tang, Z. Li, and T. Mei, “Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training,” *The AAAI Conference on Artificial Intelligence (AAAI)*, 2018., pp.7396–7403, Feb. 2018.
- [4] D. Gonzalez-Rico and G. Fuentes-Pineda, “Contextualize, Show and Tell: A Neural Visual Storyteller,” *arXiv e-prints*, arXiv:1806.00738, June 2018.
- [5] T. Kim, M.O. Heo, S. Son, K.W. Park, and B.T. Zhang, “GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation,” *arXiv e-prints*, arXiv:1805.10973, May 2018.
- [6] J. Li, H. Shi, S. Tang, F. Wu, and Y. Zhuang, “Informative visual storytelling with cross-modal rules,” *Proceedings of the 27th ACM International Conference on Multimedia, MM ’19*, New York, NY, USA, pp.2314–2322, Association for Computing Machinery, 2019.
- [7] P. Yang, F. Luo, P. Chen, L. Li, Z. Yin, X. He, and X. Sun, “Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling,” *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp.5356–5362, International Joint Conferences on Artificial Intelligence Organization, 7 2019.
- [8] L.A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, “Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data,” *arXiv e-prints*, arXiv:1511.05284, Nov. 2015.
- [9] C.C. Park and G. Kim, “Expressing an image stream with a sequence of natural sentences,” in *Advances in Neural Information Processing Systems 28*, ed. C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, pp.73–81, Curran Associates, Inc., 2015.
- [10] B. Wang, L. Ma, W. Zhang, W. Jiang, and F. Zhang, “Hierarchical Photo-Scene Encoder for Album Storytelling,” *arXiv e-prints*, arXiv:1902.00669, Feb. 2019.
- [11] L. Yu, M. Bansal, and T.L. Berg, “Hierarchically-Attentive RNN for Album Summarization and Storytelling,” *arXiv e-prints*, arXiv:1708.02977, Aug. 2017.
- [12] C.C. Hsu, S.M. Chen, M.H. Hsieh, and L.W. Ku, “Using Inter-Sentence Diverse Beam Search to Reduce Redundancy in Visual Storytelling,” *arXiv e-prints*, arXiv:1805.11867, May 2018.
- [13] X. Wang, W. Chen, Y.-F. Wang, and W.Y. Wang, “No metrics are perfect: Adversarial reward learning for visual storytelling,” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, pp.899–909, Association for Computational Linguistics, July 2018.
- [14] Q. Huang, Z. Gan, A. Celikyilmaz, D. Wu, J. Wang, and X. He, “Hierarchically Structured Reinforcement Learning for Topically Coherent Visual Story Generation,” *arXiv e-prints*, arXiv:1805.08191, May 2018.
- [15] N. Li, B. Liu, Z. Han, Y.S. Liu, and J. Fu, “Emotion reinforced visual storytelling,” *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR ’19*, New York, NY, USA, pp.297–305, Association for Computing Machinery, 2019.
- [16] Y. Liu, J. Fu, T. Mei, and C.W. Chen, “Let your photos talk: Generating narrative paragraph for photo stream via bidirectional attention recurrent neural networks,” *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI’17*, pp.1445–1452, AAAI Press, 2017.
- [17] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol.28, no.3/4, pp.321–377, 1936.
- [18] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” *Proc. 30th International Conference on*

- Machine Learning, ed. S. Dasgupta and D. McAllester, Proceedings of Machine Learning Research, vol.28, Atlanta, Georgia, USA, pp.1247–1255, PMLR, 17–19 June 2013.
- [19] W. Liu, J.L. Qiu, W.L. Zheng, and B.L. Lu, “Multimodal Emotion Recognition Using Deep Canonical Correlation Analysis,” arXiv e-prints, arXiv:1908.05349, Aug. 2019.
- [20] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.
- [21] N. Martin and H. Maes, Multivariate analysis, Academic press London, 1979.
- [22] R.-G. Cirstea, D.-V. Micu, G.-M. Muresan, C. Guo, and B. Yang, “Correlated time series forecasting using multi-task deep neural networks,” Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, New York, NY, USA, pp.1527–1530, Association for Computing Machinery, 2018.
- [23] R. Wan, S. Mei, J. Wang, M. Liu, and F. Yang, “Multivariate temporal convolutional network: A deep neural networks approach for multivariate time series forecasting,” Electronics, vol.8, no.8, 2019.
- [24] M. Schuster and K.K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Transactions on Signal Processing, vol.45, no.11, pp.2673–2681, 1997.
- [25] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.P. Morency, “Tensor Fusion Network for Multimodal Sentiment Analysis,” arXiv e-prints, arXiv:1707.07250, July 2017.
- [26] Z. Liu, Y. Shen, V.B. Lakshminarasimhan, P.P. Liang, A.B. Zadeh, and L.-P. Morency, “Efficient Low-rank Multimodal Fusion with Modality-Specific Factors,” arXiv e-prints, arXiv:1806.00064, May 2018.
- [27] H. Babii, A. Janes, and R. Robbes, “Modeling Vocabulary for Big Code Machine Learning,” arXiv e-prints, arXiv:1904.01873, April 2019.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770–778, June 2016.
- [29] D.P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv e-prints, p.arXiv:1412.6980, Dec. 2014.
- [30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” arXiv e-prints, p.arXiv:1912.01703, Dec. 2019.
- [31] J.D. Owens, M. Houston, D. Luebke, S. Green, J.E. Stone, and J.C. Phillips, “Gpu computing,” Proceedings of the IEEE, vol.96, no.5, pp.879–899, 2008.
- [32] M. Denkowski and A. Lavie, “Meteor universal: Language specific translation evaluation for any target language,” Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland, USA, pp.376–380, Association for Computational Linguistics, June 2014.
- [33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, USA, p.311–318, Association for Computational Linguistics, 2002.
- [34] R. Vedantam, C.L. Zitnick, and D. Parikh, “Cider: Consensus-based image description evaluation,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4566–4575, June 2015.
- [35] C.Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” Text Summarization Branches Out, Barcelona, Spain, pp.74–81, Association for Computational Linguistics, July 2004.
- [36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and Tell: A Neural Image Caption Generator,” arXiv e-prints, arXiv:1411.4555, Nov. 2014.
- [37] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” arXiv e-prints, arXiv:1502.03044, Feb. 2015.
- [38] X. Wang, Y.-F. Wang, and W.Y. Wang, “Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning,” Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, Louisiana, pp.795–801, Association for Computational Linguistics, June 2018.
- [39] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” Journal of Machine Learning Research, vol.9, pp.2579–2605, 2008.

natural language processing, and deep learning.

Rizal Setya Perdana received his B.Sc. degree in Informatics Engineering from the Brawijaya University and M.S. degrees in Informatics Engineering from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia in 2013 and 2015, respectively. He is currently a Doctoral student at Department of Computer Science and Engineering, Toyohashi University of Technology, Toyohashi, Japan. His research interests include multimodal learning representation, image captioning, visual storytelling,

Yoshiteru Ishida received the BA Eng., MA Eng., and Dr. Eng. degree in Applied Mathematics and Physics from Kyoto University, Japan, in 1979, 1981, and 1986, respectively. He joined the Department of Applied Mathematics and Physics and the Division of Applied System Science, Kyoto University, as Assistant Professor from 1983 to 1986 and 1986 to 1994, respectively. In April 1986 and 1993, he became a Visiting Researcher at School of Computer Science and Department of Psychology, Carnegie-Mellon University. From April 1994 to February 1998, he joined the Graduate School of Information Science, Nara Institute of Science and Technology. Since June 1998 until 2010, he became a Professor at Department of Knowledge-based Information Engineering, Toyohashi University of Technology. Currently, he is a Professor at Department of Computer Science and Engineering, Toyohashi University of Technology. He is a member of ACM, IEICE, IPSJ, SICE and JFES.