IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 29, 2021, accepted August 8, 2021, date of publication August 12, 2021, date of current version August 19, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104276

Deep Learning-Based Short Story Generation for an Image Using the Encoder-Decoder Structure

KYUNGBOK MIN, MINH DANG, AND HYEONJOON MOON
Department of Computer Science and Engineering, Sejong University, Seoul 05006, South Korea

Corresponding author: Hyeonjoon Moon (hmoon@sejong.ac.kr)

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540) and National Research Foundation of Korea (NRF) grant funded by the Korea government, Ministry of Science and ICT (MSIT) (2021R1F1A1046339).

ABSTRACT Research that applies artificial intelligence (AI) to generate the captions for an image has been extensively studied in recent years. However, the length of these captions was short, and the number of generated captions was limited. In addition, it is unknown whether a short story can be generated based on the image, because many sentences have to be connected to create a fluent short story. As a result, this study introduces an encoder-decoder framework structure to generate a short story captioning (SSCap) using a common image caption dataset and a manually collected story corpus. This manuscript has three main contributions, which include 1) an unsupervised deep learning-based framework that combines a recurrent neural network (RNN) structure and encoder-decoder model for composing a short story for an image, 2) a huge story corpus, which includes two different genres (horror and romantic), manually collected and validated. Extensive experiments demonstrated that short stories created by the proposed model show creative content compared to existing systems that can only make concise sentences. Therefore, the demonstrated framework has the potential to motivate the development of a more robust AI story writer and motivates the integration of the suggested model into practical applications to help the story writers find a new idea.

INDEX TERMS Image caption, story teller, deep learning, computer vision, context awareness.

I. INTRODUCTION

Artificial intelligence (AI) systems have become more advanced and even outperformed humans in specific tasks during the last decades. They have been increasingly applied to the domains that are previously considered human territories, such as writing novels, composing music, and communicating. As a result, the idea of adopting AI to compose the artworks, such as photos [1], [2], stories [3], and movie scenarios [4] has become a hot topic. Compared to the other domains, the story or narrative is one of the earliest and fundamental means for humankind to transfer knowledge to the following generations. The story is highly linked to a series of related events, experiences, or incidents, whether real or fictitious. Therefore, this study focuses on the story domain and aims to apply the image caption generation technique to compose a short story inspired by the input image's content.

The associate editor coordinating the review of this manuscript and approving it for publication was Longzhi Yang.

Recently, the number of images that were shared on the internet has increased significantly due to the latest advancements in technology, such as online news, advertisements, and social networks. Humans are able to understand images using the context without their description, even though some images do not contain any description or explanation. On the other hand, machines cannot understand or support humans to interpret some group of image captions without being taught through sufficient datasets. Image captioning is a trending Artificial Intelligence (AI) topic that works on interpreting and providing textual descriptions for an image, which can be applied to both computer vision (CV) and natural language processing (NLP). Computers have perceived and produced more accurate descriptions for an image due to the advancements in CV and NLP over the last decade [5], [6]. Image captioning plays a critical role in various real-world applications, such as social networking, Human-computer interaction (HCI), and applications for people with vision loss. An image caption generation system

is expected to describe the textual descriptions of an image and generate captions with appropriate linguistic meaning. However, the current image captioning systems concentrated on presenting the visual content of the image and describing actual details and information. In contrast, linguistics, such as semantics and morphology, are crucial parts of human language that show personal characteristics, feelings, and attitudes. The multi-style image captioning, which focuses on the linguistic aspects rather than the plain descriptions of the image, was in the early development stage [7].

Figure 1 demonstrates a standard image caption and a horror-style short story that is generated by the proposed model for an input image. The short story showed the artistic aspect, significantly increased the attractiveness, and stimulated the user's involvement and social relation.

The idea of creating a short story, which is based on the context of an input image, which can be regarded as a ''visual story writer,'' is an exciting and challenging topic. The main objective is to change the previous image captioning approach to short story generation with the image's context as the primary source of information to create the stories. The system requires a contextual description and a complicated language model to make the story related to the image's context from the prospect of the language consistency. The existing studies mainly concentrated on creating a short caption to describe the image thoroughly, such as what objects are in the image and how the objects interact [8], [9]. A typical approach to image captioning is to combine a convolutional neural network (CNN) model and a recurrent neural network (RNN) model. The CNN is first implemented to obtain the abstract features from the input images. RNN and its extensions are then used to generate the caption using the extracted features from CNN. The ''visual story writer'' is more complicated than the common sequence-to-sequence problem due to the challenge of learning a huge visual variance and maintaining the long-term dependency and consistency among several sentences.

This work considers the ''visual story writer'' topic an unsupervised sequence-to-sequence problem to efficiently address the ''visual story writer'' topic. The image caption generation is the nearest related topic to the ''visual story writer''. The model receives any image as an input, and then several captions that describe the image's content are generated. These captions are then used to create a sequence of sentences to make a short story as the output. An unsupervised sequence-to-sequence gated recurrent unit (GRU) framework is proposed to generate a short story of different genres using the predicted captions from an input image. The model creates stories, sentence by sentence, concerning the image context and the previously generated sentences. The proposed framework consists of an RNN decoder that is trained on the collected story datasets. After that, the sentences extracted from the datasets are mapped to a skip-thought vector. The RNN model is then conditioned on the skip-thought vector to create the encoded sentences. Finally, a visual-semantic embedding model was used to link conceptual captions and image datasets into a uniformed vector space. The approach's main goal was to force the system to create stories that include more narrative and consistent language.

The research content is separated into six primary sections. Section 2 briefly describes the current trends of caption generation and the sequence-to-sequence topics. After that, Section 3 introduces a huge story dataset containing two different genres, which were collected manually. The model that is introduced to generate the short story is explained in Section 4. The extensive experiments, which are carried out to evaluate the suggested framework's performance, are described in Section 5. Finally, the main contents of the research, remaining limitations, and some future approaches are discussed in Section 6.

II. RELATED WORK

A. IMAGE CAPTION GENERATION

Image caption generation is the problem of creating a sequential textual description for a non-sequential input image, which is different from the image classification problem that returns the non-sequential class index outputs [5]. The existing image captioning work can be categorized into two primary groups, which include the retrieval-based approach and the template-based approach. The template-based method first detects the objects, object attributes, and object interactions to fill the blank slots of the fixed templates. On the other hand, the retrieval-based approach retrieves the captions from a collection of existing captions. The captions of the visually similar images to the input image are first identified from the training set, and then they are connected using sentence templates or grammar rules. There are some apparent limitations of the mentioned approaches because hardcoded rules and manually selected features were used to produce the output text. The template-based methods have a restricted number of available visual models, and the generated sentences have limited coverage, creativity, and complexity. The retrieval-based methods fail to generate image-specific and semantically accurate captions.

The huge successes of deep learning in a wide range of application domains in the last decades have led to the introduction of many deep learning-based image caption generation frameworks that achieved remarkable results. In addition, the availability of various benchmark datasets for the image caption generation, as described in [10], such as COCO from Microsoft [11], SBU1M captions dataset [12], Deja Image-Captions [13], also motivated the development of the deep learning-based image caption generation systems. There are four primary deep learning-based image captioning trends, which include multimodal learning, encoder-decoder framework, compositional architecture, and attention-guided network.

Deep learning-based caption generation models introduced by [5], [7] proved that they could produce a human-level caption for an image. The framework presented by the two research consisted of pre-trained deep learning models (VGG [14], or AlexNet [15]) to obtain the abstract features

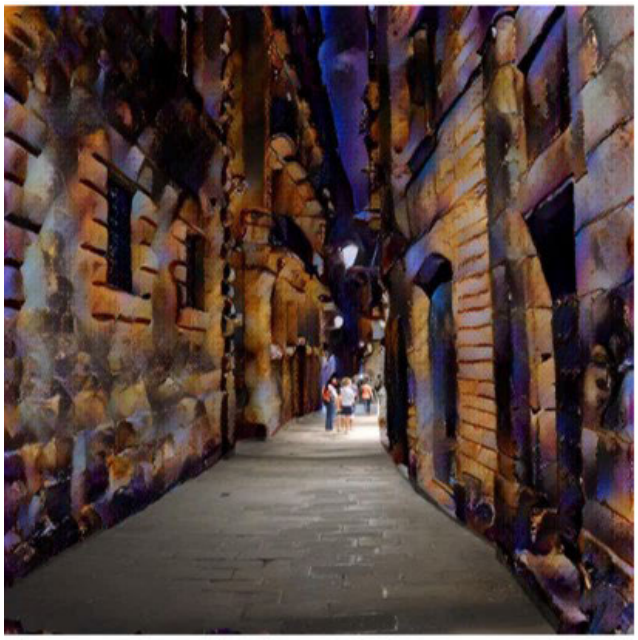


Image captions

A picture of a several individuals walking on a sidewalk in between buildings.

Visual story writer

My heart is beating so fast it is a bit shorter than my breathing. I think I'm being stalked. I don't know what happened to me, or why. I was **scared** of it, but I need to know why. I need to know what happened. Because everything is different now, yet I can't know how, I feel it. People act strangely in the **dark alley** and I feel unsafe, because I'm at a dark and unexplained scene.

FIGURE 1. An image caption and a horror-style story for an input image that was generated by the proposed model.

from the image, which is about to be captioned. After that, the features taken from the final FC layer were fed to the caption creator, which applies natural language processing technique and RNN to create the captions. These models described the caption generation process as the prediction of the next word in a sentence. Recent research on image captioning has mainly focused on describing subjects and actions in the image [8], [9]. Among them, the attention-guided models have attracted huge interest from the research community recently compared to the other approaches because of their state-of-the-art performance [16].

B. STORY-STYLE MULTIMEDIA CONTENT RETRIEVAL

Research in multimedia content retrieval focuses on the perception of narrative to enhance indexing and retrieval effectiveness. For example, the system proposed by [17] aligned crowdsourced plot synopses with shots from the videos for story-style content retrieval. The authors applied a similarity function to the sentences in the plot synopses and shots using personal identities and keywords in the captions. In [18], the authors implemented a multiple neural network structure using single and pairwise element-based predictions and exploited both text and image features. The obtained results proved that the proposed model could learn exciting aspects of common temporal sense. Although many efforts have been made to understand the narrative structure in video media, few studies have examined the possibility of utilizing the notion of narrative structure in still images. Recently, Gaur exploited and analyzed hashtags from an image to create significant anecdotes to connect to the image's essence [19]. They used a dual attention-based encoder-decoder model to create hashtags for an input image. After that, a character-level language model based on a multi-layer RNN model was trained in order to produce narratives using one of the created hashtags. Another model demonstrated by Alexander in [20] can create image captions that were visually descriptive and appropriately. The authors tried to divide semantics and style. NLP techniques and frame semantics were employed to generate concise semantic term representation. Moreover, they implemented a unified language model that decoded sentences with diverse word choices and syntax for various styles.

Our work is distinguishable from previous research in one crucial aspect, in which we deal with a more complex area of story/caption alignment. Unlike caption generation that only describes the main content of an image, the story is more verbose and might vary because it is only based on the image context. Moreover, the three-step process is introduced to create a story with the style transfer efficiently performed based on the skip-thought model.

III. DATASET

A. DATASET DESCRIPTION

The proposed research uses two separate datasets: (1) Books downloaded from the Smashwords website, which is a website where the authors share their unpublished books with the community. The books that contained over 20,000 words were crawled to reduce noise and too short stories. The dataset contains stories in 2 distinctive genres, romance (500 stories) and horror (621 stories). (2) Conceptual captions dataset, which is a huge captions dataset introduced by Google in 2018 [21]. It contains over 3.3 million pairs of images and captions, which were created by automatically filtering and extracting the caption annotations from the internet. The dataset includes an increased magnitude of captioned images compared to the human-curated MS-COCO

dataset [11]. Moreover, the dataset contains a broader diversity of image-caption pair, because the images and annotations were downloaded from billions of web pages, enabling better performance of image captioning models, which help the proposed model generate accurate captions.

B. DATASET PREPARATION

1) BOOK DATASET

An automated python crawler was created to download the books in PDF format, so we apply a python pdf2txt library to convert them into plain text. We then used the python NLP library to groom the data by removing blank rows and finally save all contents into one file.

2) CONCEPTUAL CAPTIONS DATASET

All sentences in the dataset were converted into lowercase, non-alphanumeric characters were then discarded. Finally, the words that appear over fuve times in the dataset were filtered out.

IV. METHODOLOGY

In this section, a thorough explanation of the proposed *visual story writer* model is explained thoroughly. The implementation process of the introduced model is described in Figure 2, which includes three main components 1) train both visual-semantic alignment model and skip-thought encoder on the conceptual captions dataset, 2) extract skip-thought vectors from the collected stories dataset, and 3) perform the deep style transfer to transform the caption style into the story style using the skip-thought decoder.

A. IMAGE CAPTION GENERATION

This section describes the image caption generation process for an input image based on the deep visual-semantic alignments model [7], which contains an encoder and a decoder. Firstly, a region convolutional neural network (RCNN) is applied to map the image regions and words into a joint, multimodal embedding in the encoder side. A recurrent neural network (RNN) is then implemented on the decoder side to determine the embedding representations to allow semantically related concepts across the words and image regions to be near each other.

1) IMAGE REPRESENTATION

The RCNN model used is pre-trained on the ImageNet, which involves 1000 object classes [15]. It was then fine-tuned on 200 unique object classes of the ImageNet Object Localization Challenge [15]. Following [7], the top 19 localized positions and the whole image were used to calculate the representations using the pixels I_b of each localized bounding box. The RCNN parameters θ_c include roughly 58 million parameters.

$$v = W_m[RCNN_{\theta_c}(I_b)] + b_m \tag{1}$$

where $RCNN(I_b)$ converts the pixels I_b inside the localized bounding boxes into a 2048-dimensional activation of the fully connected layer. The matrix W_m has the dimension of $h \times 2048$, where h is the multimodal embedding space size. Each image is described as a collection of h-dimensional vectors $v_i|i = 1, \ldots, 20$.

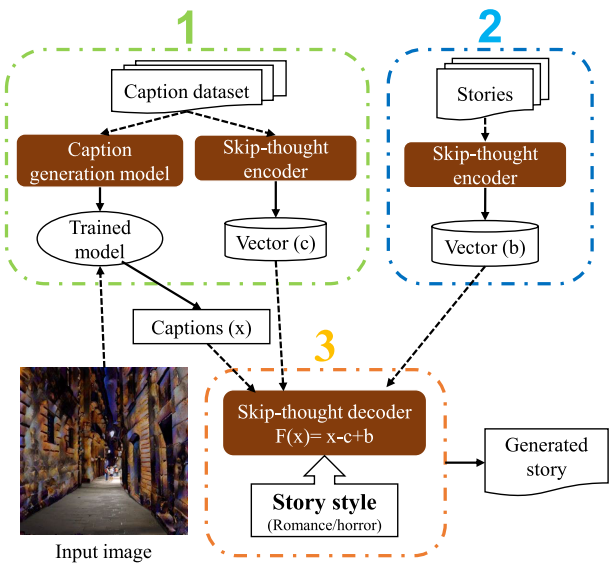


FIGURE 2. Overall system architecture of the visual story writer model.

2) SENTENCE REPRESENTATION

The RNN model receives a sequence of N words and converts them into an h-dimensional vector. In addition, each word's representation is enhanced by an unfixed-sized context surrounding it. The index $t = 1, \ldots, N$ is used to indicate the word location in a sentence. The equation of the RNN is described below.

$$x_t = W_x \prod_t \tag{2}$$

$$e_t = f(W_e x_t + b_e) \tag{3}$$

$$h_t^f = f(e_t + W_f h_{t-1}^f) + \tag{4}$$

$$h_t^b = f(e_t + W_b h_{t-1}^b) + b_b \tag{5}$$

$$s_t = f(W_d(h_t^f + h_t^b) + b_d) \tag{6}$$

where $\prod_t$ is a column vector that holds a single one at the index of the t-th word in a word vocabulary. The weights W_w define a word embedding matrix with fixed 300-dimensional word2vec weights to prevent overfitting. The RNN contains two separate streams, where the first stream moves from left to right (h_t^f) and the other stream moves from right to left (h_t^b). The final h-dimensional representation s_t for the t-th word is a function of both the word's surrounding context and the word at that location in the sentence. In general, s_t is a function that represents a collection of words in a sentence. In addition, the parameters W_e, W_f, W_b, W_d and the respective biases b_e, b_f, b_d were learned. Rectified linear unit (ReLU) was used as the activation function f, it can be calculated as $f : x \rightarrow max(0, x)$.

B. SKIP-GRAM SENTENCE ENCODER-DECODER

This section discusses a bi-directional sentence skip-gram model, which was motivated by the skip-gram structure [22] to learn the word representations based on the distributional theory. The sentences with the same context and syntax are expected to be syntactically and semantically related. Therefore, they can be encoded using an identical vector. The trained encoder can be trained to map a sentence to vector representations. Their similarity score is then computed through an inner product. The encoder depends heavily on contiguous text, where sentences in the training data follow one another in order. Therefore, we collected a large collection of books to acquire overall text representations. For example, Table 1 represents the nearest neighbors of input sentences that are not in the training data. The generated sentences confirmed that our intuition is correct because the generated sentences semantically and syntactically resembled the input sentences.

Given a sentence tuple (s_{i-1}, s_i, s_{i+1}), the model tries to encode the sentence s_i into a fixed vector. After that, based on the generated vector, it attempts to create the previous sentence s_{i-1} and the next sentence s_{i+1}. Let w_i^t indicates the $t-th$ word for the sentence s_i and let x_i^t be its word embedding. We divide the system into three main sections, which include the objective function, the encoder, and the decoder.

1) ENCODER

Suppose that $w_i^1, \ldots, w_i^N$ depicts all N words w in the current sentence s_i. The encoder task is to create a hidden state h_i^t for each time step t that encode the sequence of words from w_i^1 to w_i^t. Therefore, the last hidden state h_i^N of the sentence s_i will encode the entire contents. GRU model of the encoder generates the current hidden state h_i^t by linearly adding the previous hidden state x_i^{t-1} and the newly proposed state $\tilde{h}_i^t$

$$h_i^t = z_i^t \odot h_i^{t-1} + (1 - z_i^t) \odot \tilde{h}_i^t \tag{7}$$

where $\odot$ is element-wise multiplication and z_i^t is the update gate. The update gate z_i^t is applied to decide how much past information needs to be kept in the future, which is calculated by a linear sum between the existing state and the newly proposed state.

$$z_i^t = \sigma(W_z x_i^t + U_z h_i^{t-1} + b_i^z) \tag{8}$$

σ is the logistic nonlinearity, and $W_z(n \times m)$, $U_z(n \times m)$, $b_z(n \times 1)$ are fixed-sized parameters belong to the update gate (two weights and bias). Thanks to the sigmoid activation function, the update gate output value is in the range of zero to one. If it is one, the update gate will completely forget previous hidden states, which means $h_i^t = \tilde{h}_i^t$. On the other hand, if it is zero, then all hidden states from previous time steps will be copied over, that is $h_i^t = h_i^{t-1}$. The proposed state $\tilde{h}_i^t$ is computed equivalent to the conventional recurrent unit

$$\tilde{h}_i^t = tanh(W x_i^t + U(r_i^t) \odot h_i^{t-1}) + b_i^t) \tag{9}$$

where r_i^t is a reset gate. The reset gate r_i^t is calculated similar to the update gate but based on a set of different parameter values: $W_r(n \times m)$, $U_r(n \times m)$, $b_r(n \times 1)$.

$$r_i^t = \sigma(W_r x_i^t + U_r h_i^{t-1} + b_i^r) \tag{10}$$

When the status is off (r_i^t equals to 0), the reset gate controls the newly proposed state by working on the first word in a sequence of words, which enable it to eliminate the state $h_i^{(t-1)}$ that were computed previously.

When the status is off (r_i^t equals to 0), the reset gate controls the newly proposed state by working on the first word in a sequence of words, which enable it to eliminate the state $h_i^{(t-1)}$ that were computed previously. Although GRU preserves the memory remarkably better than RNN, it just analyzes the forward lingual context and cannot examine the backward context. Therefore, the learning process of GRU is considered partially completed, because the context of a word is influenced by its forward and backward contexts in a sentence. As a result, this paper proposes to apply a two-stream gated recurrent unit to cope with the mentioned problem, which is an upgraded version of GRU that was motivated by the bidirectional recurrent neural networks (BRNNs) in [4]. The model contains two main networks. The first network process the forward context, whereas the other network analyzes the backward context. The equations for the hidden state, the proposed hidden state, the update gate, and the reset gate of the TGRU, are described as follows.

Forward pass:

$$\vec{z}^t = \sigma(\vec{W}_z x^t + \vec{U}_z h^{t-1}) \tag{11}$$
$$\vec{r}^t = \sigma(\vec{W}_r x^t + \vec{U}_r h^{t-1}) \tag{12}$$
$$\vec{\tilde{h}}_t = tanh(\vec{W}^d x^t + \vec{U}^d(\vec{r}^t \odot \vec{h}^{t-1})) \tag{13}$$
$$\vec{h}^t = \vec{z}^t \odot \vec{h}^{t-1} + (1 - \vec{z}^t) \odot \vec{\tilde{h}}^t \tag{14}$$

The backward pass was added as an additional module of the model to exploit valuable features that were missed in the forward pass.

$$\overleftarrow{z}^t = \sigma(\overleftarrow{W}_z x^t + \overleftarrow{U}_z h^{t-1}) \tag{15}$$
$$\overleftarrow{r}^t = \sigma(\overleftarrow{W}_r x^t + \overleftarrow{U}_r h^{t-1}) \tag{16}$$
$$\overleftarrow{\tilde{h}}_t = tanh(\overleftarrow{W}^d x^t + \overleftarrow{U}^d(\overleftarrow{r}^t \odot \overleftarrow{h}^{t-1})) \tag{17}$$
$$\overleftarrow{h}^t = \overleftarrow{z}^t \odot \overleftarrow{h}^{t-1} + (1 - \overleftarrow{z}^t) \odot \overleftarrow{\tilde{h}}^t \tag{18}$$

TABLE 1. Two nearest neighbors of the input sentences that are not in the training data.

Input	Output
Lights dimmed the city, and I saw a flash of light in the distance.	It was so much more than that I could hardly believe what had happened to me. After the sun broke out on the horizon, it was as if he were in New York City, and I had no idea what to do with her. It seemed to be the most memorable night of my life
I can clearly observe that abandoned apartment from my front door.	I noticed a human body hung on a nearby tree, and it resembles a bunch of fresh bones. I am brave. I told myself. What happened? How am I still sane? I have yet to know how to manage this circumstance if it is real.

2) DECODER

The computational process of the decoder is similar to the encoder, which is based on the encoder's output h_i for the sentence s_i. The introduction of C_z, C_r, and C is used as a bias for the hidden state, update gate, and reset gate, which is computed by the encoder's output vector h_i.

Two decoders are concurrently computed based on the encode's output. The first one is applied to create the previous sentence s_{i-1}, and the other one is used to generate the next sentence s_{i+1}. Although the two decoders utilize distinctive parameter sets to calculate their hidden states, they are based on the vocabulary dictionary V that uses a hidden state to calculate the distribution over words. As a result, the two decoders are similar to an RNN language model, but they rely on the encoder's output h_i.

Assume that $\vec{h}^t$ indicates the hidden state of the decoder of the next sentence s_{i+1} at time step t. The update gate, reset gate, and hidden state of the s_{i+1} decoder is given as below.

$$z^t = \sigma(W_z x^{t-1} + U_z h^{t-1} + C_z h_i) \tag{19}$$

$$r^t = \sigma(W_r x^{t-1} + U_r h^{t-1} + C_z h_i) \tag{20}$$

$$\tilde{h_t} = tanh(W^d x^{t-1} + U^d(r^t \odot h^{t-1}) + Ch_i) \tag{21}$$

$$h^t = z^t \odot h^{t-1} + (1 - z^t) \odot \tilde{h^t} \tag{22}$$

where $x^{(t-1)}$ is the word embedding of the prior word when the decoder determine the current word using the context of the previous word. Given $h^t_{(i+1)}$ represents both ($\vec{h}^t$ and ($\overleftarrow{h^t}$), the probability of the word $w^t_{(i+1)}$ is based on the encoder vector h_i and the previous word $t-1$ is computed as follows.

$$P(w^t_{i+1}|w^{<t}_{i+1}, h_i) \propto exp(v_{w^t_{i+1}} h^t_{i+1}) \tag{23}$$

where $v_{w^t_{i+1}}$ indicates the row of V, which correspond to the word of w^t_{i+1}.

Objective

Given $(s_{(i-1)}, s_i, s_{(i+1)})$, the main goal is the sum of log-probabilities for the next and previous sentences conditioned on the representation of the encoder.

$$\sum_t logP(w^t_{i+1}|w^{<t}_{i+1}, h_i) + \sum_t logP(w^t_{i-1}|w^{<t}_{i-1}, h_i) \tag{24}$$

The complete objective is the sum of such training tuples. Adam algorithm [23] was applied to perform the optimization.

C. DEEP STYLE TRANSFER

In order to generate the final story with a specific style, a function that joins the gap between passages in novels and their retrieved image captions is required. It can be defined as function F that reflects image caption vectors X to a story content vector $F(x)$, which is then $F(x)$ fed into the decoder to get the final outputs. There is no existing mechanism described above, so F needs to be constructed differently.

F is assumed to contain a caption x, a "caption style" vector c, and a "story style" vector b, which can be explained as keeping the caption's context, and replace the captions with a story. $F(x)$ can be defined as below.

$$F(X) = x - c + b \tag{25}$$

where c is the mean of the skip-thought vectors used to train the conceptual captions. b is the mean of the skip-thought vectors used for romance story paragraphs that have a length greater than 100.

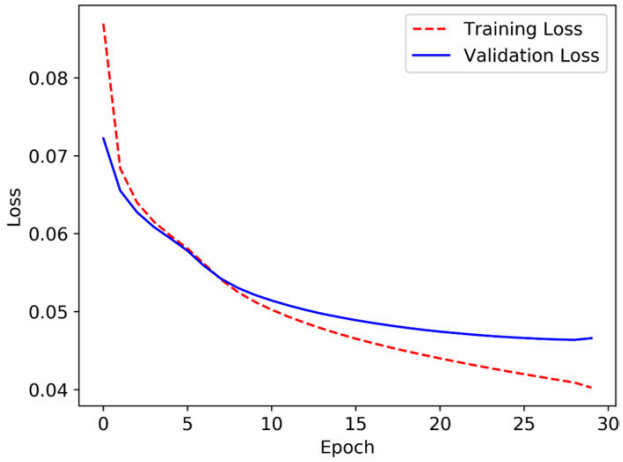


FIGURE 3. Loss curves of the training and validation processes.

V. EXPERIMENTAL RESULTS

A. DEVELOPMENT ENVIRONMENT

NVIDIA DIGITS toolbox, which was pre-installed an Ubuntu 16.04, was used to train the deep learning models in this research. It has an Intel Core i7-5930K processor, 64GB of DDR4 RAM, and four Titan X 12GB GPUs.

B. IMAGE CAPTION GENERATION RESULTS

From the conceptual captions dataset, 75% of the data were randomly used to train the caption generation model. The remaining 25% of the data was applied to validate the trained model. The training and validation process losses are represented in Figure 3, where the dotted red and the solid blue lines depict the loss of the training process and validation process, respectively. The training and validation losses are generally stable, with the validation loss steady at about 0.048 at epoch 30th.

C. STORY GENERATION RESULTS

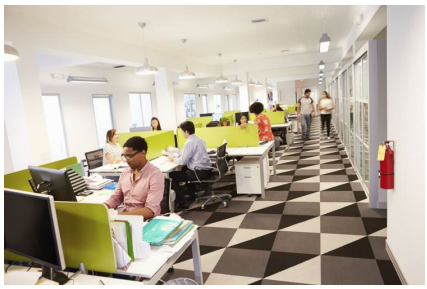
Figure 4 shows two examples of the generated stories, which include both romance stories and horror stories. The results prove that our model can create vivid stories for images. Notably, the model used deep visual-semantic alignment to create a short caption for an image. After that, this caption is used as a query to transfer the horror or romance story style to the caption using the proposed sentence encoder-decoder model. The retrieved stories are semantically understandable, and they are also based on the general content of the image.

A slightly more exciting idea is shown in Figure 5, we apply our model on the images generated by computer

A calm ocean view on a clear day day

My beach looked up to the shore, at the exact same moment. The sun had begun to rise and fall on the surface of my mind, and I breathed a sigh of relief. I don't really care what it was like to fall in love with him, so that would be the most important thing in the world.

After turning off all the lights and going to sleep, I had a horrible nightmare. I awoke in a pool of red liquid with the scent of blood. I was so sleepy, I could not move, and I could not swim anymore. I let out a gasp, and I then hear a thump thump sound out of nowhere.



In a classroom of students, some study and others socialize

They became uncomfortably aware of everyone in the classroom, and I began to study them. I was having a hard time meeting with my coworkers, but I had no idea what to do. As long as she was gone, I couldn't help but notice that Dr. Franklin Terzini stood at the end of the hall, looking for a room full of books and computers.

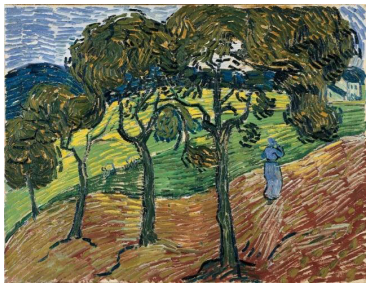
When I sat in the corner I heard a loud scraping sound coming from inside the office. I thought it was just a trick of the camera, but I saw the reflection which was still underneath the desk in the middle of the room. I was hoping that it was just my imagination.

FIGURE 4. Two examples of the story generated by the proposed model for normal photos (Red text indicates the original caption, the blue text shows the romance story, and the black text is the horror story that was created by the proposed model).

A crowded city filled with people and buildings

We drifted into the city , but I couldn't quite block out the crowds. The streets were crowded with people, and I had no idea what was going to happen. In fact, it seemed to be my own personal space. As soon as they walked through the city streets, I let out a soft sigh of relief.

It was a cold winter night and i went outside to have a smoke. As I step outside I feel something is wrong. I look around and can see a shadow all around me. The air around me gets cold as the shadow tries to embrace me. I try to run but it's too late, I'm traped in darkness.



A misty green painting of birds on a tree branch before the ruins of a castle

My inner landscape was on the ground next to it. As long as the sun rose above the horizon, I couldn't help but notice that I had fallen in love with him. He looked like a thousand years old, and as he pulled out of my line of vision, he seemed to be the most important thing in the world.

The face of a dead man slowly appeared on the photograph of the dead tree. It was difficult to make out anything, but I could see the color of the dead fabric and the mangled flesh of the child and the chains around his neck. Someone must have discovered the man while he was murdering the child.

FIGURE 5. Two examples of the story generated by the proposed model for artistic pictures (Red text indicates the original caption, the blue text shows the romance story, and the black text is the horror story that was created by the proposed model).

from the research proposed by [24]. The created captions prove that the deep visual-semantic alignment works well in describing the general contents from the images, even if the images are generated by the computer.

VI. CONCLUSION

The final goal of AI/Deep learning is to support humans in several areas without the need for monitoring continuously. To achieve this goal, many researchers have focused on creating a smarter generation of AI by using the power of deep learning. In the field of art and music, AI-based programs have been developed to reach the human level. However, in the area of writing, programs similar to humans have not yet been established, despite many attempts because writing is intuitively unrecognizable by the computer. Moreover, different languages have different grammar, words, and nuances,

which are hard to overcome by AI. This study introduces a three-step process to create a story from an input photo. The proposed model uses NLP and encoder-decoder structure to generate the stories under different genres. In addition, we also collected a large dataset of novels for each romance and horror genre. Finally, the generated stories are related to the input pictures in different ways.

There are many potential approaches to improve this topic. For example, grammar and detail context modules were not considered in the study. Those two functions can make the stories more natural and make the story longer without missing the narrative. Another weakness of the system is the deep learning model. Only a basic CNN model with the GRU method was implemented, but it was originally developed for image processing. In order to make a more advanced story creator, a more suitable machine learning model needs to be developed and integrate into the NLP methods in processing the training data in the future.

REFERENCES

[1] L. Dang, S. Hassan, S. Im, J. Lee, S. Lee, and H. Moon, "Deep learning based computer generated face identification using convolutional neural network," *Appl. Sci.*, vol. 8, no. 12, p. 2610, 2018.

[2] L. M. Dang, S. I. Hassan, S. Im, and H. Moon, "Face image manipulation detection based on a convolutional neural network," *Expert Syst. Appl.*, vol. 129, pp. 156–168, Sep. 2019.

[3] N. Peng, M. Ghazvininejad, J. May, and K. Knight, "Towards controllable story generation," in *Proc. 1st Workshop Storytelling*, 2018, pp. 43–49.

[4] Y. Chen, Y. Wang, D. S. Kirschen, and B. Zhang, "Model-free renewable scenario generation using generative adversarial networks," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3265–3275, May 2018.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[6] C. Wang, H. Yang, and C. Meinel, "Image captioning with deep bidirectional LSTMs and multi-task learning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 2, pp. 1–20, May 2018.

[7] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.

[8] X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2422–2431.

[9] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.

[10] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, and B. Plank, "Automatic description generation from images: A survey of models, datasets, and evaluation measures," *J. Artif. Intell. Res.*, vol. 55, pp. 409–442, Jan. 2016.

[11] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*. [Online]. Available: https://arxiv.org/abs/1504.00325

[12] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.

[13] J. Chen, P. Kuznetsova, D. Warren, and Y. Choi, "Déja image-captions: A corpus of expressive descriptions in repetition," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2015, pp. 504–514.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[17] M. Tapaswi, M. Bäuml, and R. Stiefelhagen, "Aligning plot synopses to videos for story-based retrieval," *Int. J. Multimedia Inf. Retr.*, vol. 4, no. 1, pp. 3–16, Mar. 2015.

[18] H. Agrawal, A. Chandrasekaran, D. Batra, D. Parikh, and M. Bansal, "Sort story: Sorting jumbled images and captions into stories," 2016, *arXiv:1606.07493*. [Online]. Available: https://arxiv.org/abs/1606.07493

[19] S. Gaur, "Generation of a short narrative caption for an image using the suggested hashtag," in *Proc. IEEE 35th Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2019, pp. 331–337.

[20] A. Mathews, L. Xie, and X. He, "SemStyle: Learning to generate stylised image captions using unaligned text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8591–8600.

[21] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2556–2565.

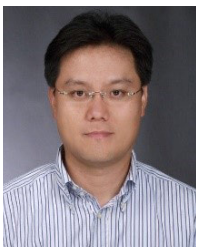
[22] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.

[23] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[24] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

KYUNGBOK MIN received the B.S. degree in electronics and computer engineering from Korea University and the Ph.D. degree in computer science and engineering from Sejong University, Seoul, South Korea, in 2021. In 2017, he joined the Computer Vision Pattern Recognition Laboratory (CVPR Lab). His current research interests include computer vision, natural language processing, and artificial intelligence.

MINH DANG received the B.S. degree in information systems from the University of Information Technology, VNU HCMC, Vietnam, in 2016, and the Ph.D. degree in computer science and engineering from Sejong University, Seoul, South Korea, in 2021. At the beginning of 2017, he joined the Computer Vision Pattern Recognition Laboratory (CVPR Lab). His current research interests include computer vision, natural language processing, and artificial intelligence.



HYEONJOON MOON received the B.S. degree in electronics and computer engineering from Korea University, in 1990, and the M.S. and Ph.D. degrees in electrical and computer engineering from the State University of New York at Buffalo, in 1992 and 1999, respectively. From January 1996 to October 1999, he was a Senior Researcher at the Electro-Optics/Infrared Image Processing Branch, U.S. Army Research Laboratory (ARL), Adelphi, MD, USA. He developed a face recognition system evaluation methodology based on the Face Recognition Technology (FERET) program. From November 1999 to February 2003, he was a Principal Research Scientist at Viisage Technology Inc., Littleton, MA, USA. He has extensive background on still image and real-time video based computer vision and pattern recognition. Since March 2004, he has been with the Department of Computer Science and Engineering, Sejong University, where he is currently a Professor and the Chairman. His main interest is on research and development is on real-time facial recognition system for access control, surveillance, and big database applications. His current research interests include image processing, biometrics, artificial intelligence, and machine learning.