ДОДАТОК А

Тези «A survey of methods of text-to-image translation»

A SURVEY OF METHODS OF TEXT-TO-IMAGE TRANSLATION

D.O. Pydorenko
Supervisor - Ph.D., Associate Professor O. Turuta
Kharkiv National University of Radio Electronics
61166, Kharkiv, Nauky ave, 14, Software Engineering Department,
e-mail: daria.pydorenko@nure.ua

The given work is devoted to the text-to-image translation. This translation can be realized in two steps. First, a large text should be transformed into a small set of captions. Second, these captions should be used to generate images. There are different methods that can help to implement both steps.

Self-publishing becomes more and more popular. And many people create cover and/or illustrations to their own texts by themselves. So, it would be very convenient to automate the process of creating images by generating them to text using certain software. The text may be too large to convert it to a specific image. First, it should be shortened, for example, by finding a description (without dialogs), generating an annotation on the whole work or finding keywords in the text. So, the task can be divided into two main steps: the compression of the text in the best way for the next processing and the generation of the image from the first step output. And both steps should get a quality assessment of conversions. The image can be converted to the caption and compared with the input of the first step.

There are several algorithms to find keywords in the text such as TF-IDF, RAKE, and TextRank. The RAKE algorithm is based on the construction of a co-occurrence graph from the words that occur most often in the text.

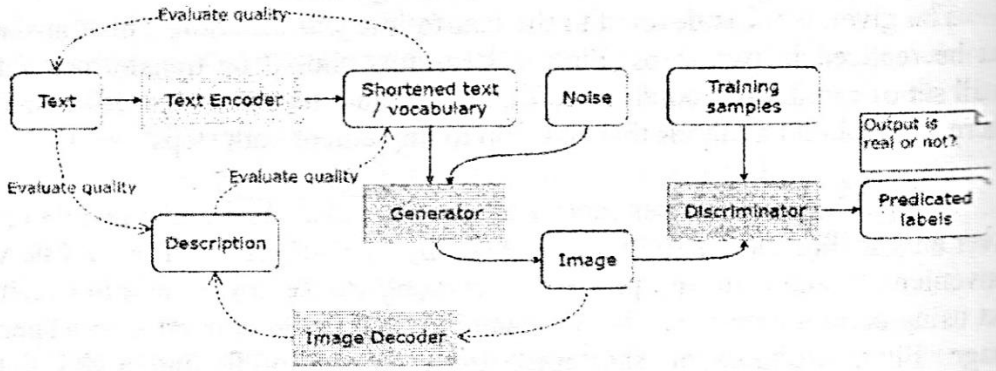
All these algorithms are represented by Python libraries. There is NLTK – a suite of libraries in Python for natural language processing (NLP) for English written text. One of the Python libraries which implements RAKE algorithm is multi-rake that helps to extract keywords from a text written in any language. They all work enough quickly but represents result in different ways. Multi-rake library should be configured to return word combinations instead of one word, because for the next step captions are needed as input. Also, the output doesn't always correspond to the most important part of the text, but the most repeatable motive. There is a problem with the language of the text on this step because texts written in English have more possibilities for their own processing.

Generation of images using neural networks is actively developing now. Different variations of Generational Competitive Networks (GAN) are used for that. GAN is represented by two networks - Generator and Discriminator, the first generates images based on the input; the second attempts to distinguish real images from generated and outputs a certain result (they are real or not).

A group of developers from the Lehigh University, Rutgers University, and Duke University and Microsoft have proposed their own solution for generating images – the Attentional Generative Adversarial Network (AttnGAN) allows multi-step change of the generated image according to the change of some

words in the text description. There is an implementation of AttnGAN on Python with PyTorch framework. But there are problems with the choice of a dataset with labels (the language of the labels, the number of images, the classes of images), and with the speed of training neural networks.

The whole process is represented in picture 1.



Picture 1 – Text to image conversion

The results below (pic. 2) were received after combining multi-rake algorithm and AttnGAN. The Last Leaf by O. Henry was chosen as a test input.

Picture 2 – Illustrations for The Last Leaf by O. Henry (1 – leaves, 2-3 – doctor)

The quality of generated images depends on datasets on which the neural network was trained. So, it is better to concentrate on a certain type of texts to improve the generation. For example, generate images of nature only and train on a dataset with such images. This may not be nature; any topic can be chosen.

The first step should be expanded with the extraction of only necessary parts of the text in this case. The text can be divided into paragraphs (or the sentences) and the number of words indicating the nature can be calculated (by the dictionary, by the classes of the dataset or its captions, etc.). Sequential paragraphs or sentences can be combined into one. All dialogs can be deleted before processing the text. If the number of words (or the percentage) exceeds a given threshold, the paragraph is sent to further processing. Thus, the program knows which input it expects and the output becomes more predictable and text-related.

ДОДАТОК Б

Стаття «A survey of methods of text-to-image translation»

A SURVEY OF METHODS OF TEXT-TO-IMAGE TRANSLATION

Pydorenko D.[1], Turuta O.[2]
[1] Master student of the Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, daria.pydorenko@nure.ua, ORCID ID: 0000-0003-0232-4634
[2] PhD in Computer Science, Associate Professor of the Department of Software Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, oleksii.turuta@nure.ua, ORCID ID: 0000-0002-0970-8617

Abstract - The given work considers the existing methods of text compression (finding keywords or creating summary) using RAKE, Lex Rank, Luhn, LSA, Text Rank algorithms; image generation; text-to-image and image-to-image translation including GANs (generative adversarial networks). Different types of GANs were described such as StyleGAN, GauGAN, Pix2Pix, CycleGAN, BigGAN, AttnGAN. This work aims to show ways to create illustrations for the text. First, key information should be obtained from the text. Second, this key information should be transformed into images. There were proposed several ways to transform keywords to images: generating images or selecting them from a dataset with further transforming like generating new images based on selected ow combining selected images e.g. with applying style from one image to another. Based on results, possibilities for further improving the quality of image generation were also planned: combining image generation with selecting images from a dataset, limiting topics of image generation.

KEYWORDS: Image generation, Text keywords, Image-to-image translation, Text-to-image translation, Text compression**.**

1. State of the Art

Online publications are currently very common. Their advantage is simplicity and accessibility for all people. A publication of their works is no longer unattainable for most people. Access to the Internet solves almost every possible problem.

One of the components of a successful, attention-grabbing publication is the illustrations that accompany it. Not everyone can create its unique illustration or buy the rights to someone else's picture. Images in the public domain often repeat and quickly become boring. How, then, can you create an illustration without art skills?

One of the options may be the generation of images by software.

There are neural networks such as GAN. GAN is a generative-competitive network, which is based on a combination of two neural networks, one of which generates candidates, and the other tries to distinguish the right candidates from the wrong ones.

There are a lot of types of GAN, and each of them does its job.

For example, the face generation is very popular now. There is even a website that creates faces of non-existing people. This creating process is based on style-based GAN. The architecture of StyleGAN leads to an automatically learned, unsupervised separation of high-level attributes (e.g., pose and identity when trained on human faces) and stochastic variation in the generated images (e.g., freckles, hair), and it enables intuitive, scale-specific control of the synthesis [1].

One of the implementations of StyleGAN architecture is provided by NVIDIA labs.

Also, NVIDIA presented GauGAN network in March 2019 and provided an interactive app that generates realistic landscape images from the layout users draw. GauGAN allows user control over both semantic and style of created images [2].

Generally, there a lot of ways to translate image to image, e.g. Pix2Pix and CycleGAN.

Pix2Pix trains to translate images basing on pairs of images {A,B}, where A and B are two different depictions of the same underlying scene. This architecture has examples of translating labels to facades, changing day to night, edges to photo, or black-white image to colorful [3].

CycleGAN is an unpaired version of Pix2Pix architecture. It can translate an image from a source domain X to a target domain Y in the absence of paired examples. This architecture is used for changing weather or season on the photo, or for applying a style of a specific artist [4].

Also, there is a neural style algorithm that can apply a style of one image to another [5].

There are a lot of GANs that can generate images basing on specific labels. They are called conditional GANs.

One of improving such GANs is BigGAN. It creates very realistic images for the specific class. It can be trained with complex datasets such as ImageNet, for example. One of the observed failures of partially-trained BigGAN models is a class leakage, where images from one class contain properties of another. Resulted images can be translated from one class to another [6].

Another improving of conditional GAN is AttnGAN. AttnGAN allows changing the generated image in a multistage manner according to the change of individual words in the text description [7].

AttnGAN uses a variational autoencoder. Autoencoders are neural networks that copy their inputs to their outputs. They work by compressing the input into a latent-space representation and then reconstructing the output from this representation. This kind of network is composed of two parts: encoder and decoder.

Variational autoencoders (unlike vanilla autoencoders) allow creating data similar to existing data or even changing it in a specific direction. Variational autoencoders are trained to extend their latent space generating possible input and feed it to the decoder to generate new data samples [8].

Based on variational autoencoders, different GANs can be built.

Also, there is a SinGAN, an unconditional generative model that can be learned from a single natural image. It can generate high quality, diverse samples that carry the same visual content as the image.

SinGAN contains a pyramid of fully convolutional GANs, each responsible for learning the patch distribution at a different scale of the image. This allows generating new realistic samples of arbitrary size and aspect ratio, that have significant variability, yet maintain both the global structure and the fine textures of the training image. In contrast to other single image GAN schemes, SinGAN approach is not limited to texture images, and is not conditional (i.e. it generates samples from noise) [9].

2. Problem Statement and Proposed Solution

There are a lot of ways to generate images basing on input images or text description. But how can this information be received from a big text?

So, the problem of generating illustrations from a text can be divided into two parts. The first one is text compression to appropriate data for input. The second one is translating text to image or selecting images from a dataset.

Consider each of the steps in more detail.

2.1 Text Compression

There are various ways to work with text: generating annotations, finding key phrases, or choosing certain words or word combinations. One of them is Rapid Automatic Keyword Extraction (RAKE) algorithm. RAKE is an algorithm to automatically extract keywords from documents [10].

Another way is natural language processing (NLP), which provides a lot of abilities to understand what text is about. For example, tokenization (the process of splitting text to words, sentences, paragraphs or other types of tokens) is one of the basic components of almost any NLP task, and it can be the first step to prepare a text for processing.

Extractive text summarization techniques perform summarization by picking portions of texts and constructing a summary, unlike abstractive techniques that conceptualize a summary and paraphrases.

There are several unsupervised graphical-based text summarizers such as Lex Rank or Text Rank.

In original TextRank the weights of an edge between two sentences is the percentage of words appearing in both of them.

Lex Rank uses IDF-modified Cosine as the similarity measure between two sentences. This similarity is used as a weight of the graph edge between two sentences. Lex Rank also incorporates an intelligent post-processing step which makes sure that the top sentences chosen for the summary are not too similar to each other.

Luhn is an algorithm that scores sentences based on the frequency of the most frequent (significant) words. It ranks sentences for summarization extracts by considering significant words and the linear distance between these words due to non-significant words.

LSA works by projecting the data into a lower-dimensional space without any significant loss of information. One way to interpret this spatial decomposition operation is that singular vectors can capture and represent word combination patterns that are recurring in the corpus. The magnitude of the singular value indicates the importance of the pattern in a document [11].

Also, there is Parts-Of-Speech tagging (POS tagging) and is also known as word classes or lexical categories. The task of POS-tagging is to labeling words of a sentence with their appropriate Parts-Of-Speech (Nouns, Pronouns, Verbs, Adjectives, etc.) [12].

For example, only nouns can be selected from the text, because nouns characterize the text most of all. Or a combination of a noun and a verb can be chosen. These word combinations show what actors are in the texts (nouns) and what actions they perform (verbs) – that is almost the main idea of the text.

2.2 Image Generation

So, images should be generated based on the keywords received in one of the ways described in paragraph 2.1.

GAN networks can be used for image generation, for example, AttnGAN.

Also based on a combination of nouns and verbs (as described in paragraph 2.1) images of people can be generated. There are many ways to generate faces, but it is better to generate the whole person who performs a specific action (so have the specific pose).

There is Pose Guided Person Image Generation network that allows synthesizing person images in arbitrary poses, based on an image of that person and a novel pose. A generation framework PG2 utilizes the pose information explicitly and consists of two key stages: pose integration and image refinement. It allows simultaneously transferring the appearance of a person from a given pose to a desired pose and keep important appearance details of the identity. Each stage is focusing on one aspect. For the first stage, several model variants can be used and, for the second stage, conditional DCGAN is used to fill in more appearance details. [13].

Another way is not to generate an image (because the result can be unpredictable), but to find specific images by labels in datasets.

The dataset should have not only images but labels, which describe each picture. Images can be selected using these labels. Image labels should match input keywords as much as possible. An example of a suitable dataset is ImageNet.

A full process of generating images for the text is shown in Fig. 1.

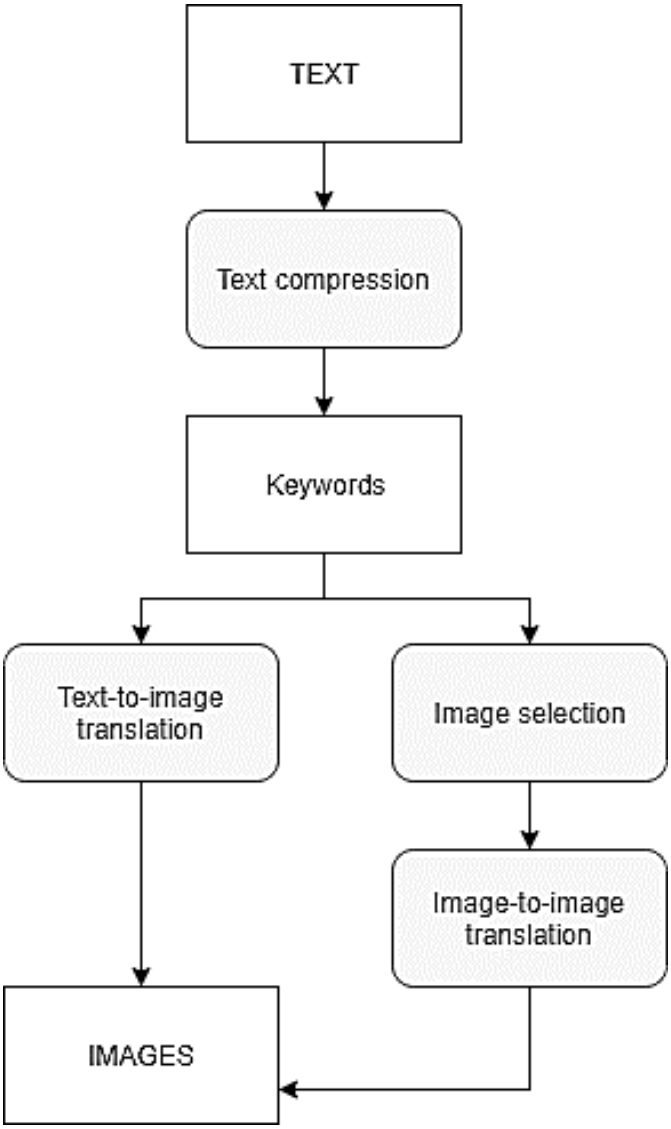


Fig. 1 – Steps of image generation for the text

3. Conclusion

First, several ways to compress text (find keywords or create summary) were considered.

Examples are created for the story “The Last Leaf” by O. Henry and the novel “The Old Man and the Sea” by Ernest Hemingway. These texts were chosen because they are short and have a lot of descriptions to generate illustrations. The obtained text compressions are shown in Table 1 and Table 2.

Table 1.

Results of Text Compression Methods (The Last Leaf)

	The Last Leaf	**Settings & Metrics**
Rake	de world mit der foolishness dot poor leetle miss yohnsy elegant horseshow riding trousers useless woollen shoulder scarf pen-and-ink drawing sue found behrman smelling strongly red eyes plainly streaming especial mastiff-in-waiting lone ivy leaf clinging wide-open eyes staring small dutch window-panes three-story brick sue	12 keywords min characters = 3, max words = 5, min frequency = 1 _____ 5 (both other texts) and 2 common key phrases
Multi Rake	michael angelo's moses beard curling de world mit der foolishness dot poor leetle miss yohnsy elegant horseshow riding trousers useless woollen shoulder scarf red eyes plainly streaming sue found behrman smelling strongly lone ivy leaf clinging brick house twenty feet busy doctor invited sue ravager strode boldly feet trod slowly	12 keywords min characters = 3, max words = 7, min frequency = 1 _____ 5 (both other texts) and 3 common key phrases
Rake NLTK	dot poor leetle miss yohnsy miss yohnsy shall lie sick brush without getting near enough de world mit der foolishness art people soon came prowling brick house twenty feet away useless woollen shoulder scarf elegant horseshow riding trousers allow dot silly pusiness sue found behrman smelling strongly last one said johnsy brick wall one ivy leaf	12 keywords _____ 5 (both other texts) and 1 common key phrases
Lex Rank	"It is the last one," said Johnsy. And then they found a lantern, still lighted, and a ladder that had been dragged from its place, and some scattered brushes, and a palette with green and yellow colours mixed on it, and - look out the window, dear, at the last ivy leaf on the wall.	2 sentences
Luhn	You may bring a me a little broth now, and some milk with a little port in it, and - no; bring me a hand-mirror first, and then pack some pillows about me, and I will sit up and watch you cook."	2 sentences

Table 1 Continuation.

	The Last Leaf	Settings & Metrics
LSA	And then they found a lantern, still lighted, and a ladder that had been dragged from its place, and some scattered brushes, and a palette with green and yellow colours mixed on it, and - look out the window, dear, at the last ivy leaf on the wall.	2 sentences
Text Rank	In a little district west of Washington Square the streets have run crazy and broken themselves into small strips called "places." "Johnsy, dear," said Sue, bending over her, "will you promise me to keep your eyes closed, and not look out the window until I am done working?	2 sentences

Table 2.

Results of Text Compression Methods (The Old Man and the Sea)

	The Old Man and the Sea	Settings & Metrics
Rake	portuguese man-of-war floating dose white tipped wide pectoral fins portuguese men-of-war long deadly purple filaments trailing purple pectoral fins set wide two-decker metal container projecting green sticks dip sharply ordinary pyramid-shaped teeth man-of-war bird thrusting all-swallowing jaws high dorsal fin knifing great erect tail slicing	12 keywords min characters = 3, max words = 5, min frequency = 1 _____ 7 common key phrases
Multi Rake	white tipped wide pectoral fins long deadly purple filaments trailing purple pectoral fins set wide portuguese man-of-war floating dose high dorsal fin knifing projecting green sticks dip sharply great erect tail slicing he'll weigh ten pounds small delicate dark terns great long white spine shark's yellow cat-like eyes small tuna's shivering pull	12 keywords min characters = 3, max words = 7, min frequency = 1 _____ 7 common key phrases
Rake NLTK	good luck old man good luck know others better que va beg keep warm old man know sleep well old man	12 keywords _____ 0 common key phrases

Table 2 Continuation.

	The Old Man and the Sea	Settings & Metrics
Rake NLTK	sorry qua va get four fresh ones one white cumulus built like friendly piles old man saw flying fish spurt old man said happily head let pedrico chop hands well old man day ends let us hope	
Lex Rank	He cannot know that it is only one man against him, nor that it is an old man. But there was not much of it.	2 sentences
Luhn	After that he began to dream of the long yellow beach and he saw the first of the lions come down onto it in the early dark and then the other lions came and he rested his chin on the wood of the bows where the ship lay anchored with the evening off-shore breeze and he waited to see if there would be more lions and he was happy. He took all his pain and what was left of his strength and his long gone pride and he put it against the fish's agony and the fish came over onto his side and swam gently on his side, his bill almost touching the planking of the skiff and started to pass the boat, long, deep, wide, silver and barred with purple and interminable in the water.	2 sentences
LSA	How would you like to see me bring one in that dressed out over a thousand pounds?" "I'll get the cast net and go for sardines. No matter what passes I must gut the dolphin so he does not spoil and eat some of him to be strong.	2 sentences
Text Rank	He saw the phosphorescence of the Gulf weed in the water as he rowed over the part of the ocean that the fishermen called the great well because there was a sudden deep of seven hundred fathoms where all sorts of fish congregated because of the swirl the current made against the steep walls of the floor of the ocean. After that he began to dream of the long yellow beach and he saw the first of the lions come down onto it in the early dark and then the other lions came and he rested his chin on the wood of the bows where the ship lay anchored with the evening off-shore breeze and he waited to see if there would be more lions and he was happy.	2 sentences

Second, AttnGAN was used to generate images from resulted keywords.

Examples were created based on results from the previous step, so the same texts were used. The obtained images are shown in Table 3 and Table 4.

Table 3.

Results of Image Generation (The Last Leaf)

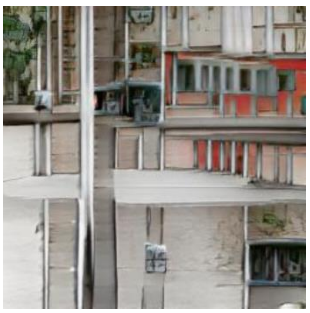
	The Last Leaf
brick wall one ivy leaf	
lone ivy leaf clinging	
In a little district west of Washington Square the streets have run crazy and broken themselves into small strips called "places."	
And then they found a lantern, still lighted, and a ladder that had been dragged from its place, and some scattered brushes, and a palette with green and yellow colours mixed on it, and - look out the window, dear, at the last ivy leaf on the wall.	

Table 4.

Results of Image Generation (The Old Man and the Sea)

	The Old Man and the Sea
small tuna's shivering pull	
purple pectoral fins set wide	
long yellow beach	
After that he began to dream of the long yellow beach and he saw the first of the lions come down onto it in the early dark and then the other lions came and he rested his chin on the wood of the bows where the ship lay anchored with the evening off-shore breeze and he waited to see if there would be more lions and he was happy.	

As part of further researches, the freedom of GAN image generation can be combined with an easy and exact selection of pictures. If the GAN receives a specific template for

creating an image, the resulting images will be more logical and realistic. It is necessary to think where this template can be obtained and how it can be used in the GAN.

GAN freedom should be limited. The ability to generate any image complicates the training of the network and its further use. Perhaps for each topic (type of image) networks can be trained separately with some specific parameters and additional restrictions or conditions.

Also, existing images can be translated using other images, as the StyleGAN mentioned in paragraph 1 does, for example.

A combination of a text-to-image translation and an image-to-image translation can be a way to provide better results of an image from text generation.

References

Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2019. – P. 4401-4410.

Park T., Liu M.-Y., Wang T.-C., Zhu J.-Y. Semantic Image Synthesis with Spatially-Adaptive Normalization // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2019. – P. 2337-2346.

Isola P., Zhu J.-Y., Zhou T., Efros A. A. Image-to-Image Translation with Conditional Adversarial Networks // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2017. – P. 1125-1134.

Zhu J.-Y., Park T., Isola P., Efros A. A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks // IEEE International Conference on Computer Vision (ICCV) – 2017. – P. 2223-2232.

Gatys L. A., Ecker A. S., Bethge M. A Neural Algorithm of Artistic Style // arXiv e-prints, arXiv:1508.06576v2 – 2015. – P. 1.

Brock A., Donahue J., Simonyan K.. Large Scale GAN Training for High Fidelity Natural Image Synthesis // arXiv e-prints, arXiv:1809.11096v2 – 2019. – P. 8.

Xu T., Zhang P., Huang Q., Zhang H., Gan Z., Huang X., He X. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2018. – P. 1316-1324.

Shafkat I. Intuitively Understanding Variational Autoencoders // Medium – 2018. – URL: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

Shaham T. R., Dekel T., Michaeli T. SinGAN: Learning a Generative Model from a Single Natural Image // IEEE International Conference on Computer Vision (ICCV) – 2019. – P. 4570-4580.

Rose, S., Engel, D., Cramer, N., & Cowley, W. Automatic Keyword Extraction from Individual Documents // M. W. Berry & J. Kogan (Eds.), Text Mining: Theory and Applications – John Wiley & Sons, Hoboken, 2010. – P. 1-20.

Pranay M., Aman G., Aayush Y. Text Summarization in Python: Extractive vs. Abstractive techniques revisited // Rare Technologies. – 2017. – URL: https://rare-technologies.com/text-summarization-in-python-extractive-vs-abstractive-techniques-revisited/

Naskar A. Extract Custom Keywords using NLTK POS tagger in python // ThinkInfi. – 2018. – URL: https://www.thinkinfi.com/2018/10/extract-custom-entity-using-nltk-pos.html

Ma L., Jia X., Sun Q., Schiele B., Tuytelaars T., Gool L. V. Pose Guided Person Image Generation // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2019. – P. 2337-234

ДОДАТОК В

Стаття «A text-to-image translation: nature illustrations»

A TEXT-TO-IMAGE TRANSLATION: NATURE ILLUSTRATIONS

Daria Pydorenko, Oleksii Turuta
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, daria.pydorenko@nure.ua
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, oleksii.turuta@nure.ua

The given work considers ways of text-to-image translation using RAKE (algorithm for finding keywords in a text) and GANs (generative adversarial networks). An example of a topic for image translation is nature. Images are selected from a dataset by keywords, transformed using their segmentation, GauGAN and StyleGAN. Results are compared with image generation using AttnGAN. Based on results, possibilities for further researches are suggested: merging images with different types, for example, nature images with generated people images.

NATURE, LANDSCAPE, IMAGE GENERATION, TEXT KEYWORDS, IMAGE-TO-IMAGE TRANSLATION, TEXT-TO-IMAGE TRANSLATION, TEXT COMPRESSION

1. State of the Art

In the age of modern technology, there are a lot of opportunities to create images with the help of software. For example, people without an aptitude for drawing can illustrate something by themselves not to buy images or use stock ones.

There are neural networks such as GAN. GAN is a generative-competitive network, which is based on a combination of two neural networks, one of which generates candidates, and the other tries to distinguish the right candidates from the wrong ones.

There are a lot of types of GAN, and each of them does its job. But usually generated images are not so good as people can draw. The larger set of images network can generate – the less intelligible is the picture.

So how to make networks to generate illustrations for a really extensive but still specific topic like nature?

NVIDIA presented GauGAN network in March 2019 and provided an interactive app that generates realistic landscape images from the layout users draw [1]. So, this network allows creating a specific type of illustrations (a nature) if the user can provide layout (or segmentation) for it.

There are a lot of algorithms for image semantic segmentation, most of which are improvements of Fully Convolutional Network (FCN).

One of them is Mask R-CNN. The Mask R-CNN is a Faster R-CNN with 3 output branches: the first one computes the bounding box coordinates, the second one computes the associated class and the last one computes the binary mask[3] to segment the object. The binary mask has a fixed size and it is generated by an FCN for a given RoI.

The Faster R-CNN architecture for object detection uses a Region Proposal Network (RPN) to propose bounding box candidates. The RPN extracts Region of Interest (RoI) and a RoIPool layer computes features from these proposals to infer the bounding box coordinates and the class of the object. Mask R-CNN also uses a RoIAlign layer instead of a RoIPool to avoid misalignments due to the quantization of the RoI coordinates.

The particularity of the Mask R-CNN model is its multi-task loss combining the losses of the bounding box coordinates, the predicted class and the segmentation mask [2].

These segmentations can be used as a layout in a GauGAN network.

Also, there is a neural style algorithm that can apply a style of one image to another [3]. Usually, such networks are used to apply the style of some artists. But it doesn't matter which image is used as an input – a piece of art, photography or something else.

There are a lot of GANs that can generate images basing on specific labels. They are called conditional GANs.

One of improving of conditional GANs is AttnGAN. AttnGAN allows changing the generated image in a multistage manner according to the change of individual words in the text description [4].

AttnGAN uses a variational autoencoder. Autoencoders are neural networks that copy their inputs to outputs. They work by compressing the input into a latent-space representation and then reconstructing the output from this representation. This kind of network is composed of two parts: encoder and decoder. Variational autoencoders (unlike vanilla autoencoders) allow creating data similar to existing data or even changing it in a specific direction. Variational autoencoders are trained to extend their latent space generating possible input and feed it to the decoder to generate new data samples [5].

Also, there is a SinGAN, an unconditional generative model that can be learned from a single natural image. It can generate high quality, diverse samples that carry the same visual content as the image.

SinGAN contains a pyramid of fully convolutional GANs, each responsible for learning the patch distribution at a different scale of the image. This allows generating new realistic samples of arbitrary size and aspect ratio, that have significant variability, yet maintain both the global structure and the fine textures of the training image. In contrast to other single image GAN schemes, SinGAN approach is not limited to texture images, and is not conditional (i.e. it generates samples from noise) [6].

2. Problem Statement and Proposed Solution

Text-to-image translation can be divided into two steps: text compression and image generation. The result of the first step must be controlled to be suitable for the second step.

If the resulting keywords are too abstract, any GAN won't be able to generate images, regardless of which dataset it was trained on.

In this case, it can be done oppositely. Only keywords that match the dataset and the trained GAN model accordingly should be selected.

2.1 Text Compression

A text compression can be obtained using the Rapid Automatic Keyword Extraction (RAKE) algorithm. RAKE is an algorithm to automatically extract keywords from documents [7].

But how can be controlled what exactly is received?

For example, there are a dataset of images with labels. So, the list of all possible labels can be received. These labels can be called a dictionary.

Then the text can be checked for the presence of words from the dictionary and only those sentences or paragraphs (text units) which contain them should be selected. And only after this step the text received should be compressed. The resulting key phrases can also be analyzed for the presence of words from the dictionary.

The number of matching words or the percentage of such words in a text unit can be calculated and only those units where the number exceeds a certain threshold value should be selected.

Paragraphs can be analyzed separately, or adjacent paragraphs can be combined into one.

Using WordNet, not only the presence of words from the dictionary can be checked, but also their synonyms (choose a certain percentage of similarity). WordNet is a large lexical database of English. Nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept [8].

Depending on the topic, dialogs may or may not be counted. For example, descriptions of nature can be found without dialogues at all.

The "extra" punctuation marks can also be removed.

2.2 Image Generation

Based on the keywords obtained in paragraph 2.1, images from the dataset should be selected.

The dataset should have not only images but labels, which describe each picture. Images can be selected using these labels. Image labels should match input keywords as much as possible.

An example of a suitable dataset is ImageNet.

Further, the received images can be converted in a certain way. There are several ways to do it.

First, different autoencoders can be used. Autoencoders copies their input to output, and variational autoencoders can generate new output based on input.

Second, GauGAN. The first step will be receiving image segmentation, and then generating a new image based on segmentation.

Third, StyleGAN. In this case, a new image will be created based on two selected images: will be taken the segmentation first and applied the second style. Or even first GauGAN can be applied to one of the images, and then the styles of the other images can be applied to the resulting image.

Fourth, SinGAN. Preliminarily models can be trained for each image in the dataset (but it will be very time consuming). After, new images can be generated based on models of the selected images. Also, SinGAN allows transforming paints to images. So, an image segmentation can be used like in the case when GauGAN is used.

Thus, a large number of new images can be obtained, almost unique and quite diverse.

3. Conclusion

First, RAKE algorithm was used to compress text (find keywords). The obtained text compressions are shown in Table 1.

Table 1.

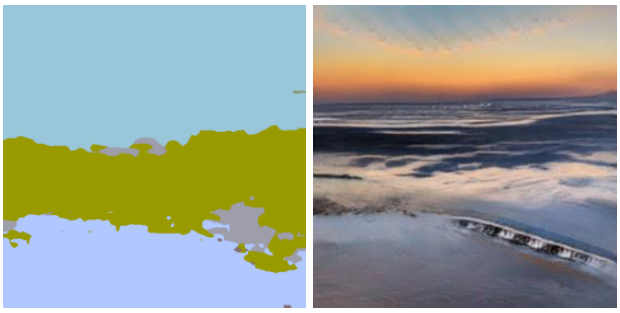
Results of Text Compression Method (The Old Man and the Sea)

	The Old Man and the Sea	**Settings**
Rake	[('man heard dip push oars', 25.0), ('boats beaches', 4.0), ('moon hills', 4.0), ('sea', 1.0)] [('clouds building trade wind looked ahead', 36.0), ('sea', 1.0)] [('signs sky days ahead sea', 25.0), ('hurricane', 1.0)] [('sleep moon sun sleep', 16.0), ('days current flat calm', 16.0), ('ocean sleeps', 4.0)] [('late afternoon', 4.0), ('sea sky', 4.0)]	12 keywords min characters = 3, max words = 7, min frequency = 1 By paragraphs, paragraphs are connected

Second, several images were found. For example, the keywords 'late afternoon' and 'sea sky' can be chosen to find images. A dataset which images below were selected from is called Earth Porn and was taken from Reddit [9].

Then GauGAN was used to translate images from images. The obtained images are shown in Figures 1-4.

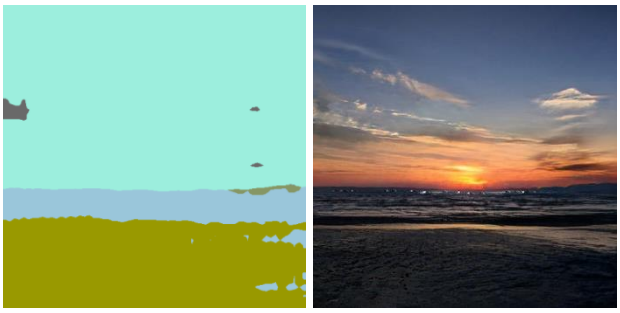
a) Input (found image)



b) Segmentation c) Output (generated image)

Figure 1. Image-to-image translation

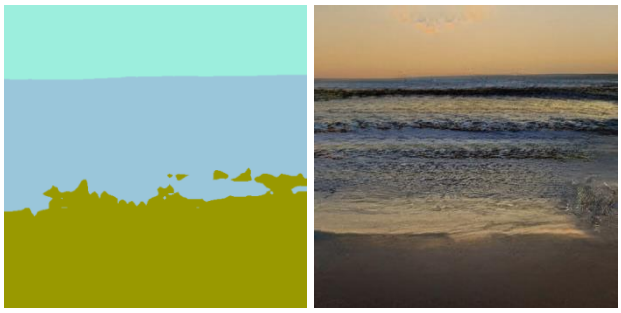
a) Input (found image)



b) Segmentation c) Output (generated image)

Figure 2. Image-to-image translation

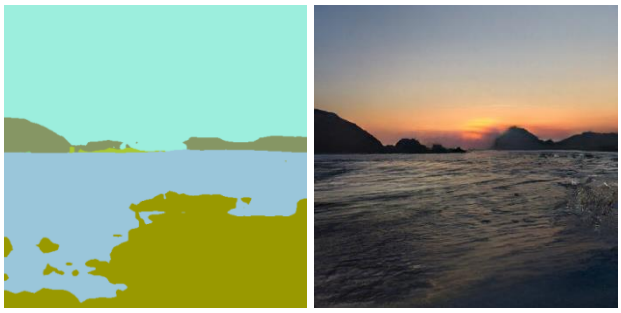
a) Input (found image)



b) Segmentation c) Output (generated image)

Figure 3. Image-to-image translation

a) Input (found image)



b) Segmentation c) Output (generated image)

Figure 4. Image-to-image translation

Also, template segmentation can be used. The result is shown in Figure 5 (different input images can be applied for a neural style transfer).

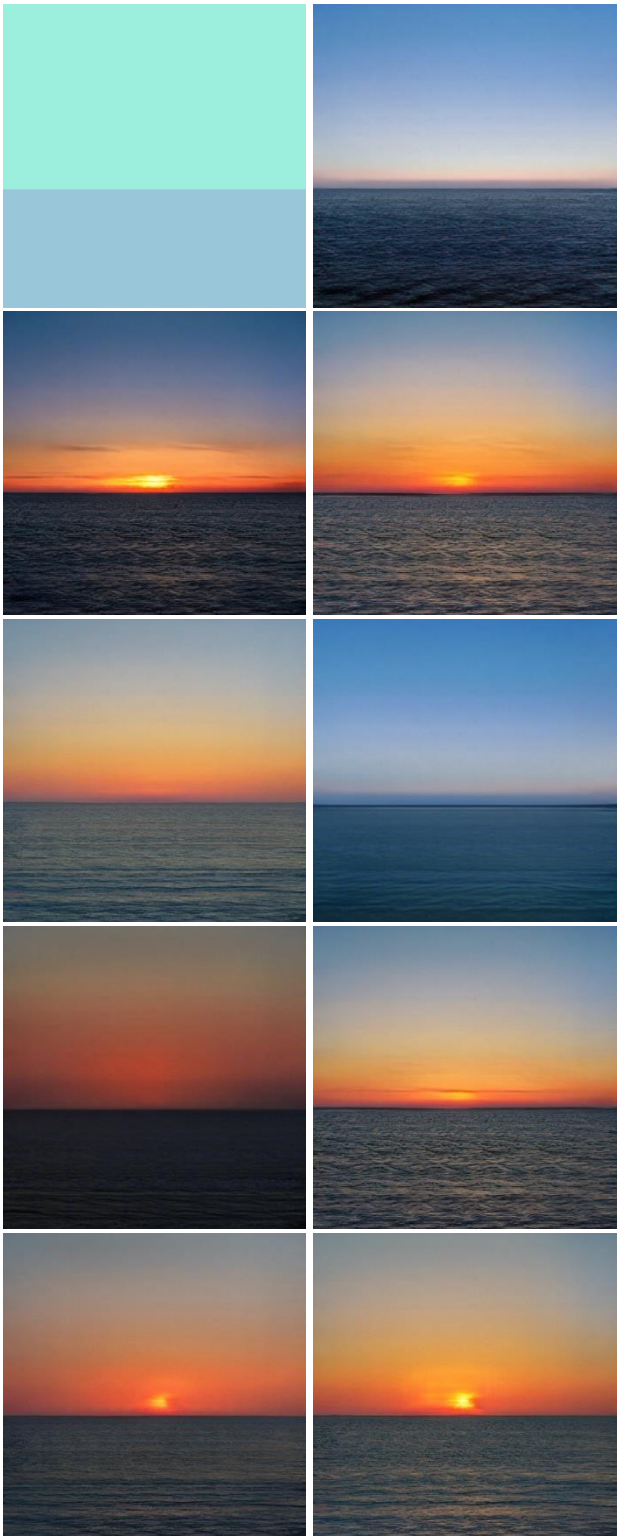


Figure 5. Image-to-image translation

For each selected image models were trained using SinGAN. Then new images were generated. Received results are shown in Table 2.

Table 2.

Results of SinGAN Image Generation (The Old Man and the Sea)

The Old Man and the Sea		
		Scale = 5
		Scale = 3
		Scale = 3
		Scale = 4
		Scale = 4
		Scale = 3
		Scale = 5

A result can be compared with images generated by AttnGAN, look at Table 3.

Table 3.

Results of AttnGAN Image Generation (The Old Man and the Sea)

	The Old Man and the Sea
It was getting late in the afternoon and he saw nothing but the sea and the sky.	
late afternoon	
sea sky	

So, a combination of a text-to-image translation and an image-to-image translation with GauGAN provides better results of an image from text generation than just text-to-image translation with AttnGAN.

As part of further researches, images generated for different topics can be merged. For example, people's images can be generated separately and background (like landscape or interior) separately.

There is a Pose Guided Person Image Generation network that allows synthesizing person images in arbitrary poses, based on an image of that person and a novel pose [10]. Combinations of a noun and a verb can be obtained from a text. Images can be selected from the special dataset by these word combinations showing what actors are in the texts (first group of images with people's looks) and what actions they perform (second group of images with people's poses). And based on that, people's images can be generated.

And then an image of a person (or people if several images are combined) can be merged with a landscape image to receive one resulting image.

References

Park T., Liu M.-Y., Wang T.-C., Zhu J.-Y. Semantic Image Synthesis with Spatially-Adaptive Normalization // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2019. – P. 2337-2346.

Ouaknine A. Review of Deep Learning Algorithms for Image Semantic Segmentation // Medium – 2018. – URL: https://medium.com/@arthur_ouaknine/review-of-deep-learning-algorithms-for-image-semantic-segmentation-509a600f7b57

Gatys L. A., Ecker A. S., Bethge M. A Neural Algorithm of Artistic Style // arXiv e-prints, arXiv:1508.06576v2 – 2015. – P. 1.

Xu T., Zhang P., Huang Q., Zhang H., Gan Z., Huang X., He X. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2018. – P. 1316-1324.

Shafkat I. Intuitively Understanding Variational Autoencoders // Medium – 2018. – URL: https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

Shaham T. R., Dekel T., Michaeli T. SinGAN: Learning a Generative Model from a Single Natural Image // IEEE International Conference on Computer Vision (ICCV) – 2019. – P. 4570-4580.

Rose, S., Engel, D., Cramer, N., & Cowley, W. Automatic Keyword Extraction from Individual Documents // M. W. Berry & J. Kogan (Eds.), Text Mining: Theory and Applications – John Wiley & Sons, Hoboken, 2010. – P. 1-20.

WordNet // Princeton University – URL: https://wordnet.princeton.edu/

EarthPorn // Reddit – URL: https://www.reddit.com/r/EarthPorn/

Ma L., Jia X., Sun Q., Schiele B., Tuytelaars T., Gool L. V. Pose Guided Person Image Generation // IEEE Conference on Computer Vision and Pattern Recognition (CVPR) – 2019. – P. 2337-2346.

ДОДАТОК Г

Слайди презентації

Дослідження методів генерування ілюстрацій на основі тексту

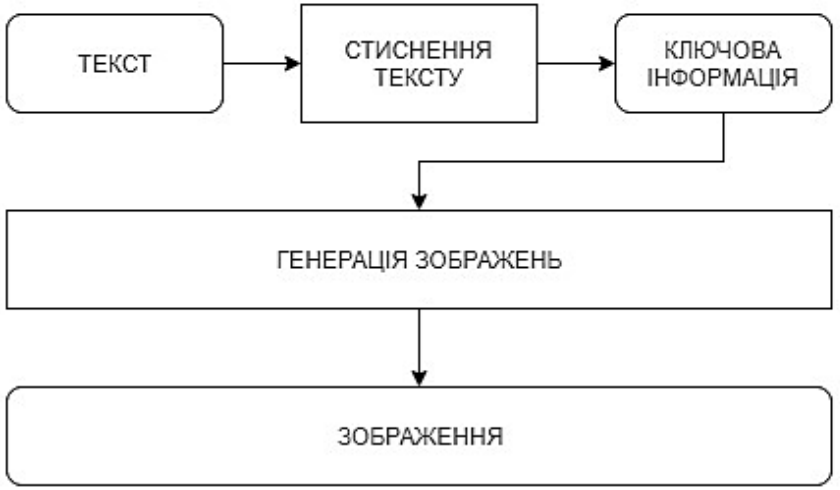
Виконала:
ст. гр. ПЗСм-18-1
Пидоренко Д.О.

Керівник:
доцент каф. ПІ, к.т.н., доцент Турута О.П.

Мета роботи

Створити ілюстрації до текстів:

- дослідити методи стиснення тексту;
- дослідити методи генерації зображень;
- оцінити отриманий результат.



Аналіз предметної галузі – RAKE

Tulips and dandelions are flowers.
Tulips are red and dandelions are yellow.
I like this red tulip.

---

red tulip: 3.5
red: 1.5
tulips: 1.0
dandelions: 1.0
flowers: 1.0
yellow: 1.0

https://pypi.org/project/multi-rake/

Frequency:
{'tulips': 2, 'dandelions': 2, 'flowers': 1,
'red': 2, 'yellow': 1, 'tulip': 1})

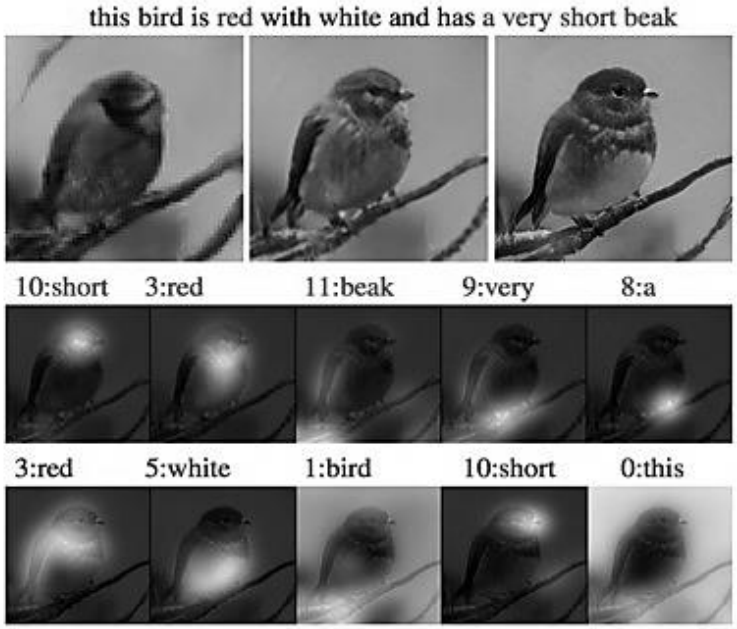
Degree:
{'tulips': 0, 'dandelions': 0, 'flowers': 0,
'red': 1, 'yellow': 0 , 'tulip': 1})

Sum = frequency + degree:
{'tulips': 2, 'dandelions': 2, 'flowers': 1, 'red': 3,
'yellow': 1, 'tulip': 2}

Frequency / sum:
{'tulips': 1.0, 'dandelions': 1.0, 'flowers': 1.0, 'red':
1.5, 'yellow': 1.0, 'tulip': 2.0})

Аналіз предметної галузі

AttnGAN



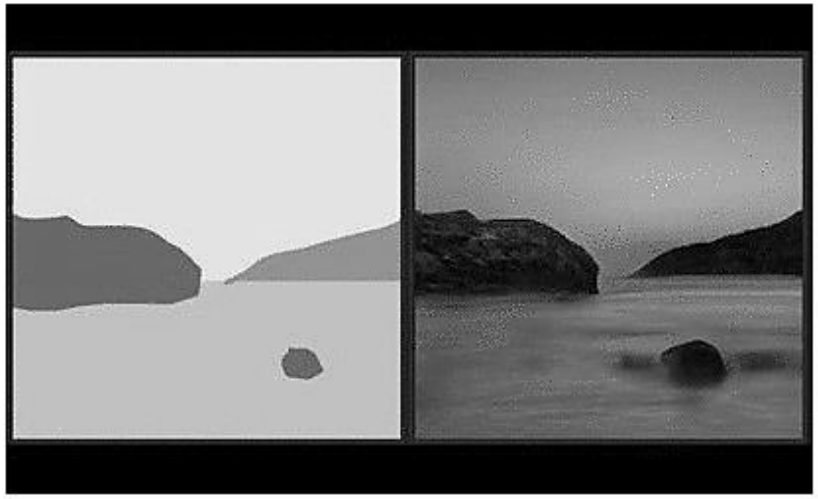
https://github.com/taoxugit/AttnGAN

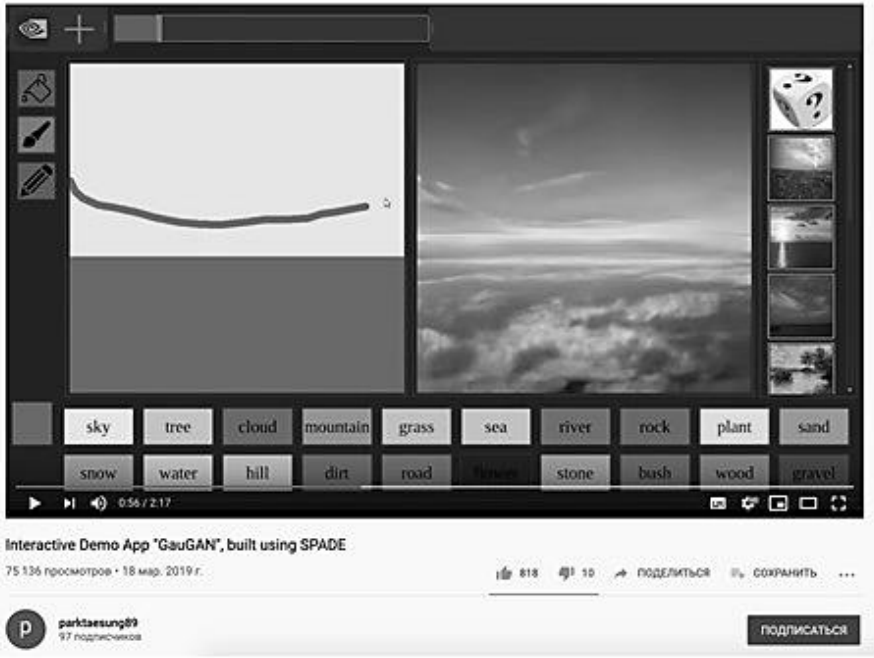
Neural Style

https://github.com/jcjohnson/neural-style

Аналіз предметної галузі

GauGAN

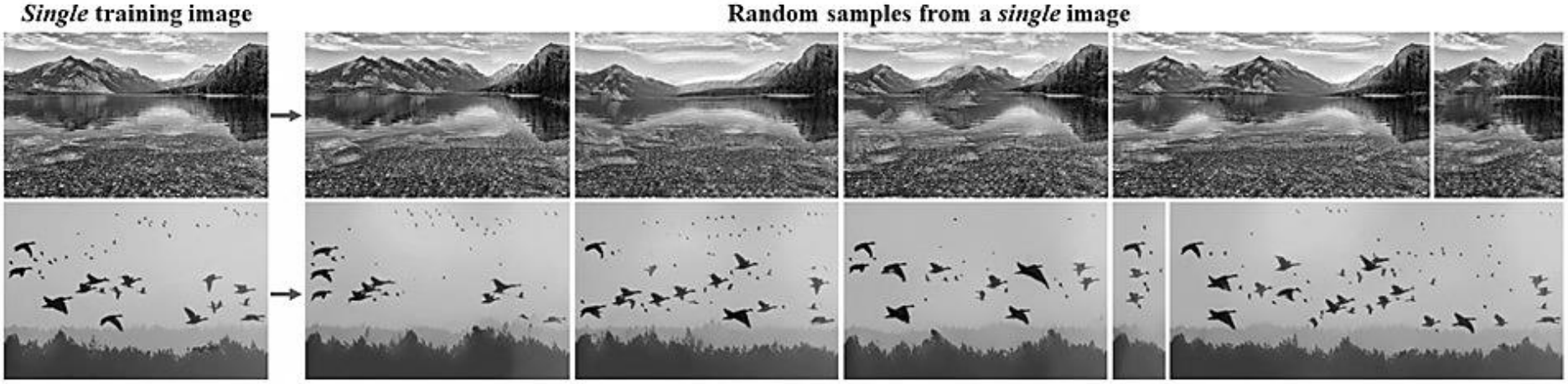




https://github.com/NVlabs/SPADE

Аналіз предметної галузі

SinGAN



https://github.com/tamarott/SinGAN

Постановка задачі

- Дослідити методи стиснення тексту: RAKE, TextRank, LexRank, Luhn, LSA.
  Можливі проблеми: налаштування методів скорочення тексту для української мови, складність обрати саме такі ключові слова, які можуть бути проілюстровані.

- Дослідити GAN-моделі для генерації зображень: створення зображення з тексту або із зображення.
  Можливі проблеми: відсутність датасету зображень з мітками/описом для української мови.

- Оцінити результати за наступними факторами: суб’єктивна метрика, унікальність, відповідність отриманого зображення тексту.

Реалізація

python

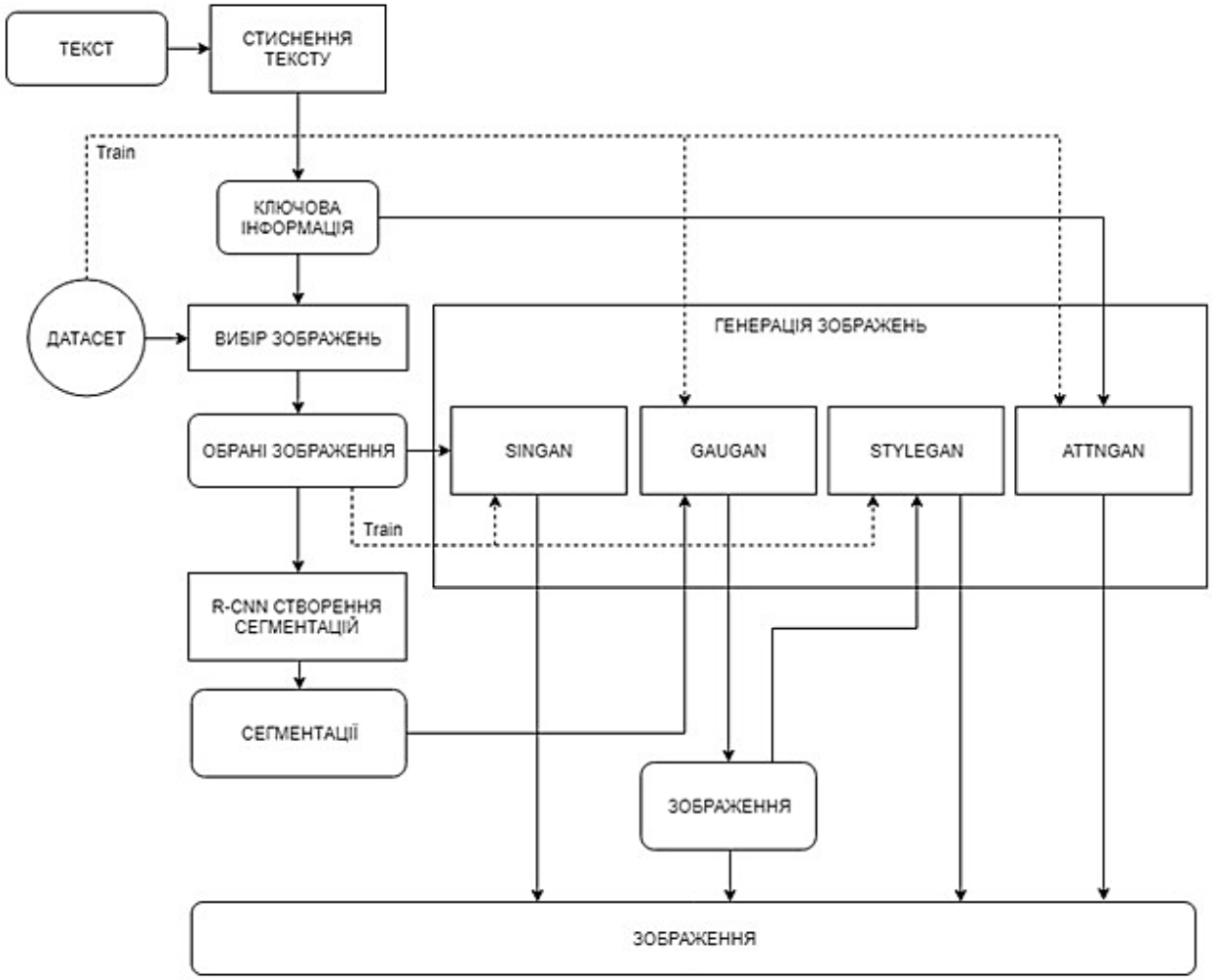
PyTorch

WORDNET

NLTK

IMAGENET

COCO
Common Objects in Context

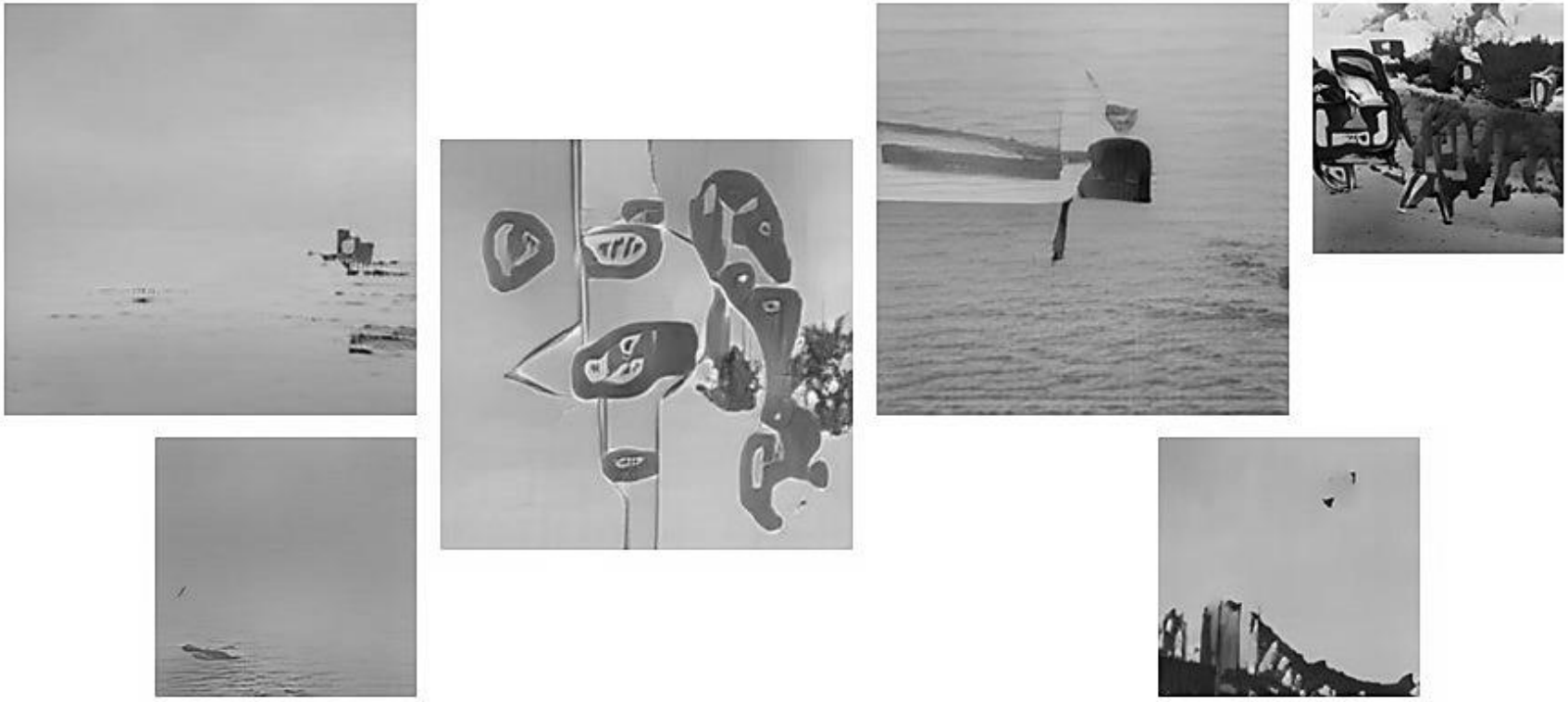


Стиснення тексту – RAKE (multi-rake)

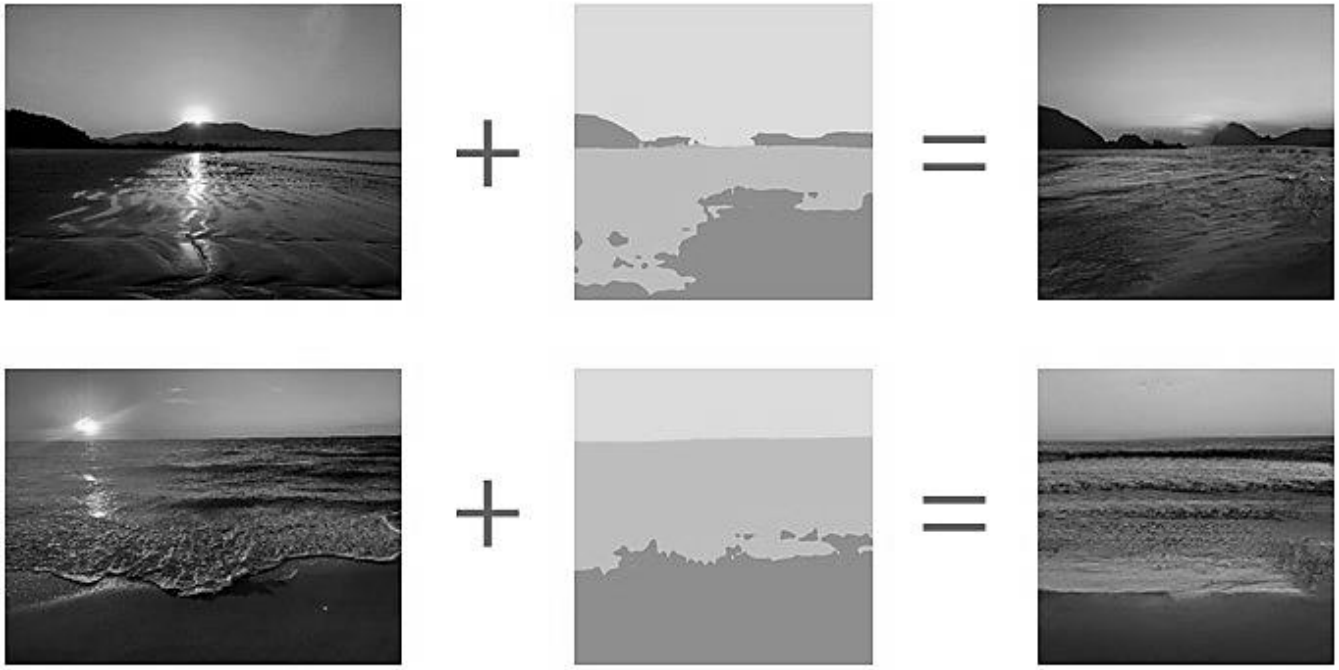
Старий і море (Е. Хемінгуей)

[('море великого ді маджо— мовив старий— кажуть', 49.0), ('— хотів би', 9.0), ('взяти', 1.0), ('батько', 1.0), ('рибалкою', 1.0)]
[('бачив бо місяць уже сховався', 25.0), ('розтинають воду їхні весла', 16.0), ('різних боків', 4.0), [('залишивши позаду запахи берега кермував', 25.0), ('звечора надумав вийти ген', 16.0), ('свіжий вранішній дух океану', 16.0), ('відкрите море', 4.0), ('старий', 1.0)]
[('подумки називав море la mar', 25.0), ('кажуть по-іспанському', 4.0), ('любить', 1.0)]
[('місяць розтривожує море', 9.0), ('— думав', 4.0), ('само', 1.0), ('жінку', 1.0), ('старий', 1.0)]
[('макрель ішла стрімко випереджаючи', 16.0), ('зрештою мала перейняти рибу', 16.0), ('знов пірне', 4.0), ('водою', 1.0), ('літ', 1.0), ('море', 1.0)]
[('високому вересневому небі видніли тонкі смужки пірчастих хмаринок', 64.0), ('небо й побачив білі купчасті хмари подібні', 49.0), ('одну апетитних кульок морозива а', 25.0), ('накладених одна', 4.0), ('подивився', 1.0)]
[('заходив вечір а старий', 16.0), ('бачив', 1.0), ('море', 1.0), ('небо', 1.0)]
[('безкрає море —', 9.0), ('друзі й вороги', 9.0), ('ньому', 1.0)]
…

Генерація зображень – AttnGAN

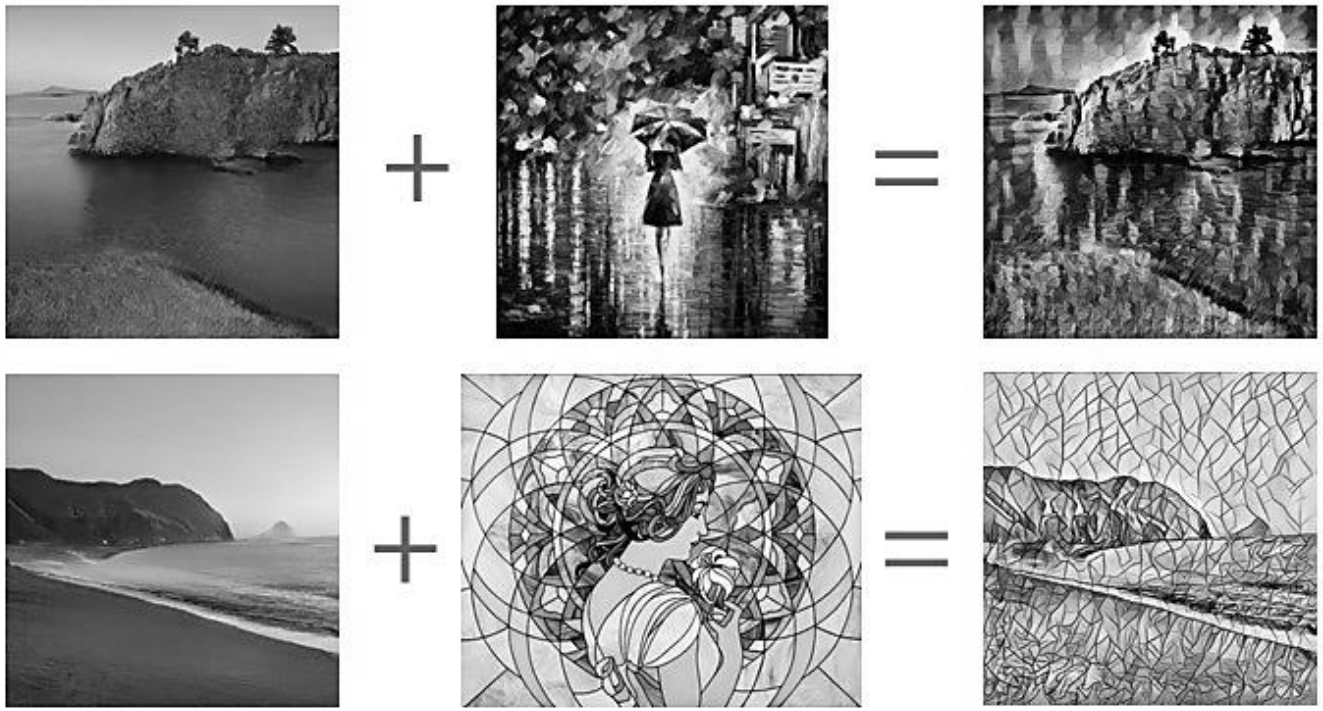


Генерація зображень - GauGAN



https://reddit.com/r/EarthPorn/

Генерація зображень - StyleGAN



https://reddit.com/r/EarthPorn/

Генерація зображень – GauGAN + StyleGAN

https://reddit.com/r/EarthPorn/

Генерація зображень - SinGAN

https://reddit.com/r/EarthPorn/

Генерація зображень – Порівняння

Критерії	AttnGAN	GauGAN	StyleGAN	SinGAN
Вхідні дані	Слово або словосполучення, або речення	Зображення або його сегментація	Зображення	Натренована модель зображення
Тренування	Одна модель	Одна модель	Модель для кожного стилю	Модель для кожного зображення
Достатній час тренування	Декілька днів	Декілька днів	Декілька годин	Декілька годин
Датасет для тренування	Зображення з описом	Зображення з мітками	Одне зображення	Одне зображення
Результат	Зображення	Зображення	Зображення	Зображення
Стиль результату	Не має чіткої відповідності	На основі одного із входів	На основі одного із входів	На основі входу
Сегментація результату	Не має чіткої відповідності	На основі входу	На основі іншого із входів	На основі входу

Оцінка отриманого результату

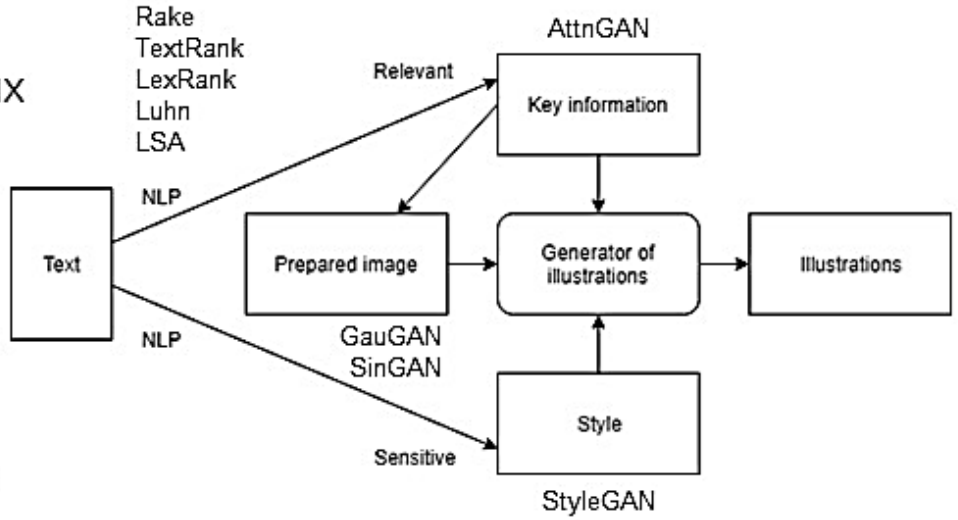
Критерії	AttnGAN	GauGAN	StyleGAN	SinGAN
Суб'єктивна оцінка	2.25	4	3.6	3.5
Унікальність (MSE та SSIM)	Висока унікальність (нема базового зображення)	MSE: 3348.113 SSIM: 0.44	MSE: 2604.122 SSIM: 0.422	MSE: 796.45 SSIM: 0.55
Відповідність тексту	Достатній рівень	Високий рівень	Поганий рівень	Достатній рівень

MSE – середній квадрат помилки $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

SSIM – індекс структурної подібності $\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$

Висновки

- Були досліджені задачі стиснення тексту та генерації зображень.
- Був створений датасет розмічених зображень на тему природи.
- Були згенеровані зображення природи за допомогою різних моделей GAN та їх комбінацій.
- Далі потрібно дослідити інші моделі (наприклад, PoseGAN для генерації людей) та збагатити датасет для розширення можливих тем генерації.



ДОДАТОК Д

Відгук та рецензії

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ РАДІОЕЛЕКТРОНІКИ
Факультет комп'ютерних наук

ВІДГУК

на атестаційну роботу магістра
Пидоренко Дар'ї Олександрівни, ПЗСм-18-1,
спеціальність 121 - Інженерія програмного забезпечення
освітньо-професійна програма «Програмне забезпечення систем»
Тема атестаційної роботи «Дослідження методів генерування ілюстрацій на основі тексту»

В атестаційній роботі розглянуті актуальні задачі аналізу природної мови та генерації зображень, комбінація таких методів дозволила розробити інноваційне рішення для генерації ілюстрацій, додавання стилю на основі текстів різної довжини.

Студент Пидоренко Д. проявила себе як самостійній дослідник, провела дослідження методів стиснення тексту, генерації зображень. На основі проведених досліджень, в перший рік виконання роботи, були отримані ілюстрації низької якості та умовної відповідності до вхідного тексту. Системна робота Дар'ї Пидоренко привела до отримання належного результату.

В результаті виконання атестаційної роботи автором були отримані наукові результати, які представлені на Молодіжному форумі, підготовлені тези конференції та публікація в фаховому виданні для технічних наук.

Магістрант гр. ПЗСм-18-1 Пидоренко Д.О. готовий до самостійної інженерної діяльності. Атестаційну роботу можна подати до захисту в ЕК за спеціальністю 121-«Інженерія програмного забезпечення», освітньо-професійною програмою «Програмне забезпечення систем».

«_____» _____ 2019 р.
підпис Керівник атестаційної роботи магістра
Турута О.В., доцент каф. ПІ., к.т.н., доцент

Рецензія

на атестаційну роботу магістра
студента групи ПЗСм-18-1 Пидоренко Д.О.
спеціальність – 121 - Інженерія програмного забезпечення
освітньо-професійна програма «Програмне забезпечення систем»
«Дослідження методів генерування ілюстрацій на основі тексту».

Атестаційна робота складається з пояснювальної записки __ сторінок; графічної частини __ аркушів; програмне застосування у вигляді набора файлів та моделей ___ файлів загальним обсягом ____ Кбайт.

В роботі проводится мультидисциплінарне дослідження, яке поєднує методи обробки природної мови та генерації зображень. Актуальним є методи генерації зображень на основі ключових слів та попередньо опрацьованого шаблону.

Атестаційна робота складається з розділів аналізу предметної області, постановки задачі, стиснення тексту, генерація зображень, оцінка отриманого результату, висновків та додатків. В розділах стиснення тексту проведений аналіз існуючих методів виділення ключової інформації на основі частотної характеристики слів та на основі граматичних основ. Дослідження проводяться для корпусів текстів, що представлені англійської та українською мовами. В розділі генерації зображень порівнюються методи створення зображень на основі претренерованих моделей нейронних мереж на текстовому наборі даних та на основі попередньо сегментованих зображень. Додатково розглядається метод передачі стилю від базових зображень.

Розроблене програмне застосування є автоматичним рішенням створення ілюстрацій для текстів різної довжини. Проведені дослідження підтверджують можливість створення ілюстрацій для текстів будь-якої довжини та тематики, але найкращі результати отримані для створення пейзажів.

Виконані дослідження та проведена робота та отримані результати повністю описані в пояснювальній записці. По результатам дослідження опубліковані тези доповіді та підготовлена публікація.

В якості недоліків слід вказати різну якість отриманих ілюстрацій та їх відповідність до вхідного тексту.

Атестаційна робота магістранта групи ПЗСм-18-1 Пидоренко Д.О. відповідає вимогам до атестаційних робіт і заслуговує оцінки «відмінно – 96 А». Атестаційну роботу можна представити для захисту в ЕК за спеціальністю 121 - Інженерія програмного забезпечення, освітньо-професійною програмою «Програмне забезпечення систем».

Рецензент

Рецензія

на атестаційну роботу магістра
студента групи ПЗСм-18-1 Пидоренко Д.О.
спеціальність – 121 - Інженерія програмного забезпечення
освітньо-професійна програма «Програмне забезпечення систем»
«Дослідження методів генерування ілюстрацій на основі тексту».

Атестаційна робота складається з пояснювальної записки __ сторінок; графічної частини __ аркушів; програмне застосування у вигляді набора файлів та моделей ___ файлів загальним обсягом ____ Кбайт.

Робота присвячена аналізу природної мови та генерації зображень на основі проаналізованого тексту. Актуальність теми обумовлена використанням сучасних підходів визначення ключової інформації, інтонації тексту та алгоритмів нейронних мереж для переносу стилю зображення та генерації зображень.

Атестаційна робота складається з наступних розділів: аналіз предметної області, постановка задачі, стиснення тексту, генерація зображень, оцінка отриманого результату, висновків та додатків. В розділі аналізу предметної області розглядаються існуючі алгоритми стиснення тексту та генерації зображень. В постановці задачі досліджень сформульована актуальна задача створення ілюстрацій для тексту. В наступних розділах розглядаються алгоритми стиснення тексту, визначення ключових слів, граматичних основ, семантичного відтінку тексту; алгоритми генерації зображень, трансферу стилю зображень, аналіз підходів для генерації зображень; розробка методу оцінки якості згенерованих зображень та відповідності первісного тексту. В додатках приводяться дві публікації студента.

В роботі проводиться порівняння методів стиснення тексту, результати порівняння наведені в таблиці 4.2. Проведено порівняння методів генерації зображень та визначенні ефективні підходи для створення відповідних зображень. Розроблений додаток автоматично створює ілюстрації на основі текстів українською та англійською мовами. Дослідження проведені в необхідному обсязі, дана оцінка якості та відповідності зображень до вихідного тексту. На основі проведених досліджень реалізоване працездатне програмне забезпечення та впроваджено у вигляді сервісу. Пояснювальна записка оформлена відповідно до вимог та стандартів.

В якості недоліків слід зазначити недостатній обсяг набору зображень описаних української мовою, що обмежує можливості генерації зображень на основі текстів українською мовою; відсутність інтеграції в існуючі текстові редактори (наприклад, MS Word).

Атестаційна робота магістранта групи ПЗСм-18-1 Пидоренко Д.О. відповідає вимогам до атестаційних робіт і заслуговує оцінки «відмінно – 96 А». Атестаційну роботу можна представити для захисту в ЕК за спеціальністю 121 - Інженерія програмного забезпечення, освітньо-професійною програмою «Програмне забезпечення систем».

Рецензент