UDC 004.89 DOi 10.30837/ bi.2019.2(93).13

Konarieva I.[1] , Pydorenko D.[2], Turuta O.[3]

[1]Master student of the University of Computense, Madrid (UCM), Spain,
iulikona@ucm.es, ORCID ID: 0000-0001-9266-9877

[2] 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

[3] 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

A SURVEY OF METHODS OF TEXT-TO-IMAGE TRANSLATION

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.

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 multi-stage 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 [13-15].

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.

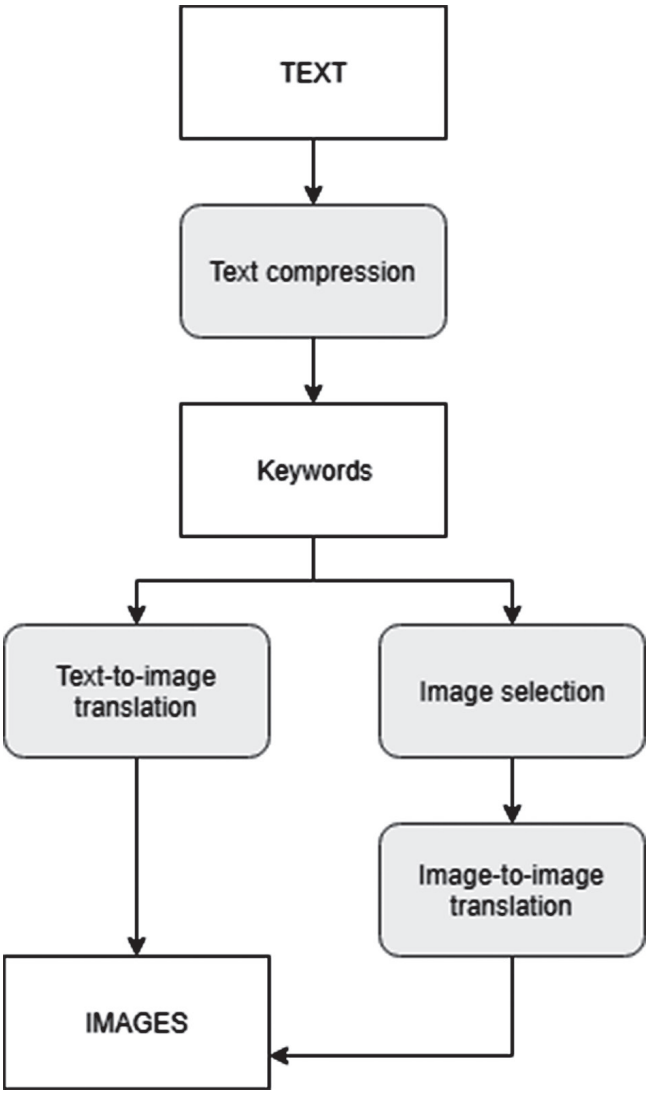


Fig. 1. Steps of image generation for the text

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
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)

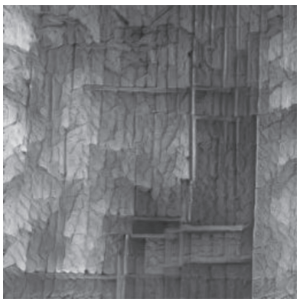
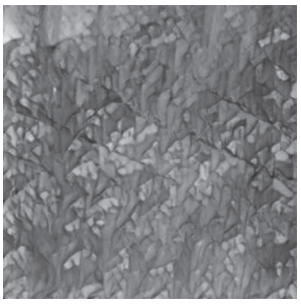
	The Old Man and the Sea	**Settings & Metrics**
Rake	portuguese man-of-war floating dose white tipped wide pectoral fins portuguese men-of-war long deadly purple filaments trailing purple pectoral fins set wide two-decker metal container projecting green sticks dip sharply ordinary pyramid-shaped teeth 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
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.»	

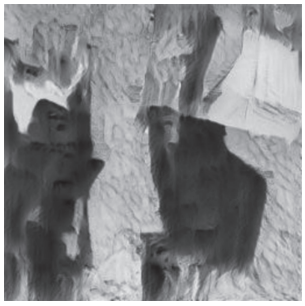
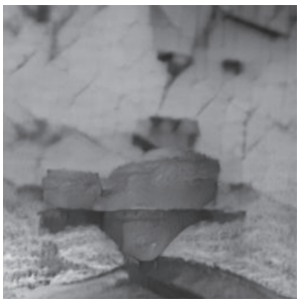
And then they found a lantern, still lighted, and a ladder that had been dragged from its place, and some scattered brushes, and a palette with green and yellow colours mixed on it, and - look out the window, dear, at the last ivy leaf on the wall.	

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

[1] *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.

[2] *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.

[3] *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.

[4] *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.

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

[6] *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.

[7] *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.

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

[9] *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.

[10] *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.

[11] *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/

[12] *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

[13] *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.

[14] *Bondarenko M., Konoplyanko Z., Chetverikov G.* Analiz problemy sozdaniya novich technicheskich sredstv dlya realizataii lingvisticheskogo interfeisan // Proc.of the 10th International Conference KDS–2003, Varna, Bulgaria, – 2003. – P. 3–15.

[15] *Bondarenko M.F., Konoplyanko Z..D., Chetverikov G.G.* Theory fundamentals of multipli-valued structures and coding in artificial intelligence systems.iKharkiv: Factor-druk. 2003.– 336 p.

The article was delivered to your editory stuff on the 20.11.2019