Towards a Visual Language Using Neural Networks

Luís Gonçalo, João M. Cunha, and Penousal Machado
University of Coimbra, CISUC, Department of Informatics Engineering
{lgoncalo,jmacunha,machado}@dei.uc.pt

Abstract

In recent years computer simulations have proven to be useful in the study of the origin and evolution of communication. In this paper, we present a system that is able to evolve image-based communication protocols to transmit information. We trained an encoder and a decoder in an architecture similar to a model of communication where the generator transforms a message into an image and the decoder tries to reconstruct the original message. This way the two networks are stimulated to work together to establish some type of communication.

We used the GloVe word to vector dataset to generate images for concepts and took advantage of its linear properties to generate new concepts. We analyse the results by comparing images of similar concepts and demonstrated that our system is capable of creating similar images for related concepts, distancing itself from different concepts.

Introduction

Even before the existence of a formal writing system, the human species developed ways to communicate knowledge using proto-writing which consisted of ideographic symbols that represented a limited number of concepts (Schmandt-Besserat, 2014).

In recent years, there have been proposals to design a universal written language based on ideograms. One of the examples is a system of Blissymbols proposed by (Bliss, 1965) which was conceived as a writing system where each basic symbol represents a concept and can be combined to represent new concepts. There also have been many approaches to study the origin and the evolution of language using computer simulations. However, the majority of these approaches are focused on the evolution of communication based on symbols (Sukhbaatar, szlam, and Fergus, 2016; Foerster et al., 2016; Mordatch and Abbeel, 2018) or textual communication (Das et al., 2017; Lewis et al., 2017).

In this paper, we propose a system that generates images that represent single concepts. We trained two neural networks to create a vocabulary composed of concept representative images through collaboration. In addition, we present a thorough analysis of the results obtained taking into account the word to vector linear properties.

Related work

There have been recent approaches to generate concept-representative symbols in the field of Computational Creativity. One example, the Emojinating system (Cunha et al., 2020) which generates visual representations of concepts through visual blending emoji. Xiao and Linkola (2015) also presented a similar system based on the combination of images. This system takes a conceptual task described in text and generates visual compositions taking into account the semantic associations and using different visual combination operations.

However, in all of these approaches the initial input, already exist and the system only generates a combination of them. Some approaches try to tackle these limitations by using machine learning techniques to generate images such as Generative Adversarial Networks (GAN) (Goodfellow et al., 2014).

These networks have been used to simulate the model of communication proposed by Shannon (1948). The author states that the general communication system can be divided into five elements, the information source which produces the message, the transmitter which creates the signal, the channel which is the medium that carries the information, the receiver transforms the signal received to the original message and the destination for which the message is intended.

Simon[1], Trenaman[2] and Murdock[3] used this model of communication together with GAN to develop a small set of experiments inspired by early proto-writing systems such as cuneiform and hieroglyphs. In these approaches, the encoder turns a message into an image that the decoder tries to decode back into the original message. Noise is added in the communication channel and increased over time so the encoder and decoder are encouraged to evolve the language to be robust to noise. Park (2020) presents a similar approach but focused on symbols. In this approach, the encoder is enforced to come up with a set of distinctive symbols that

[1] https://www.joelsimon.net/dimensions-of-dialogue.html, retr. 2021

[2] https://github.com/noahtren/GlyphNet, retr. 2021

[3] https://rynmurdock.github.io/2020/02/05/CCN.html, retr. 2021

resemble the human-made glyphs. The authors adopted a GAN-based neural painter trained with a synthetic brushstroke dataset so the encoder focuses on generating a set of symbols with diverse shapes that resemble human-made glyphs.

Approach

Some approaches use this type of model combined with neural networks to evolve a visual language, however, in our approach, we not only evolve a visual language, but each artefact that is transferred between networks has an underlying concept associated. However, our system aims to generate images that visually represent concepts by jointly training an encoder and a decoder to transfer a representation of a word through a noisy channel, following the Shannon model of communication (Shannon, 1948).

To obtain a vector representation of concepts, we used a word to vector dataset called Global Vectors for Word Representation (GloVe) (Pennington, Socher, and Manning, 2014) These representations showcase interesting linear substructures of the word vector space. For example, the underlying concept that distinguishes man from woman may be equivalently specified by various other word pairs, such as king and queen or brother and sister.

For the generation of the message to be transmitted, which in this case is an image, we used two neural networks based on the Deep Convolutional Generative Adversarial Network architecture (Radford, Metz, and Chintala, 2016), provided by Pytorch[4]. To implement the architecture proposed by Shannon (1948), we modified the decoder implementation so its output is a reconstruction of the input vector of the encoder. The training process is similar to the one used in autoencoders, however as the communication channel is noisy the signal received by the decoder differs from the original one.

First, the encoder tries to encode a set of word vectors from the GloVe dataset in RGB or grayscale images. Then, some type of noise is applied to the generated images based on a set of transformations. These transformations consist of a set of rotations, translations and normalization of the pixel data. Finally, the decoder tries to reconstruct the original vectors based on the images received.

The quality of the encoder and the decoder is assessed by evaluating how well the decoder is able to reconstruct the original vector. This way, the two networks are forced to cooperate to be able to converge to a vocabulary that is understandable by both.

Experimental setup

In this section, we describe the setup used in our experimentation. As previously mentioned, the encoder network is based on the original implementation of the DCGAN with some modifications on the last layers of the network.

The loss value is calculated using the mean squared error function which measures the average squared difference between the estimated values and the actual value. In our case, it measures the difference between the reconstructed latent vector z and the original vector used in the encoder. The networks were trained during 50 epochs. The batch size used was 64 and the learning rate was set to 0.001 which was divided by 10 at epoch 25, 35 and 43. Finally, the vector size used to set the size of the input layer of the encoder and the output layer of the decoder was set to 100 to be the same size as the word vectors of the dataset.

[4] https://pytorch.org/tutorials/beginner/dcgan_faces_tutorial.html

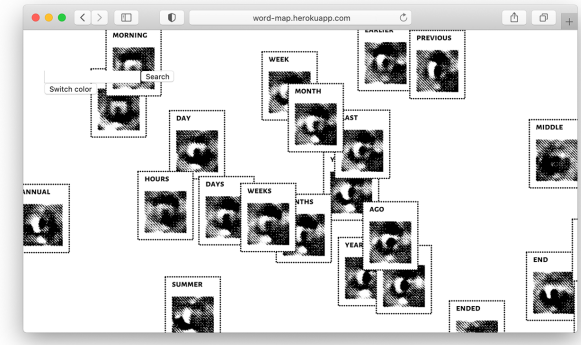

Figure 1: A visualisation of the website developed for this project.

Visualization Tool

To more easily explore the results obtained we developed a webpage [5] that presents the generated images. We used a t-distributed stochastic neighbour embedding (van der Maaten and Hinton, 2008) to transformed our higher dimensional vectors into a representation that we can visualize, in our case two dimensions to use as x and y value and created a 2D world by placing the images in the corresponding positions. Figure 1 shows a screenshot of the developed website.

On the website, it is also presented a version in which the images were created using three channels (red, green and blue). The website also provides a search tool to more easily find the images for each concept and compare them with other images and a different distribution based on the similarities between the images instead of the word vectors.

Results and discussion

Using the training process described in the previous section, we trained the networks to generate images for concepts taken from the GloVe dataset. Firstly, we removed the English stop words and words that may not have any concepts by themselves, such as 'he', 'she' and 'it'. In the end, we choose the 5120 most popular words of the dataset.

Analysis of similar and distinct words

Usually, the quality of autoencoders Makhzani et al. (2016) can be analysed based on the similarity of the vectors produced for each image. If the original images have strong similarities, the vectors produced must have similarities.

[5] available at http://word-map.herokuapp.com/

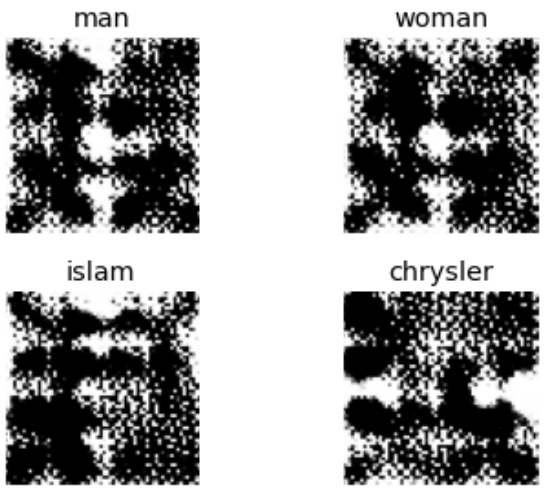

Figure 2: Comparison between similar images (‘man’ and ‘woman’) and different images (‘islam’ and ‘chrysler’).

However, when two images are different the autoencoder must be able to create distant vectors for each image. So, we decided to compare images of similar and different concepts to infer if the model is capable of creating similar images for related concepts while distancing them from different concepts.

Figure 2 presents two pairs of images, a similar pair (‘man’ and ‘woman’) and a different pair (‘islam’ and ‘chrysler’). As it is possible to observe, the two images that represent the words ‘man’ and ‘woman’ present very similar characteristics. Even though they are two antonyms, the context where they emerge is similar which results in similar images. The word to vector training is focused on the word associations, not on the meaning of the words, therefore, the word vectors that are closer are the ones that emerge in similar contexts rather than similar words. So, it is expected to find significant similarities between the images that represent ’man’ and ’woman’. The second pair (‘islam’ and ‘chrysler’) was selected based on the two words from the dataset with the biggest distance between them. As it is possible to observe, the two images are very different from each other, which is expected as they have very different word vectors. This shows that our model can emerge a vocabulary that approximate images that represent similar concepts while distancing images for concepts that are not related.

Analysis based on vector operations

One of the properties of a word to vector architecture is the linear substructures of the word vector space which can capture multiple different degrees of similarity between words. Mikolov, Yih, and Zweig (2013) found that semantic and syntactic patterns can be reproduced using vector arithmetic. Patterns such as man are to woman as king is to queen can be generated through algebraic operations on the vector representations of the words. In this analysis, we investigated if the semantic and syntactic patterns of the word to vector can be reproduced in our model by obtaining vectors of concepts that should be similar to existing concepts in the dataset. Then, we produced the images for both vectors and compare them visually. This way, we can infer the quality of the model on the generation of concepts that may not be available in the original dataset.

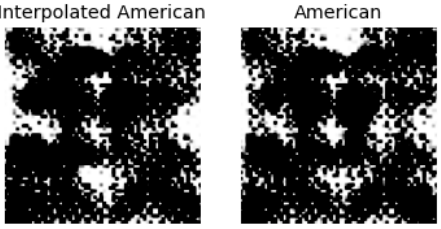

Figure 3: Comparison between the images generated with the ‘american’ produced using vector arithmetic and the ‘american’ from the dataset.

In the first experiment, we calculated the distance from the vector that represents the word ‘china’ to the vector that represents the word ‘chinese’, which represents the distance that goes from the country to the inhabitant in that country. Then, we created another word vector by adding this distance to a vector that represents a different country and generated the corresponding image. We used the following formula to synthetically create the vector for the word ‘american’ (referred to as ‘interpolated american’) and Figure 3 presents the results obtained.

$$chinese - china + america = american$$

As it is possible to observe in Figure 3 the image generated using the real vector (‘american’) and the image generated using the word vector created using the method previously described (‘interpolated american’) are very similar, which indicates that our model can produce images similar to ones of existing vectors through vector operations. This can be used to produce images for concepts that are not available in the dataset.

We expanded our analysis beyond countries and nationalities to assess if these properties can also be observable in verbal tenses. First, we calculated the distance that goes from a verb in the present tense to its past tense, for example from ‘go’ to ‘went’, and synthetically created past tenses for other verbs.

Figure 4 presents the comparison between both images, the image generated using the vector that we synthetically created and the image generated by the original vector that represents the word ‘took’. As it is possible to observe, our model is also able to generate images to represent the past tenses using the present tense. This might be useful when the past tense of a verb is not available but the present tense is, so we can generate the word vector synthetically and then the image that represents it.

Analysis based on the visualization tool

With the help of the visualisation tool, it is possible to observe some behaviours that can be unexpected as some con-

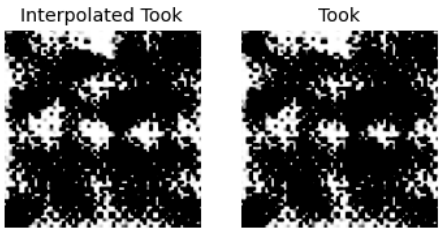

Figure 4: Comparison between the images generated with the 'took' produced using vector arithmetic and the 'took' from the dataset.

cepts may have similar images while for us they may not be related. One example of this behaviour is related to the word 'one'. In our visualisation tool, it is possible to observe some groups of words that are formed with related words, for example, the group related to sports, where words such as 'football', 'team', 'player' are located separated from the rest. The same behaviour is observed in the numbers. However, the word 'one' is not in the same place as the other numbers. As the word 'one' emerges more associated as a single unit or individual, like the word 'only' or 'another'. So, as it is more used in these contexts, the work 'one' is placed farther from the other numerals.

Conclusion and Future Work

Over the past years, we observed the adoption of computer models and artificial intelligence on language evolution. In our approach, we used two neural networks to evolve communication protocols based on images where each image represents a different concept, where similar concepts lead to images with similarities. We also explored the linear properties of the word vector by using vector arithmetic to create images for concepts. This approach proved to be useful to generate concepts that the dataset does not contain or even new non-existing concepts using the available concepts.

One of the limitations of the developed approach is the low resolution of the generated images. As the output size of the neural network is 64x64 pixels. Also, the images could be adapted to use graphic objects instead of raw pixel data. This way the images would be scalable without loss of quality and could be adapted to create a writing system more similar to the one we use.

It is also important to highlight the potential of our approach, it evolves a communication medium based on the visualization of a word embedding. It can be used in the generation of visually aesthetic QR codes which can only be deciphered using the correct decoder. It is also possible to add constraints to the generated images to enforce the emergence of customizable characteristics, such as controlling the amount of the black or white colour in each image.

However, this approach is focused on computer to computer communication that humans can observe and analyse. It represents the first step toward the creation of systems capable of evolving language, where we can use abstractions from visual features of the world.

References

Bliss, C. 1965. *Semantography (Blissymbolics): A Logical Writing for an Illogical World*. Semantography Blissymbolics Publ.

Cunha, J. M.; Lourenço, N.; Martins, P.; and Machado, P. 2020. Visual blending for concept representation: A case study on emoji generation. *New Generation Computing 38 (4), 739-771*.

Das, A.; Kottur, S.; Moura, J. M. F.; Lee, S.; and Batra, D. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. 2970–2979.

Foerster, J.; Assael, I. A.; de Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks. *Advances in Neural Information Processing Systems* 3.

Lewis, M.; Yarats, D.; Dauphin, Y.; Parikh, D.; and Batra, D. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proc. of the 2017 Conf. on Empirical Methods in Natural Language Processing*, 2443–2453. Copenhagen, Denmark: Association for Computational Linguistics.

Makhzani, A.; Shlens, J.; Jaitly, N.; and Goodfellow, I. 2016. Adversarial autoencoders. In *Int. Conf. on Learning Representations*.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. *Linguistic Regularities in Continuous Space Word Representations*. Association for Computational Linguistics. 746–751.

Mordatch, I., and Abbeel, P. 2018. Emergence of grounded compositional language in multi-agent populations. In *AAAI*.

Park, S.-w. 2020. Generating Novel Glyph without Human Data by Learning to Communicate.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Association for Computational Linguistics.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR* abs/1511.06434.

Schmandt-Besserat, D. 2014. The evolution of writing. *Denise Schmandt-Besserat. University of Texas at Austin* 25.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal* 27(3):379–423.

Sukhbaatar, S.; szlam, a.; and Fergus, R. 2016. Learning multiagent communication with backpropagation. In

Advances in Neural Information Processing Systems, volume 29. Curran Associates, Inc.

van der Maaten, L., and Hinton, G. 2008. Viualizing data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605.

Xiao, P., and Linkola, S. M. 2015. Vismantic: Meaning-making with images. In *Proc. of the Sixth Int. Conf. on Computational Creativity*.