



Generative Adversarial Networks for text-to-face synthesis & generation: A quantitative–qualitative analysis of Natural Language Processing encoders for Spanish

Eduardo Yauri-Lozano ^a, Manuel Castillo-Cara ^{b,*}, Luis Orozco-Barbosa ^c, Raúl García-Castro ^d

^a Universidad Nacional de Ingeniería, Lima, Peru

^b Universidad Nacional de Educación a Distancia, Madrid, Spain

^c Universidad de Castilla-La Mancha, Albacete, Spain

^d Universidad Politécnica de Madrid, Madrid, Spain

ARTICLE INFO

Keywords:

Image synthesis
CelebA dataset
RoBERTa transformer
Spanish
cDCGAN
Text-to-face generation
Text-to-image synthesis

ABSTRACT

In recent years, the development of Natural Language Processing (NLP) text-to-face encoders and Generative Adversarial Networks (GANs) has enabled the synthesis and generation of facial images from textual description. However, most encoders have been developed for the English language. This work presents the first study of three text-to-face encoders, namely, the RoBERTa pre-trained model and the Sent2Vec and RoBERTa models, trained with the CelebA dataset in Spanish. It then introduces customised and fine-tuned conditional Deep Convolutional Generative Adversarial Networks (cDCGANs) trained with the CelebA dataset for text-to-face generation in Spanish. To validate the results obtained, a qualitative evaluation was carried out with a visual analysis and a quantitative evaluation based on the IS, FID and LPIPS metrics. Our findings show promising results with respect to the literature, improving the numerical metrics of FID and LPIPS by 5% and 37%, respectively. Our results also show, through a quantitative–qualitative comparison of the cDCGAN training epochs, that the IS metric is not a reliable objective metric to be considered in the evaluation of similar works.

1. Introduction

In recent years, advances in Artificial Intelligence (AI) for text-to-image synthesis have seen enormous growth in novel techniques with promising results (Deorukhkar, Kadamala, & Menezes, 2022; Tao et al., 2022). Specifically, the development of AI techniques for the synthesis and generation from text-to-face has been in constant expansion (Agnese, Herrera, Tao, & Zhu, 2020; Nasir et al., 2019). To this end, Generative Adversarial Networks (GANs) are increasingly used in various computer vision tasks, including the generation of realistic images (Goodfellow et al., 2020). This task has several potential applications, such as creating realistic avatars, generating images of missing persons, and improving the quality of facial sketches used by law enforcement agencies (Ma et al., 2020).

Despite significant progress in the field, there are still many challenges, including the need to balance the accuracy of the images generated with their diversity and the difficulty of obtaining large-scale datasets of textual descriptions and corresponding facial

* Corresponding author.

E-mail addresses: eduardo.yauri@uni.pe (E. Yauri-Lozano), manuelcastillo@dia.uned.es (M. Castillo-Cara), luis.orozco@uclm.es (L. Orozco-Barbosa), r.garcia@upm.es (R. García-Castro).

URL: <http://www.manuelcastillo.eu> (M. Castillo-Cara).

<https://doi.org/10.1016/j.ipm.2024.103667>

Received 4 October 2023; Received in revised form 18 December 2023; Accepted 20 January 2024

0306-4573/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

images (Zhang et al., 2017; Zhao, Xie, Wang, Cao, & Zhang, 2019). The former is mainly due to GANs being unable to process the desired characteristics for the images generated (Agnese et al., 2020; Qin et al., 2022). In response to this deficiency, conditional Generative Adversarial Networks (cGANs) that comprise additional inputs to the original GANs are being developed (Goodfellow et al., 2020). The new inputs enable the model to be trained with additional information, such as class labels or other conditioning variables (Li et al., 2019). Unlike common GANs, which use dense intermediate layers in their generator and discriminator, cDCGANs use convolutional and deconvolutional ones known as transposed convolutional layers (Vasquez-Espinoza, Castillo-Cara, & Orozco-Barbosa, 2021). Hence, cDCGANs have made significant progress in two main domains: (i) the use of Convolutional Neural Networks (CNN) instead of fully-connected networks, as they are more suitable for images (Agnese et al., 2020; Talla-Chumpitaz, Castillo-Cara, Orozco-Barbosa, & García-Castro, 2023); and (ii) the development of various techniques to generate the information vectors that enter the network for the generation of synthetic images (Parmar, Zhang, & Zhu, 2022).

Furthermore, sentence embedding techniques have radically changed the field of Natural Language Processing (NLP) in recent years because they make it possible to encode text fragments as fixed-size vectors (Guan, Mondal, Dai, & Bao, 2023; Pagliardini, Gupta, & Jaggi, 2018). The development in this area has been constant and intensive in implementations, e.g. Sent2vec (Zhao et al., 2021). They have provided a major boost to many fields of AI, most notably conditional image synthesis, because the only way to feed information into neural networks is by mapping it onto vectors of real numbers (Pagliardini et al., 2018).

To address the limitations mentioned above, the challenge in text-to-image generation for spoken/written portraits and the progress of the previously described deep learning models (Vasquez-Espinoza et al., 2021) have motivated the implementation of a generative architecture using Sentence-BERT (SBERT) as an encoder (Reimers & Gurevych, 2019). This architecture includes the mechanisms to use descriptive text as input to be processed by an encoder and a cDCGAN model to generate synthetic images of faces (Zhang et al., 2017; Zhao et al., 2019). Furthermore, the task describes the implementation of images based on a textual description of the physical characteristics in Spanish as input. For instance, the architecture consists of comparing different Spanish-trained encoders, i.e., Sent2vec and SBERT, and a cDCGAN as a deep learning-based generative model (Parmar et al., 2022).

The present work implements the encoder by performing a comparative study between Sent2vec, and RoBERTa-large-bne¹ (Fandiño et al., 2022) (here on in referred to as RoBERTa) as a pre-trained model (Liu et al., 2019). For the transformer, customised training was carried out to increase the performance and generalisation of the model (Ding et al., 2023; Qin et al., 2022). In the training process, the facial feature descriptor text corpus, the CelebA dataset² (Xia, Yang, Xue, & Wu, 2021) (here on in referred to as CelebA), was used to translate it into Spanish. Finally, a generative model cDCGAN was developed, which was quantitatively evaluated using Inception Score (IS) (Barratt & Sharma, 2018), Frechet Inception Distance (FID) (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang, Isola, Efros, Shechtman, & Wang, 2018) metrics. Then it was qualitatively evaluated by visualising the synthetic images generated. Note that the customised encoder training was carried out in Spanish (Fandiño et al., 2022). In fact, this is one of the main contributions of this work, since to the best of the authors' knowledge, there is no pre-trained model for this specific task in Spanish.

Consequently, this work presents a quantitative (numerical) and qualitative (visual) comparison between Sent2vec trained with CelebA (here on in referred to as Sent2vec+CelebA), the RoBERTa baseline model (Liu et al., 2019) and our own trained model, taking RoBERTa baseline model with CelebA (here on in referred to as RoBERTa+CelebA) in the generation of sentence embedding vectors (Tao et al., 2022). These encoders are evaluated in the implemented cDCGAN and can be used as a valuable tool to develop the spoken portrait of a person, having multiple areas of application (Ma et al., 2020). In addition, the development of both the cDCGAN and the encoders is conducted in Spanish (Fandiño et al., 2022). The contributions of this paper are as follows.

- We perform a customised training of the RoBERTa and Sent2vec encoders with the CelebA corpus developed in Spanish.
- We carry out a comparative evaluation of the encoders to determine the best settings for the cDCGAN.
- We conduct quantitative (IS, FID and LPIPS) and qualitative (visual) analysis of the synthetic images generated by cDCGAN as a function of the encoder: Sent2vec+CelebA, RoBERTa and RoBERTa+CelebA.
- We demonstrate that IS is an unreliable metric for text-to-face generation and should be replaced by more robust metrics, such as FID and LPIPS.
- We show that our promising qualitative–quantitative results improve on those in the literature in terms of text-to-face generation. Furthermore, this is the first work in Spanish in this research area.

The remainder of this paper is organised as follows. Section 2 reviews the recent GAN for the text-to-image synthesis literature in three sections: Sentence embedding, text-to-image synthesis, and text-to-face generation. Moreover, we discuss how this work has developed with respect to advances in these areas. Section 3 specifies the dataset used in this investigation, the numerical evaluation metrics in the cDCGAN, and the guidelines. Subsequently, Section 4 describes, step-by-step, the details of the phases used to train the encoders. This section also covers the implementation and evaluation of the cDCGAN model. Section 5 presents our first set of results using quantitative metrics, i.e., IS, FID, and LPIPS, and a qualitative evaluation through a visual analysis of the generated images. Finally, Section 6 presents our conclusions and future work directions.

¹ <https://huggingface.co/PlanTL-GOB-ES/roberta-base-bne>

² <https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>

2. Related work

In this section, we review related work on the synthetic generation of images from textual description using GANs (Xia et al., 2021; Zhang et al., 2017). This review serves as a starting point and a contribution in developing the work presented, which covers the solution of problems in several areas of study of this research, such as word embedding, text-to-image synthesis and text-to-face generation.

2.1. Sentence embedding

Sentence embedding emerged as an extension of word embedding methods. In 2017, the InferenceNet algorithm outperformed most unsupervised methods, such as Skip-Thought (Conneau, Kiela, Schwenk, Barrault, & Bordes, 2017). Its goal was to find a directional relationship between text fragments through textual implications under labels. Its architecture consists of: (i) a sentence encoder that takes word vectors and generates their respective encoded vector; and (ii) a Natural Language Inference (NLI) classifier that takes the encoded vectors and generates a class label that may be the implication, contradiction or neutral. In 2018, the Universal Sentence Encoder (USE) trained a transformer network or a Deep Averaging Network (DAN) and improved unsupervised learning by training with the Stanford Natural Language Inference (SNLI) database.

In 2019 Sent2vec combined the techniques and fundamentals of Word2vec and FastText algorithms, combining the Continuous Bag of Words (CBOW) approach and n-grammes to generate the result vector (Pagliardini et al., 2018). During training, an unsupervised method that was simple, efficient and fast was used because of its low computational complexity. The training results showed that it outperformed most unsupervised and supervised models, with the robustness of the results being highlighted. In the same year, the transformer-based SBERT encoder enabled the generation of high-quality sentence embeddings compared to the models studied previously (Reimers & Gurevych, 2019). SBERT generates sentence embeddings using a Siamese and triplet network during training (Reimers & Gurevych, 2019). Its architecture consists of positional coding, self-attention and multi-headed attention.

2.2. Text-to-image synthesis

All work related to text-to-image synthesis was focused on the development and refinement of the text encoder and GANs. To this end, StackGAN was presented, which was able to generate images conditional on descriptive sentences passed to the model (Zhang et al., 2017). The model divides all tasks into a set of subtasks that are easier to perform through a sketch-refinement process. Phase 1 sketches the primitive shape and colours of the object based on the text description, which generates low-resolution images. Phase 2, takes the results of the previous phase and, together with the text descriptions as input, generates high-resolution images with high-quality details. In this phase, defects in the Phase 1 results can be rectified and convincing details can be added to the refinement process.

In the same year, AttnGAN improved on the StackGAN model in multi-stage refinement. This model can synthesise fine-grained details in different subregions of the image, paying attention to the significant words in the description (Xu et al., 2018). Moreover, a Deep Attentional Multimodal Similarity Model (DAMSM) was proposed to compute a fine-grained image-text matching loss to train the generator. This model learns from two neural networks that map subregions of images and sentence words to a common semantic space. This novel technique allows us to measure the similarity of image and text at the word level to calculate a fine-grained loss for image generation. The encoder used is an LSTM that extracts semantic vectors from the text description. In the bidirectional LSTM, each word corresponds to two hidden states, one for each direction. Thus, its two hidden states are concatenated to represent the semantic meaning of a word.

Finally, MirrorGAN achieved significant visual realism and semantic coherence compared to the models previously studied (Qiao, Zhang, Xu, & Tao, 2019). The architecture proposed leverages the idea of learning text-to-image generation through the description. It consists of three modules: (i) a Semantic Text Embedding Module (STEM) with an RNN; (ii) a Global-Local collaborative Attention Module (GLAM) in cascaded image generators; and (iii) a Semantic Text Regeneration and Alignment Module (STREAM). Hence, STEM generates embedding vectors at the word and sentence level, and GLAM has an architecture to generate target images from coarse to fine scales. Leveraging both local word and global sentence attention to progressively improve the diversity and semantic consistency of the images generated, STREAM seeks to regenerate the text description from the generated image, which is semantically aligned with the given text description.

2.3. Text-to-face generation

Generating face images is more complex and error-prone than generating other types of images, and, hence, a number of models and applications have been developed for this task.

In this area, a significant advance was presented in 2019, proposing a different approach to the traditional way of generating conditional faces (Zhao et al., 2019). In this approach, sketches were used instead of full images which, together with the textual description, were the inputs to the generator. The contributions were the following: (i) the process of completing the sketch to generate the real image is similar to the process of face super-resolution reconstruction; (ii) semantic features are extracted from the attribute vector and resized to the same size as the input image sketch; (iii) by using the Skip-connection method, the authors succeeded in reducing the number of layers of the generative network without losing efficiency, albeit increasing the computational

Table 1

Comparison of resources used by the literature in terms of face generators, Encoder and Neural Network Architecture, numerical metrics, Inception Score (IS), Frechet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS), number of images used in the testing phase (No. images) and the language used as a base. The last three rows (*This work*) show the characteristics and results of the work presented in this article. The best results are shown in bold.

Article	Encoder	GAN Arch.	IS	FID	LPIPS	No. images	Language
Zhao et al. (2019)	RNN	GLAM & STREAM	–	–	–	–	English
Nasir et al. (2019)	Skip-Thought	cDCGAN	1.4	–	–	–	English
Xia et al. (2021)	LSTM bidirectional	DFGAN	–	137.60	0.581	50	English
Xia et al. (2021)	LSTM bidirectional	DMGAN	–	131.05	0.544	50	English
Oza et al. (2021)	Skip-Thought	cDCGAN	–	128.46	0.590	250	English
Oza et al. (2021)	LSTM	DAMSM	–	125.98	0.512	250	English
Oza et al. (2021)	RNN	AttGAN	–	116.32	0.522	250	English
Oza et al. (2021)	Visual-linguistic similarity module	StyleGAN	–	106.37	0.465	50	English
Oza et al. (2021)	BERT	ACM	–	105.73	0.449	250	English
Deorukhkar et al. (2022)	SBERT	cDCGAN	2.732	90.268	–	5	English
Deorukhkar et al. (2022)	SBERT	SAGAN	2.855	95.052	–	5	English
Deorukhkar et al. (2022)	SBERT	DFGAN	3.455	88.748	–	5	English
<i>This work</i>	Sent2vec+ CelebA	cDCGAN	2.738	105.561	0.3036	1000	Spanish
<i>This work</i>	RoBERTa	cDCGAN	2.714	89.3567	0.2932	1000	Spanish
<i>This work</i>	RoBERTa+ CelebA	cDCGAN	2.789	84.221	0.2851	1000	Spanish

complexity; and (iv) the architecture uses a subframe structure (A for image input and B for attribute input) in the feature extraction phase, allowing the generated image to behave better in terms of texture, colour and structure.

In the same year, a cGAN called Text2faceGAN generated face images from fine-grained textual descriptions. For the tests, CelebA images and a set of descriptor texts were implemented for each image using a proprietary algorithm (Nasir et al., 2019). This research pioneered the use of an encoder for embedding whole sentences using Skip-Thought. The main contributions focus on the preparation of training data, the proposed GAN architecture, and the efficiency measure defined for this particular case.

Subsequently, in 2021, a unique architecture, Semantic Text-to-Face GAN, started from generic text entries and allowed them to be modified from another auxiliary text entry using the same network (Oza, Chanda, & Doermann, 2021). The architecture proposed consists of two complementary blocks and CelebA. The first block generates low-resolution images from a text description encoded with an encoder-decoder network composed of a ResNet block and an Affine Combination Module (ACM) structure to link inputs. In the second block, the generator encodes the low-resolution input images into a feature vector using Inception-v3 and then concatenates them with textual feature vectors encoded by BERT. Finally, it has an image-generating network to produce the modified output with the desired resolution. Using the FID, LPIPS, and Accuracy and Photorealism metrics, the results were compared with other models, such as TediGAN (Xia et al., 2021) or Text2FaceGAN (Nasir et al., 2019); the best results were observed in the latter. Likewise,

Finally, AnyFace (Sun et al., 2022) presented a novel two-stream framework for face image synthesis and manipulation given arbitrary descriptions of the human face. Specifically, one stream performs text-to-face generation and the other conducts face image reconstruction. Additionally, Deorukhkar et al. (2022) compared the performance of three GAN models in synthetic face generation, using FID and IS. The algorithms evaluated were cDCGAN, Self-Attention Generative Adversarial Network (SAGAN) and Deep Fusion Generative Adversarial Networks (DFGAN) (Tao et al., 2022) and CelebA. As in the Text2FaceGAN model, the authors created their own corpus with descriptive sentences using a proprietary algorithm from the attributes listed in the dataset. Finally, they used a pre-trained SBERT model in English to encode the descriptive sentences of the images in the dataset. The work showed that SBERT provided better results and quality images compared to the original BERT model and other algorithms such as Skip-Thought.

2.4. Implications

Dividing the study of the presented work into three main areas allows us to focus on the evaluation and improvements that have been added with each recent research. Table 1 summarises the model specifications and numerical results obtained in the literature and compares them with the present work (*This work* in Table 1). For these results, a comparative study was carried out with three different encoders under the same cDCGAN architecture. It is important to note that any numerical comparison between related work must be made under the same dataset for the encoder and the cDCGAN model; otherwise, the results will be completely different. Therefore, all the results shown were obtained with CelebA. Note that all work reviewed in the state of the art was conducted in English. Furthermore, to the authors' knowledge, the present work is pioneering for Spanish in this area.

Regarding the results, it is observed that, numerically, our best approximation, that is, RoBERTa+CelebA, achieves better results in the FID and LPIPS. Furthermore, lower FID and LPIPS scores and a higher IS score theoretically indicate the generated images are of a better quality. Nevertheless, as will be discussed in Section 3.2 and demonstrated in this paper, IS is not a reliable metric for text-to-face generation and should be replaced by more robust metrics, such as FID and LPIPS.

Furthermore, the column “No. images” shows the number of images generated to calculate the metrics score. In other words, all the evaluations performed by the models studied consist of applying IS, FID, and LPIPS on a generated dataset where the number of evaluated test images directly influences the numerical accuracy of the metrics. Therefore, the greater the number of images evaluated, the greater is the reliability of the metric with respect to the actual value given to a dataset. Hence, we can see that our work achieves fully reliable metrics because a larger number of images is used compared to previous works. In our case, our model under study, cDCGAN, was evaluated using 1000 images generated by the cDCGAN, obtaining great robustness in both quantitative and qualitative results.

In summary, our findings show promising results with respect to the literature by improving the quantitative metrics of FID and LPIPS by 5% and 37%, respectively. Likewise, in this work we use a greater number of evaluated images that consolidate the results obtained in Spanish. Additionally, the most relevant aspects of these investigations will be discussed below.

2.4.1. Sentence embedding

In the area of sentence embedding, all the works reviewed approach the problem in different ways. For example, [Conneau et al. \(2017\)](#) trained the encoder by entering the vector generated into a classifier network and comparing the result with the class labels defined. Meanwhile, [Pagliardini et al. \(2018\)](#) combined and expanded the Word2vec and FastText approaches, which, initially designed for words, generated complete sentence vectors. This study highlighted the ease and speed of training compared to the others. Finally, [Reimers and Gurevych \(2019\)](#), using transforms, achieved the best results to date in the performance and quality of sentence embedding. Following the line of using transformers, the present work improves the performance of the RoBERTa baseline model ([Fandiño et al., 2022](#)) by training it with the descriptive corpus of CelebA in Spanish.

2.4.2. Text-to-image synthesis

Text-guided face generation is a specialised task derived from the general task of creating images. Work on this began in 2016, with a number of GAN models having been created with different ways of encoding sentences that have improved over time. [Xu et al. \(2018\)](#) and [Zhang et al. \(2017\)](#) improved the encoding of the input text using an LSTM. The former used two processing phases, while the latter used a multimodal similarity model focused on subparts or regions of images paying attention to key features. Finally, [Qiao et al. \(2019\)](#) used a combination of three modules: an RNN to encode text, a network to generate images and a module to regenerate descriptive text from an image so that they were semantically aligned.

2.4.3. Text-to-face generation

The works focused on face generation improved the encoded descriptive text and the design of GANs. Hence, [Zhao et al. \(2019\)](#) considered developing an initial sketch in addition to the descriptor text as input into the network, using a proprietary algorithm to encode the sentences. Moreover, [Nasir et al. \(2019\)](#) employed an encoder called Skip-Thought and a cDCGAN based on convolutional networks, ReLU and LeakyRelu, and batch normalisation according to the basic GAN enhancement recommendations for image synthesis.

Finally, [Deorukhkar et al. \(2022\)](#) and [Oza et al. \(2021\)](#) highlighted the use of more recent transformers, such as BERT and SBERT, for the encoding of descriptive sentences. RestNet was used to obtain the latent space of the images, and Inception-v3 was used to obtain the feature vectors. Furthermore, [Oza et al. \(2021\)](#) proposed an architecture that allows images to be generated, which could be modified based on a subsequent description in the second step, while, [Deorukhkar et al. \(2022\)](#) numerically compared efficiency across several well-known GANs, using FID and IS.

3. Background: Dataset, metrics, tools and guidelines

The following section describes the main instruments used in the research, i.e., the dataset, metrics, hardware and guidelines.

3.1. Dataset

CelebA³ is a large-scale database of celebrity images widely used in text-to-face generation with GANs ([Deorukhkar et al., 2022](#); [Nasir et al., 2019](#); [Oza et al., 2021](#)). The images cover a wide range of postures, views and angles of each person ([Xia et al., 2021](#)). Another important feature is that different versions of the images can be downloaded. This work uses images of people in their “cropped and centred” version. It has 202,599 images of 10,177 different identities.

Furthermore, each image has a set of 40 physical attributes that allow it to be described. The attribute file contains a table of two possible values: 1 indicates that an attribute corresponds to the image, and -1 that it does not. The main descriptive characteristics include beard, hair colour, eyebrow type, face shape, hair style, eye colour and appearance-enhancing attributes.

3.2. GANs evaluation metrics

GANs have gained notable effectiveness in generating high-quality synthetic images. However, instead of being trained alone, as in most of these networks, the generator models are trained together with a second model, called a discriminator, which learns to differentiate real images from fake (generated) ones. Hence, there is no objective function or score for the generator or for the discriminator; rather, the images are evaluated after being generated with the metrics described ([Li et al., 2019](#)).

³ <https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>

3.2.1. Inception score

IS (Salimans et al., 2016) involves the use of a previously trained neural network model, InceptionV3 (trained with the ImageNET dataset) (Keras Official Documentation, 2023). The model classifies a large number of generated images and predicts the probability of an image belonging to each of the classes defined in ImageNET. The higher the IS score indicates, the higher is the image quality. Mathematically, the highest possible score is infinite; in practice, however, it is equal to the number of classes contained in the database where it was trained, being 1000 in 2014. It is defined by the following equation:

$$IS = \exp(\mathbb{E}_x KL(p(y|x) \| p(y)))$$

The following steps were implemented:

- We resized the input images to a size of $299 \times 299 \times 3$, required by the model.
- We used a single set of 1000 images generated by cDCGAN. The images are pre-processed and their label is calculated using InceptionV3.
- We calculated the initial mean and standard deviation scores using InceptionV3. The objective function takes an array of images with expected sizes and values of pixels in the range of 0 to 255.
- We calculated the IS score for five different sets of images and setting the final result by averaging them all.

However, IS is being replaced because it has certain disadvantages described in Barratt and Sharma (2018), Sommer and Iosifidis (2020) and Xu et al. (2018). Similarly, we also demonstrate that it is not a reliable metric for text-to-face generation.

3.2.2. Frechet inception distance

FID (Heusel et al., 2017) computes the distance between the feature vectors calculated for real and synthetic images and uses InceptionV3. The objective of FID is to evaluate the quality of a set of synthetic images in terms of their statistical indicators compared to a set of real images from the same target domain. The lower the FID score, the higher is the image quality. It is defined by the following equation:

$$FID = \left| \mu_r - \mu_g \right|^2 + T_r \left(\sum_r + \sum_g - 2 \left(\sum_r \sum_g \right)^{\frac{1}{2}} \right)$$

The Python library, `pytorch-fid` (Torch Metrics, 2023a), was used with the following considerations:

- Two sets of 1000 images were used. The first is that of the original images, and the second that of the images generated by cDCGAN. The InceptionV3 model is only used to extract the vector of features of the images.
- FID performs a total of five scores for five different sets of images and obtains the final result by averaging all the scores.

3.2.3. Learned perceptual image patch similarity

We evaluated LIPS (Zhang et al., 2018) in which, given two images x and x_0 , the similarity between two image slices is calculated using the cosine distance. The lower the LPIPS score, the higher is the image quality. It is defined by the following equation:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left| w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l) \right|_2^2$$

A Python library, `lpips` (Torch Metrics, 2023b), was used with the following considerations:

- We calculate the similarity between the activation of two images for a predefined Alex model.
- Each of the images is transformed into tensors using the `im2tensor` module before entering Alex. The final result is given by applying the mean and standard deviation to the set of results obtained from each of the pairs of images.
- Two sets of images are used for execution. The first one is the original images, and the second one is the images generated by cDCGAN. item LPIPS performs a total of five scores for five different sets of images and obtains the final result by averaging all the scores.

3.3. Tools

For the encoder training process designed, Sent2vec+CelebA and RoBERTa+CelebA, and the cDCGAN network, as well as the evaluation of quality metrics, we used a high-performance server located at the Albacete Research Institute of Informatics (I³A) of the Universidad de Castilla-La Mancha (UCLM). The main features are 220 GB of RAM, 500 GB of hard disk, 32 CPU cores, 1 NVIDIA Tesla T4 with 16 GB and 2 cores and Ubuntu Server 20.04.

3.4. Guidelines

Taking these works and background as a reference, the current experiment sought to generate synthetic images of human faces from a textual description in Spanish. Due to its ease of implementation and the balance between performance and complexity, the architecture implemented in Nasir et al. (2019) will be taken as the cDCGAN base.⁴ Furthermore, to improve efficiency, an encoder

⁴ <https://github.com/midas-research/text2facegan>

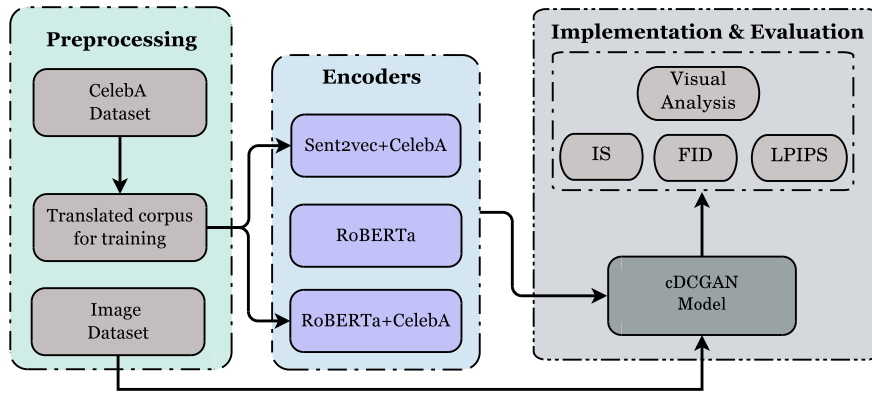


Fig. 1. Overall schema with the three main phases implemented in this work.

based on RoBERTa (Fandiño et al., 2022) was used, which was previously trained with the CelebA corpus in Spanish. Likewise, the encoders were evaluated as input to a cDCGAN in order to generate the faces of people based on a description of this, evaluating the model with IS, FID and LPIPS. Significant results were obtained, compared to those in the literature (see Table 1).

4. Methodology on image synthesis with cDCGAN

This section studies the methodology and the different phases implemented in this experiment. Furthermore, the pre-processing applied to the CelebA descriptive corpus and images is detailed, as well as the cDCGAN. Fig. 1 represents the methodology with the three main phases:

- Phase 1: Data pre-processing (see “Pre-processing” in Fig. 1).
- Phase 2: Training the encoders (see “Encoders” in Fig. 1).
- Phase 3: Implementation and evaluation of cDCGAN (see “Implementation & Evaluation” in Fig. 1).

4.1. Phase 1: Data preprocessing

In this first phase (see “Pre-processing” in Fig. 1), we prepared the corpus used to train the encoders, i.e., Sent2vec+CelebA and RoBERTa+CelebA, and cDCGAN. Note that the RoBERTa baseline model did not need a corpus as it is a pre-trained model.

4.1.1. Sent2vec+CelebA corpus

The pre-processing of the corpus for the encoders, i.e., Sent2vec+CelebA and RoBERTa+CelebA, was performed using a simple methodology shown in Fig. 2. For Sent2vec+CelebA, the first step was to translate the CelebA captions into Spanish with the algorithm used in Text2FaceGAN (Nasir et al., 2019) (see “Original dataset” and “Translated dataset” in Fig. 2). Subsequently, a new corpus with translated information was created with the same structure as the original English version.

Finally, a training corpus was generated for Sent2vec+CelebA by performing an information-cleaning process on the dataset generated in the previous step. In particular, all the sentences were combined to generate a larger corpus (see “Sent2vec final training corpus” in Fig. 2).

4.1.2. RoBERTa+CelebA corpus

RoBERTa+CelebA was trained using a Siamese network that evaluates the similarity of embeds generated by the transformer. For this purpose, the cosine similarity metric was used and compared with the similarity score in the training corpus. Each input of the training data consists of a pair of sentences A and B in Spanish and their respective similarity in the range of 0 to 1.

First, and like Sent2vec+CelebA, the original English text was translated into Spanish (see “Original dataset” and “Translated dataset” in Fig. 2). Subsequently, the document structure defines, in each line (input), a pair of Spanish sentences and their respective similarity value between 0 and 1 calculated by the Spacy library. However, since Spacy works only with English entries, the similarity between two Spanish sentences was matched with their respective English pairs (see “Spacy” in Fig. 2). Finally, the final training corpus for RoBERTa+CelebA is defined by the Spanish text and the English similarity score (see “RoBERTa final training corpus” in Fig. 2). To this end, we implemented Algorithm 1.

As a result, a corpus of 250,000 entries was composed of a pair of Spanish sentences and their respective similarity score was achieved. Subsequently, it was divided into training-validation splits.

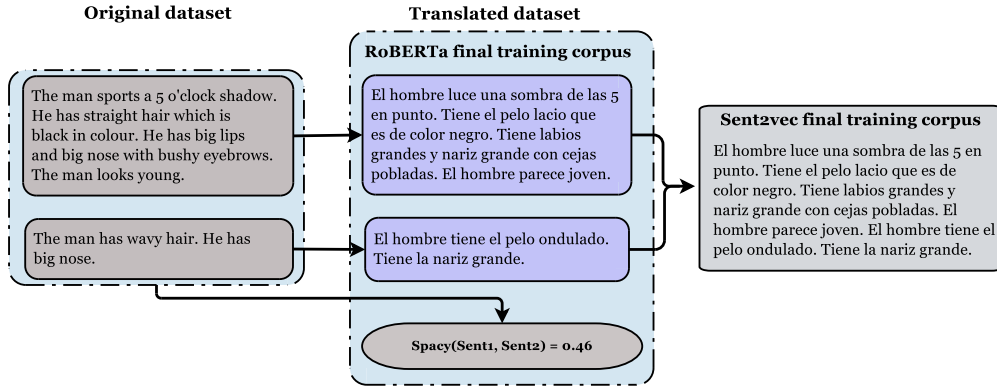


Fig. 2. Overall scheme of the corpus generation process for the Sent2vec+CelebA and RoBERTa+CelebA encoders.

Algorithm 1 Roberta+CelebA corpus generation

```

1: Input: Corpus of sentences in English and Spanish
2: Output: Transformer training corpus
3:  $ListInputSp \leftarrow \{Sent_{Sp_1}, Sent_{Sp_2}, Sent_{Sp_3}, \dots\}$ 
4:  $ListInputEn \leftarrow \{Sent_{En_1}, Sent_{En_2}, Sent_{En_3}, \dots\}$ 
5:  $Nmax \leftarrow 250000$ 
6:  $Emax \leftarrow 192050$ 
7:  $n \leftarrow 1$ 
8: while  $n \leq Nmax$  do
9:    $x \leftarrow random(0, Emax)$ 
10:   $y \leftarrow random(0, Emax)$ 
11:   $sentence1 \leftarrow ListInputEn[x]$ 
12:   $sentence2 \leftarrow ListInputEn[y]$ 
13:   $simil \leftarrow evaluateSimilarity(sentence1, sentence2)$ 
14:  Write :  $ListInputSp[x] + ListInputSp[y] + simil$ 
15:   $n \leftarrow n + 1$ 
16: end while

```

Table 2

Questions and answers for the attribute groups of an image. Note that the sentences were translated into Spanish for training/testing purposes, although this document shows them in English for better understanding.

Questions: Facial groups.	Answer: Facial attributes.
What is the structure of the face?	Chubby face, double chin, oval face, high cheekbones.
What facial hair does the person have??	5 o'clock shadow, goatee, moustache, sideburns.
What sort of hair does the person have??	Bald, straight hair, black hair, blond hair, brown hair, grey hair, fringes, wavy hair, receding hairline.
What is the description of the other facial features?	Large lips, large nose, pointed nose, narrow eyes, arched eyebrows, bushy eyebrows, slightly open mouth.
What are the attributes that enhance appearance?	Youthful, attractive, smiling, pale skin, rosy cheeks pale, rosy cheeks, lots of make-up.
What are the accessories worn by the person?	Earrings, hat, necklace, tie, glasses, lipstick.

4.1.3. cDCGAN corpus




The cDCGAN corpus was implemented by translating the corpus generated in Nasir et al. (2019) into Spanish. In order to form descriptive sentences from the initial attribute file, six groups of features were created in response to six questions that progressively describe the face, from contour features to appearance-enhancing facial features. These groups can be seen in Table 2.

Once the facial groups have been defined, the algorithm generates sentences for each of the groups. A queue is generated, in which each member of the list of attributes defined for a specific group is added. Additionally, connectors or intermediate words are added at the beginning of the sentence. As the corpus generated was originally written in English, each descriptive sentence was translated into Spanish.

The final translated corpus contained 192,248 descriptive sentences of images. During training, 10,351 faces did not have, so the label "This is a person and nothing else" (in Spanish: "Esta es una persona y nada más") was assigned, although this description was not written in the final corpus file. Table 3 shows a sample of CelebA images with their respective description in Spanish. It can be seen that the length and level of detail of the sentences are completely different in each sample.

Table 3

Sample of images and their respective description of characteristics translated into Spanish. Note that the sentences were translated into Spanish for training/testing purposes, although this document shows them in English for better understanding.

Image	Description
	The woman has wavy brown hair. She has large lips with arched eyebrows and a slightly open mouth. The attractive young woman is wearing lot of make-up. She is wearing earrings and lipstick.
	The man has straight hair with black tones. He is a handsome young man and is smiling.
	The chubby man with a double chin has high cheekbones. His hair is black in colour. He has a large nose, narrow eyes with bushy eyebrows and a slightly open mouth. The man is smiling. He is wearing a tie.

4.2. Phase 2: Training of the encoder

One important component of many deep learning architectures is an encoder that maps input data into a lower-dimensional representation. Training an encoder involves optimising its parameters to minimise a predefined objective function that measures the discrepancy between the encoded representation and the target output. For training, a subdivision was made into the two main encoders used in this work: RoBERTa and Sent2vec (see “Encoders” in Fig. 1).

4.2.1. Training of Sent2Vec+CelebA

Sent2vec can be used directly for English texts (Parmar et al., 2022). However, since the present work uses Spanish text, it was necessary to train it previously using the generated corpus (see Section 4.1) according to the following steps:

- Pre-processing the Spanish corpus. For this purpose, each of the entries of the original corpus was saved in a new file and other components, e.g. symbols, were removed. A total of 192,209 sentences were available for training.
- Applying a second pre-processing consisting of removing accents. Stop words and connectors were used as part of the sentence structure during training.
- Configuring the libraries, e.g., Sent2vec and FastText, and their parameters empirically, being: 4800 feature vector dimensions, 5000 epochs, 200 threads, 2 n-grammes and 0.05 learning rate.
- The loss function used for the Sent2Vec+CelebA encoder was skip-gram with negative sampling (Levy & Goldberg, 2014).

4.2.2. Training of RoBERTa+CelebA

In order to improve the performance of the RoBERTa encoder (SentenceTransformers Documentation, 2023), the model was trained using the corpus specified in Section 4.1. We used a Siamese network together with the cosine similarity loss function (Reimers & Gurevych, 2019) with the following actions:

- Dividing the corpus into 249,000 sentences for training and 1000 sentences for validation.
- Loading the training/validation data for the model. Two lists were generated for the information and, in each of them, the entries were composed of a pair of descriptive sentences and their similarity value.
- Implementing the RoBERTa encoder for training.
- Training with a Siamese network to evaluate the similarities of their generated embedding vectors. Subsequently, the metric was compared with the real similarity value obtained from the training corpus. The performance of the model during training was calculated using Spearman’s correlation between the real similarity vector and the calculated similarity vector. Note that the RoBERTa+CelebA encoder therefore does not use a loss function in its training phase (SentenceTransformers Documentation, 2023).

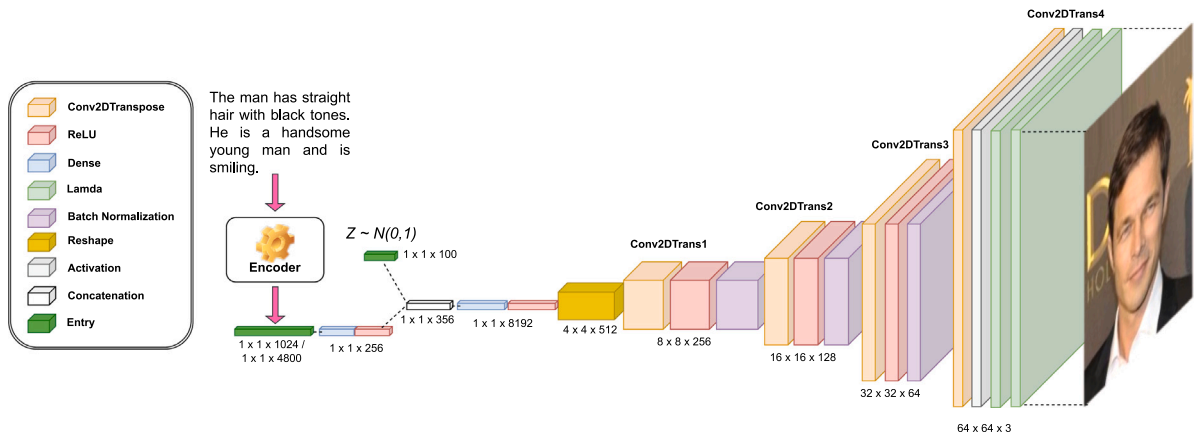


Fig. 3. Schematic of the generator architecture for the developed cDCGAN. Note that the sentences were translated into Spanish for training/testing purposes, although this document shows them in English for better understanding.

4.3. Phase 3: Training of the cDCGAN

Text-to-face synthesis using GANs involves training a generator network to generate an image that matches the given textual description, while a discriminator network is simultaneously trained to distinguish between real and generated images. The generator network learns to generate images that are as realistic as possible, while the discriminator network provides feedback to the generator by evaluating the quality of the generated images. This feedback loop allows the generator to continuously improve its output until it produces images that are indistinguishable from real ones. Therefore, the cDCGAN architecture implemented is based on the Text2faceGAN model (Nasir et al., 2019). Nevertheless, it was empirically redesigned in training to maximise the results (see “Implementation & Evaluation” in Fig. 1). It consists of two integrated underlying architectures: a generator and a discriminator.

4.3.1. Generator architecture

The generator network consists of a series of transposed convolution layers together with activation layers with the following components (see Fig. 3):

- The network has two inputs, the first for the vector of features generated by the encoder and the second for the noise vector. The dimension for RoBERTa is 1024 and for Sent2vec is 4800; in both cases, through a dense layer, its dimension is lowered to 256. The second input has a fixed dimension of 100.
- The two input layers are concatenated and pass through a dense layer of 512 neurons before entering the intermediate layers.
- It has four transposed convolution layers, with 256, 128, 64 and 3 filters, respectively, to implement the upsampling of the images.
- Between the dense and transposed convolution layers, the ReLU activation layers are set with their default parameters.
- A batch normalisation layer with momentum = 0.2.
- All transposed convolution layers have initialisation of weights to a normal distribution.
- TanH activation layer after the initial concatenation layer and the last transposed convolution layer.
- Two Lambda layers: the first performs a division by 2 on the output tensor, and the second performs an addition of 0.5 to the tensor.
- The generator uses the sigmoid activation, Adam optimiser with learning_rate = 0.0002 and beta1 = 0.5 and loss = binary_crossentropy.
- Finally, the output generates $64 \times 64 \times 3$ images that match the description entered into the network.

Therefore, during training, cDCGAN generates images according to the textual description indicated by the user. Subsequently, these images are passed on to the discriminator network.

4.3.2. Discriminator architecture

The second component of our cDCGAN is the discriminator network (see Fig. 4). This network interacts with the generator, since it is in charge of evaluating the quality of the generated images, and has the following components:

- Like the generator model, this model has 2 inputs. The first receives the image created by the generator, which has dimension of $64 \times 64 \times 3$; and the second is the vector of features of the text whose dimension is 1024 for RoBERTa and 4800 for Sent2vec.

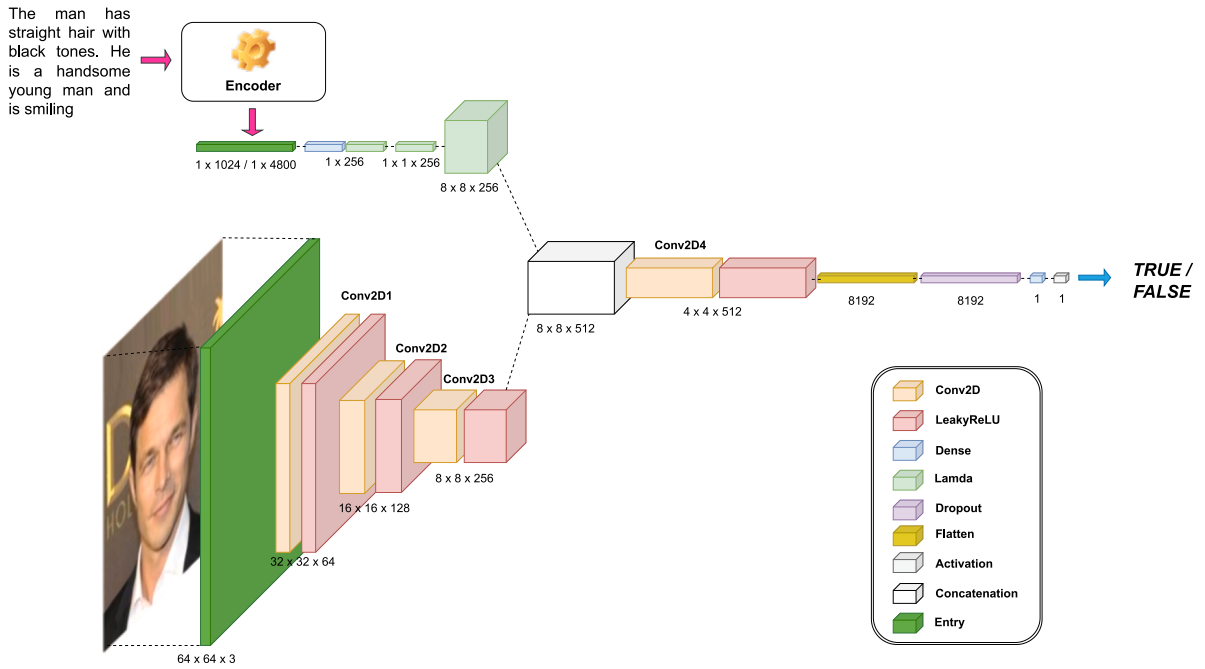


Fig. 4. Schematic of the discriminator architecture for the developed cDCGAN. Note that the sentences were translated into Spanish for training/testing purposes, although this document shows them in English for better understanding.

- In the first input, it passes through three normal convolutional layers, with filters of 64, 128 and 256, respectively, to implement the downsampling of the input image.
- Between each of the convolutional layers, the LeakyReLU layers are added with their default parameter values.
- For the second input, the feature vector passes through one dense layer, to decrease its length, and three Lambda layers, to change its dimension.
- The inputs are concatenated and pass through a 512-filter convolutional layer and a LeakyReLU layer.
- Then it passes through a Flatten layer to prevent overfitting at training time, and a dense layer to classify the image as True/False.
- Finally, the generator uses the sigmoid activation, Adam optimiser with $\text{learning_rate} = 0.0002$ and $\text{beta1} = 0.5$ and $\text{loss} = \text{binary_crossentropy}$.

Once the cDCGAN was studied, the quality of the generated images was evaluated quantitatively (numerical results) and qualitatively (visual analysis).

5. Experimental results & evaluation

The following section shows the quantitative and qualitative results obtained by the cDCGAN, taking into account the three encoders under study: Sent2vec+CelebA, RoBERTa and RoBERTa+CelebA.

5.1. Encoder models

The first phase is the training of the encoders with CelebA, i.e., Sent2vec+CelebA and RoBERTa+CelebA.

5.1.1. Sent2vec+CelebA

In contrast to RoBERTa+CelebA, in which the RoBERTa baseline model is used and the tuning is performed with CelebA, Sent2vec+CelebA, training must be performed from scratch. For this purpose, the training lasted 7 h (see Table 5) with the hardware defined in Section 3.3 and the specifications given in Section 4.2. Note that the training process had to be carried out in Spanish, as was done with RoBERTa.

5.1.2. RoBERTa+CelebA

In the case of RoBERTa+CelebA, the RoBERTa baseline model was taken and the tuning was performed with CelebA. For this purpose, the training was carried out with 145 epochs according to the specifications given in Section 4.2, having a duration of 42 days (see Table 5) with the hardware defined in Section 3.3.

Table 4

Results of Spearman's correlation coefficient for the transformers. The best result is shown in bold.

Model	Epoch	Spearman correlation
RoBERTa	Baseline model	0.827176427
RoBERTa+CelebA	-	0.811153072
RoBERTa+CelebA	74	0.999913276

Table 5

Training computation time of the different models under study. The first two columns show the computation time in the training phase carried out by the encoders, and the next three columns show the computation time in the training phase carried out by cDCGAN with the encoders.

Encoders		cDCGAN with the encoders		
Sent2Vec+CelebA	RoBERTa+CelebA	Sent2Vec+CelebA	RoBERTa	RoBERTa+CelebA
7 h	1008 h (42 days)	98 h 11 min (4.1 days)	98 h 02 min (4.1 days)	99 h 31 min (4.15 days)

In order to compare with RoBERTa, an evaluation was made with the Spearman's correlation coefficient between actual vector and the calculated vector using the cosine similarity for 1000 test sentences. Table 4 shows that RoBERTa+CelebA receives a lower correlation than RoBERTa at the beginning of training. However, as training progresses, it is observed that RoBERTa+CelebA achieves a better coefficient and thus improves the performance of the encoder.

5.2. cDCGAN model

For training, cDCGAN used the training corpus described in Section 4.1 and the architecture described in Section 4.3. For the final evaluation, the model was trained using the Sent2vec+CelebA, RoBERTa and RoBERTa+CelebA encoders. All training and evaluation iterations were performed on the server described in Section 3.3.

Some important assumptions were taken into account in the training: (i) setting the number of epochs between 50 and 600 with intervals of 50; (ii) using 50,000 images in the training, instead of 202,599; this is an optimal size with respect to the hardware used, time and results obtained; and (iii) setting the batch size to 512. The training time was very similar across the three encoders, e.g., for epoch 600 Sent2vec+CelebA took 98 h 11 min (4.1 days), RoBERTa took 98 h 02 min (4.1 days), and RoBERTa+CelebA took 99 h 31 min (4.15 days) (see Table 5).

Once the three models have been trained, a quantitative and qualitative evaluation of the generated synthetic images is carried out as described in Section 3.2.

5.3. Metric evaluation

To implement the numerical evaluations, 1000 images were taken randomly together with their respective descriptive sentences (see Section 3.2). Note that an adequate amount of data must always be used to obtain reproducible results, so while a smaller number of images may seem sufficient, in our case we use 1000 to add confidence and robustness to the models and their final quantitative results (see Table 1). Fig. 5 shows the variation of the IS, FID and LPIS with respect to the number of epochs:

- IS (see Fig. 5(a)): It can be observed that RoBERTa+CelebA produces a better result than Sent2vec+CelebA and, in turn, is better than RoBERTa. It can also be seen that Sent2vec has a high divergence from epoch 150. When comparing the transformers, RoBERTa shows a small improvement compared to RoBERTa+CelebA in the final epochs, although RoBERTa+CelebA has a metric stabilisation and RoBERTa has a high divergence. Comparison of these numerical results with the visual analysis (see Table 7) confirms that IS is not a fully accurate metric, as discussed in Section 3.2, since there is no direct relationship between a high IS score and the quality of the images. Note that the higher the IS score, the better is the quality of the images generated.
- FID (see Fig. 5(b)): It can be seen that the transformers produce better results than Sent2vec+CelebA, which has a high divergence in epoch 150 that affects the visual quality of the images (see Table 7). When comparing the transformers, RoBERTa+CelebA yields better results in almost all measurements, obtaining the best result in epoch 550 and affecting a significant improvement in the visual quality of the images (see Table 7). Note that the lower the FID score, the better is the quality of the images generated.
- LPIPS (see Fig. 5(c)): It can be observed that Sent2vec+CelebA shows the worst result in most epochs throughout the training, showing a notable worsening from epoch 200 onwards. Meanwhile, RoBERTa+CelebA decreases faster and maintains a better score at all times than RoBERTa. When LPIPS is compared to the visual evaluation of the images (see Table 7), it is noted that the images have a direct and correlating relationship, coinciding with FID but in contrast to IS. Note that the lower the LPIPS score, the better is the quality of the images generated.

In summary, Table 6 shows the best results obtained in each of the metrics for the different encoders. It can be observed that RoBERTa+CelebA obtains the best results for all the metrics, i.e., IS in epoch 50, and FID and LPIPS, in epoch 550. Additionally, for

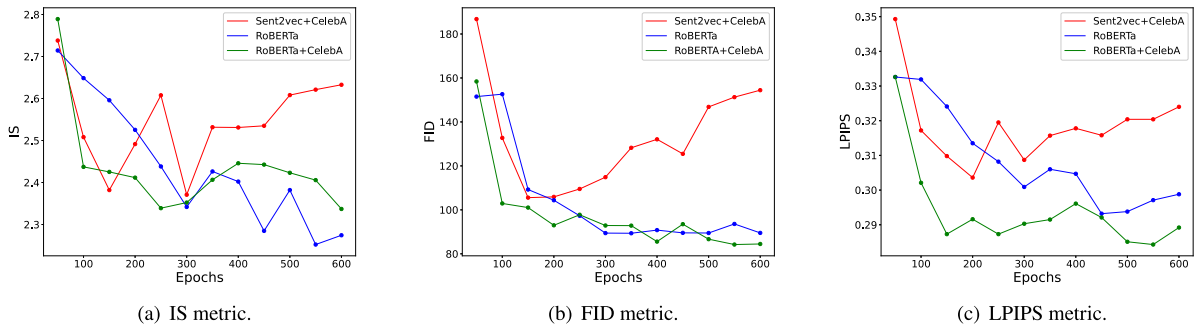


Fig. 5. Variation of the different metrics with respect to the number of training epochs for the three encoders under study: Sent2vec+CelebA, RoBERTa and RoBERTa+CelebA.

Table 6

Best values of quantitative metrics using the three encoders: Sent2vec+CelebA, RoBERTa and RoBERTa+CelebA. Best results are shown in bold.

Encoder	IS	FID	LPIPS
Sent2vec+CelebA	2.7380 ± 0.1979	105.5604	0.3036 ± 0.0758
RoBERTa	2.7143 ± 0.1769	89.3566	0.2932 ± 0.0776
RoBERTa+CelebA	2.7891 ± 0.1659	84.2212	0.2851 ± 0.0769

FID and LPIPS, these best results are achieved after a prolonged training time (during the last training periods) and in accordance with the visual results of the images generated (see Table 7). In contrast, the IS achieves the best scores during the first epochs (in epoch 50). Therefore, it can be stated that the effectiveness in assessing the image quality of the IS for CelebA faces is significantly lower than that of FID and LPIPS.

5.4. Visual analysis

Visual analysis allows us to complement the results of the metrics and compare whether they are directly related to the visual appearance of the generated images. For this purpose, seven descriptive sentences were randomly selected. Furthermore, the images generated for textual descriptions were evaluated using the best training weights obtained in FID.

Table 7 shows different images generated by cDCGAN. The descriptive text used for the evaluation is: “*She has large lips with arched eyebrows and a slightly open mouth. The young and attractive smiling woman is wearing a lot of makeup. She is wearing earrings and lipstick*”. (note that the sentences were translated into Spanish for training/testing purposes, although this document shows them in English for better understanding). The following can be observed:

- Transformer-based models, i.e., RoBERTa and RoBERTa+CelebA, yield a better visual rendering in the images generated in all epochs.
- The training time directly influences the quality of the images generated (and in accordance with the metrics), e.g., Sent2vec+CelebA gradually decreases the quality from epoch 200 onwards.
- In terms of image rendering, this is consistent with the features described in the text. RoBERTa+CelebA generates more detailed images according to the features described as input.
- The visual analysis of the images generated is directly related to FID and LPIPS and is inversely related to IS (compare with Fig. 5).

Furthermore, Table 8 shows the different outputs of faces generated from a textual description as input. The generator was boosted with the weights of the best FID result. The following can be observed:

- The resolution and quality of the images generated improve in direct proportion to the number of sentences in the description; see Items 1, 3, 5 and 6. On the contrary, in Items 2 and 4, which have only two descriptive sentences, the images have less detail. It can also be seen that Sent2vec+CelebA generates the most distortion when it has little information and RoBERTa+CelebA the least.
- In the cases where not all the given features are defined as an input, the generator assigns random features. For example, in Item 4, the face features are assumed by the cDCGAN based on the learning process in training.
- There is difficulty in generating images with descriptive features that were not considered during the training. Images have distortions or approximate features. For example, for Item 6, the feature “*She has short hair that is blonde in colour*” causes an image with distortion to be generated using Sent2vec+CelebA. However, RoBERTa and RoBERTa+CelebA generated higher-quality images.

Table 7

Best values across the epochs for IS, FID and LPIPS, using the three encoders under study: Sent2vec+CelebA, RoBERTa and RoBERTa+CelebA.

Model / Epochs	100	200	300	400	500	600
Sent2vec+CelebA						
RoBERTa						
RoBERTa+CelebA						

Table 8

Some examples of generated images based on a descriptive text as an input using the three encoders under study: Sent2vec+CelebA, RoBERTa and RoBERTa+CelebA. Note that the sentences were translated into Spanish for training/testing purposes, although this document shows them in English for better understanding.

Item	Description	Sent2vec+CelebA	RoBERTa	RoBERTa+CelebA
1	The woman has an oval face and high cheekbones. Her hair is blonde in colour. She has large lips with arched eyebrows and a slightly open mouth. The young and attractive smiling woman is wearing a lot of make-up. She is wearing earrings and lipstick.			
2	The man has high cheekbones. He has a 5 o'clock shadow. His hair is black. He has big lips and a big nose. He is a handsome young man and is smiling.			
3	The woman has high cheekbones. She has straight black hair. She has large lips with arched eyebrows and a slightly open mouth. The attractive, smiling young woman is wearing a lot of make-up. She is wearing lipstick.			
4	The man has straight black hair and a pointed nose. He has a pointed nose. The man looks young. He is wearing a tie.			
5	The woman's hair is black. She has large lips and a large, pointed nose with arched eyebrows. The attractive young woman is wearing a lot of make-up. She is wearing earrings, a necklace and lipstick.			
6	The woman has low cheekbones. She has short hair which is blonde in colour. She has arched eyebrows. The attractive young and smiling woman is wearing a lot of make-up. She is wearing lipstick.			

- The most complicated attributes to generate are the descriptions of accessories such as a tie, lipstick or hat. Despite this, it can be seen that RoBERTa+CelebA generates the most complete images of these accessories.

In summary, it can be confirmed that RoBERTa+CelebA generates images of higher quality than RoBERTa and Sent2vec+CelebA and according to reliable and robust metrics.

5.5. Discussion

Once the cDCGAN was implemented with the three encoders under study, the model was trained and the quality of the generated images was evaluated. We can make the following observations:

- The training time for our cDCGAN was quite costly in all cases, as the generator and discriminator seek a complex and fault-prone balance. Moreover, instead of using the baseline layers in the trained model, we used specialised convolutional layers in the images, which presents a higher cost in time and computational resources.
- As can be seen in Table 4, RoBERTa+CelebA generates better results than RoBERTa applying Spearman's correlation for a set of sentences. This improvement can be seen numerically in the metrics, i.e., IS, FID and LPIPS (see Fig. 5). Furthermore, our metrics improve on the results seen in the literature (see Table 1).
- Regarding IS, we obtain a significant score above 2 in all encoders. Similarly, it can be seen that there is an improvement in this metric with respect to the baseline work, obtaining an IS of 1.4 ± 0.7 (see Table 1) (Nasir et al., 2019).
- The IS score increases or decreases in the opposite proportion to the other two metrics and to the visual quality of the generated images. The main problem is that IS is limited by the dataset used during classifier training. Therefore, if not trained with similar images, it scores low and provides little ability for the classifier to detect visual features in defining the concept of image quality, i.e., poor quality images may obtain high scores.
- FID and LPIPS show reliable results that go hand in hand with the qualitative visual appearance of the images generated. This is an indicator of reliability, as they are much more accurate and suitable metrics for image evaluation than IS. Our findings show promising results with respect to the literature by improving the numerical metrics of FID by 5% and LPIPS by 37% (see Table 1).
- Table 8 shows that transformers generate much sharper images with characteristics similar to the descriptive text, even if the descriptive text is not included in the training. This is because Sent2vec+CelebA wholly depends on the corpus used for training. In contrast, RoBERTa uses a pre-trained model rich in contextual information that allows the lack of descriptive features to be overcome by replacing them with similar values included in their pre-training.
- The transformer model developed by the authors, i.e., RoBERTa+CelebA, generates better quantitative and qualitative results than RoBERTa and Sent2vec+CelebA.

6. Conclusions and open challenges

In the current research, a cDCGAN was implemented that generates faces guided by a descriptive text in Spanish. Three different encoders were tested: Sent2vec+CelebA, RoBERTa and the model developed in this research, RoBERTa+CelebA. In order to have a robust model, a descriptive corpus from CelebA was generated in Spanish, with which the Sent2vec encoder and the research model, i.e., RoBERTa+CelebA, were trained.

Likewise, to evaluate the generalisation of the developed model, i.e., RoBERTa+CelebA, a numerical evaluation was carried out using the IS, FID and LPIPS metrics, and a qualitative evaluation was performed by visually verifying the resulting synthetic face images. In both cases, it can be concluded that RoBERTa+CelebA generates both quantitatively (numerically) and qualitatively (visually) better results than Sent2vec and RoBERTa.

Furthermore, we show that IS is not an objective metric to be taken into account in this type of experiment because of its completely random performance in the quantitative–qualitative evaluation relationship. Consequently, we recommend using FID and LPIPS, which showed more reliable, stable and robust performance throughout the training, and in accordance with the generated images for the cDCGAN. Our findings show promising results with respect to the literature by improving the numerical metrics of FID and LPIPS by 5% and 37%, respectively.

About the open challenges in this research are to train the cDCGAN model with a larger number of images and their respective captions for a longer time, and to test with a larger training corpus for Sent2vec and RoBERTa encoders, e.g., *Spanish Unannotated Corpora* (Cañete et al., 2020).⁵ A further research line is to develop a quantitative and qualitative comparative study on the efficiency of RoBERTa with respect to other architectures previously developed in the literature. Likewise, it is important to evaluate the quality of images in greater detail, specifically focusing on the quality of facial features. Finally, we recommend expanding the generation of text-to-face in Spanish by testing other robust architectures, such as SBERT+DFGAN.

CRedit authorship contribution statement

Eduardo Yauri-Lozano: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Manuel Castillo-Cara:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Luis Orozco-Barbosa:** Investigation, Methodology, Project administration, Supervision, Validation, Writing – original draft, Writing – review & editing, Resources. **Raúl García-Castro:** Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

⁵ <https://doi.org/10.5281/zenodo.3247731>

Data availability

This scientific experiment contains the following extended data:

- The RoBERTa corpus (CelebA dataset) in Spanish (Yauri-Lozano, Castillo-Cara, Orozco-Barbosa, & García-Castro, 2023a).
- The Sent2vec corpus (CelebA dataset) in Spanish (Yauri-Lozano, Castillo-Cara, Orozco-Barbosa, & García-Castro, 2023b).
- The Sent2vec+CelebA encoder Model trained with CelebA dataset in Spanish (Yauri-Lozano, Castillo-Cara, Orozco-Barbosa, & García-Castro, 2023d).
- The RoBERTa+CelebA encoder model trained with CelebA dataset in Spanish (Yauri-Lozano, Castillo-Cara, Orozco-Barbosa, & García-Castro, 2023c).
- All Python code used in this work, i.e., Algorithm 1, Sent2vec+CelebA and RoBERTa+CelebA (Yauri-Lozano & Castillo-Cara, 2023).

Acknowledgements

The research leading to these results was co-funded by the CYTED (ref. 520rt0011) and the Spanish Ministry of Science, Education and Universities, the European Regional Development Fund and the State Research Agency, Grant No. PID2021-123627OB; also by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with the Universidad Politécnica de Madrid in the Excellence Programme for University Teaching Staff, in the context of the V PRICIT (Regional Programme of Research and Technological Innovation).

References

- Agnese, J., Herrera, J., Tao, H., & Zhu, X. (2020). A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *WIREs Data Mining and Knowledge Discovery*, 10(4), Article e1345. <http://dx.doi.org/10.1002/widm.1345>, URL: <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1345>.
- Barratt, S., & Sharma, R. (2018). A note on the inception score. <http://dx.doi.org/10.48550/ARXIV.1801.01973>, arXiv:1801.01973.
- Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.
- Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 670–680). Copenhagen, Denmark: Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/D17-1070>, URL: <https://aclanthology.org/D17-1070>.
- Deorukhkar, K., Kadamala, K., & Menezes, E. (2022). FGTD: Face Generation from Textual Description. In G. Ranganathan, X. Fernando, & F. Shi (Eds.), *Inventive communication and computational technologies* (pp. 547–562). Singapore: Springer Nature Singapore, http://dx.doi.org/10.1007/978-981-16-5529-6_43.
- Ding, H., Sun, Y., Wang, Z., Huang, N., Shen, Z., & Cui, X. (2023). RGAN-EL: A GAN and ensemble learning-based hybrid approach for imbalanced data classification. *Information Processing & Management*, 60(2), Article 103235. <http://dx.doi.org/10.1016/j.ipm.2022.103235>, URL: <https://www.sciencedirect.com/science/article/pii/S0306457322003363>.
- Fandiño, A. G., Estapé, J. A., Pàmies, M., Palao, J. L., Ocampo, J. S., Carrino, C. P., et al. (2022). MarIA: Spanish language models. *Procesamiento del Lenguaje Natural*, 68, <http://dx.doi.org/10.26342/2022-68-3>.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139–144. <http://dx.doi.org/10.1145/3422622>.
- Guan, M., Mondal, S. K., Dai, H.-N., & Bao, H. (2023). Reinforcement learning-driven deep question generation with rich semantics. *Information Processing & Management*, 60(2), Article 103232. <http://dx.doi.org/10.1016/j.ipm.2022.103232>, URL: <https://www.sciencedirect.com/science/article/pii/S0306457322003338>.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local NASH equilibrium. *Advances in Neural Information Processing Systems*, 30.
- Keras Official Documentation (2023). InceptionV3. URL: <https://keras.io/api/applications/inceptionv3/>. [Online: accessed 18-mar-2023].
- Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, 27.
- Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., et al. (2019). STORYGAN: A sequential conditional GAN for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6329–6338).
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. <http://dx.doi.org/10.48550/ARXIV.1907.11692>.
- Ma, D., Liu, B., Kang, Z., Zhou, J., Zhu, J., & Xu, Z. (2020). Two birds with one stone: Transforming and generating facial images with iterative GAN. *Neurocomputing*, 396, 278–290. <http://dx.doi.org/10.1016/j.neucom.2018.10.093>.
- Nasir, O. R., Jha, S. K., Grover, M. S., Yu, Y., Kumar, A., & Shah, R. R. (2019). Text2faceGAN: Face generation from fine grained textual descriptions. In *2019 IEEE fifth international conference on multimedia big data (BigMM)* (pp. 58–67). IEEE.
- Oza, M., Chanda, S., & Doermann, D. (2021). Semantic text-to-face GAN-ST \wedge 2FG. <http://dx.doi.org/10.48550/ARXIV.2107.10756>.
- Pagliardini, M., Gupta, P., & Jaggi, M. (2018). Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 528–540). New Orleans, Louisiana: Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/N18-1049>.
- Parmar, G., Zhang, R., & Zhu, J. (2022). On aliased resizing and surprising subtleties in GAN evaluation. In *2022 IEEE/CVF conference on computer vision and pattern recognition* (pp. 11400–11410). Los Alamitos, CA, USA: IEEE Computer Society, <http://dx.doi.org/10.1109/CVPR52688.2022.01112>, URL: <https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01112>.
- Qiao, T., Zhang, J., Xu, D., & Tao, D. (2019). MirrorGAN: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1505–1514).
- Qin, Z., Chen, Q., Ding, Y., Zhuang, T., Qin, Z., & Choo, K.-K. R. (2022). Segmentation mask and feature similarity loss guided GAN for object-oriented image-to-image translation. *Information Processing & Management*, 59(3), Article 102926. <http://dx.doi.org/10.1016/j.ipm.2022.102926>, URL: <https://www.sciencedirect.com/science/article/pii/S0306457322000504>.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982–3992). Hong Kong, China: Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/D19-1410>.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 29.

- SentenceTransformers Documentation (2023). Training overview. URL: <https://www.sbert.net/docs/training/overview.html>. [Online: accessed 18-mar-2023].
- Sommer, W. L., & Iosifidis, A. (2020). Text-to-image synthesis method evaluation based on visual patterns. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing ICASSP*, (pp. 4097–4101). IEEE.
- Sun, J., Deng, Q., Li, Q., Sun, M., Ren, M., & Sun, Z. (2022). AnyFace: Free-style text-to-face synthesis and manipulation. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 18666–18675). <http://dx.doi.org/10.1109/CVPR52688.2022.01813>.
- Talla-Chumpitaz, R., Castillo-Cara, M., Orozco-Barbosa, L., & García-Castro, R. (2023). A novel deep learning approach using blurring image techniques for bluetooth-based indoor localisation. *Information Fusion*, 91, 173–186. <http://dx.doi.org/10.1016/j.inffus.2022.10.011>.
- Tao, M., Tang, H., Wu, F., Jing, X., Bao, B.-K., & Xu, C. (2022). DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis. In *2022 IEEE/CVF conference on computer vision and pattern recognition CVPR*, (pp. 16494–16504). <http://dx.doi.org/10.1109/CVPR52688.2022.01602>.
- Torch Metrics (2023a). Frechet inception distance (FID). URL: https://torchmetrics.readthedocs.io/en/stable/image/frechet_inception_distance.html. [Online: accessed 18-mar-2023].
- Torch Metrics (2023b). Learned perceptual image patch similarity (LPIPS). URL: https://torchmetrics.readthedocs.io/en/stable/image/learned_perceptual_image_patch_similarity.html. [Online: accessed 18-mar-2023].
- Vasquez-Espinoza, L., Castillo-Cara, M., & Orozco-Barbosa, L. (2021). On the relevance of the metadata used in the semantic segmentation of indoor image spaces. *Expert Systems with Applications*, 184, Article 115486. <http://dx.doi.org/10.1016/j.eswa.2021.115486>.
- Xia, W., Yang, Y., Xue, J.-H., & Wu, B. (2021). TediGAN: Text-guided diverse face image generation and manipulation. In *2021 IEEE/CVF conference on computer vision and pattern recognition CVPR*, (pp. 2256–2265). <http://dx.doi.org/10.1109/CVPR46437.2021.00229>.
- Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., et al. (2018). AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE/CVF conference on computer vision and pattern recognition CVPR*, (pp. 1316–1324). Los Alamitos, CA, USA: IEEE Computer Society, <http://dx.doi.org/10.1109/CVPR.2018.00143>, URL: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00143>.
- Yauri-Lozano, E., & Castillo-Cara, M. (2023). Generative adversarial networks for text-to-face synthesis & generation: A quantitative-qualitative analysis of natural language processing encoders for spanish. <http://dx.doi.org/10.5281/zenodo.7791805>, URL: <https://github.com/eduar03yauri/DCGAN-text2face-forSpanish>.
- Yauri-Lozano, E., Castillo-Cara, M., Orozco-Barbosa, L., & García-Castro, R. (2023a). CelebA_RoBERTa_Sp (revision b642d1c). <http://dx.doi.org/10.57967/hf/0447>, URL: https://huggingface.co/datasets/oeg/CelebA_RoBERTa_Sp.
- Yauri-Lozano, E., Castillo-Cara, M., Orozco-Barbosa, L., & García-Castro, R. (2023b). CelebA_Sent2Vect_Sp (Revision 171f21f) . <http://dx.doi.org/10.57967/hf/0446>, URL: https://huggingface.co/datasets/oeg/CelebA_Sent2Vect_Sp.
- Yauri-Lozano, E., Castillo-Cara, M., Orozco-Barbosa, L., & García-Castro, R. (2023c). RoBERTa-CelebA-Sp (Revision 745a00c) . <http://dx.doi.org/10.57967/hf/0464>, URL: <https://huggingface.co/oeg/RoBERTa-CelebA-Sp>.
- Yauri-Lozano, E., Castillo-Cara, M., Orozco-Barbosa, L., & García-Castro, R. (2023d). Sent2vec_CelebA_Sp (Revision cd82a57). <http://dx.doi.org/10.57967/hf/0465>, URL: https://huggingface.co/oeg/Sent2vec_CelebA_Sp.
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition CVPR*, (pp. 586–595). Los Alamitos, CA, USA: IEEE Computer Society, <http://dx.doi.org/10.1109/CVPR.2018.00068>, URL: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00068>.
- Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., et al. (2017). StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *2017 IEEE international conference on computer vision ICCV*, (pp. 5908–5916). <http://dx.doi.org/10.1109/ICCV.2017.629>.
- Zhao, D., Wang, J., Lin, H., Chu, Y., Wang, Y., Zhang, Y., et al. (2021). Sentence representation with manifold learning for biomedical texts. *Knowledge-Based Systems*, 218, Article 106869. <http://dx.doi.org/10.1016/j.knsys.2021.106869>, URL: <https://www.sciencedirect.com/science/article/pii/S0950705121001325>.
- Zhao, J., Xie, X., Wang, L., Cao, M., & Zhang, M. (2019). Generating photographic faces from the sketch guided by attribute using GAN. *IEEE Access*, 7, 23844–23851.