# $m^3GAN$: Multi-Modal Matching with Generative Adversarial Networks

**Ntavelis, Evangelos**
Master of Artificial Intelligence
KULeuven
evangelos.ntavelis@student.kuleuven.be

**Collell, Guillem**
Computer Science Department
KULeuven
gcollell@kuleuven.be

**Moens, Marie-Francine**
Computer Science Department
KULeuven
sien.moens@cs.kuleuven.be

*Abstract*– **Humans' understanding of the world is grounded on all of our senses. Accordingly, being able to create a representation that incorporates information of more than one modality has many applications in a great variety of Natural Language Processing and Computer Vision Tasks. Recently, Generative Adversarial Networks have been proposed to tackle the problem of cross-modal retrieval. Yet, these works are limited to the domain of representations. We propose the $m^3GAN$ architecture as a way to create representations by simultaneously training a network to caption images and generate images given a caption. To train our network we use two pathways from one modality to the other. Each modality's input is encoded in a representation vector which will be then decoded to an output of the other modality, and two discriminator networks will test the quality of the produced results. At the same time, we use the encoder networks previously used on the input data to create a vector representation of the produced outputs. This way we can also measure both an inter- and an intra-modal reconstruction loss at the representation level. The resulted representations are then tested on a cross-modal retrieval task.**

## 1 Introduction

There is evidence that the nature of human memory is reconstructive [1, 2]. That means that our neural system breaks down the stimuli it receives from the outside world into an internal representation and it recreates it when the need arises. Moreover, Barsalou [3] argues that ground cognition draws upon the senses. It's normal, thus, to aim to computationally recreate this multi-modal decomposition, coding and composition procedure.

In the previous years we can observe a lot of approaches tackling this problem using auto-encoders, both combining audio and visual features [4] and visual and text features [5]. The encoders were fed with pre-learned features independently for both modalities, which were then combined *a posteriori*.

More recently, a generative family of models, Generative Adversarial Networks, have gained popularity. Among other things, they were used as an alternative method to design multi-modal embeddings in video hyperlinking with promising results [6].

Here, we propose an architecture that tries to match representations created from one modality with the corresponding ones of an other. The main idea is that given one modality we try to create the other. So, given an image we are trying to create the text, or *caption*, that describes its contents, while at the same time we are trying to generate an image given a text description. The discriminator networks will ensure that the produced results are realistic and the imposed representation reconstruction losses will establish that our representations our networks produce are able to meaningfully incorporate information.

The paper is organized in the following way. The next session describes a plethora of papers related to the task at hand. Related work and background are discussed in the second section. We later analyze the notation and architecture of our model. Then, we report on the experimental results before concluding

with the last section.

## 2 Related Work and Background

Peng et al 2017 [7] and Wang et al [8] seem to follow a similar approach to utilize an adversarial setting [9] for *Cross-Modal Representation Learning*. In both articles we observe the usage of initial representations derived from a unimodal setting:

- Visual Embedding: 4,096 dimension feature vector from the fc7 layer of a VGG-Net trained on ImageNet for both cases
- Word Embeddings:
  - $n \times x$ matrix, with n the number of words and k the number of Word2Vec Features [10]
  - Bag of Words vector with TF-IDF weighting

They both use fully connected layers to arrive to a common representation and use discriminators intermodally and intra-modally. Yet the approaches differ in their implementations and also the acmr method uses a label predictor as an additional discriminator network.

Since the introduction of Generative Adversarial Networks in 2014 by Ian Goodfellow [9] the architecture has produced a lot of promising results in a variety of tasks. Especially after the introduction of DCGANs we observed significant results in unsupervised representation learning [11]. The performance of GANs was later improved by the stabilization of their training [12, 13].

These networks have also been used in tasks than combine language and vision. Reed et. al [14] introduced an architecture to synthesize images given a text description. Using stacked layers of GANs and, additionally in their last work, incorporating attention Tao Xu, Han Zhang et al. [15–17] created realistic images. Moreover, on the image captioning task, following the *Show, Attend and Tell* model [18] Chen et al. created the *Show, Adapt and Tell* model for cross-domain image captioning [19].

GANs have produced interesting results both in image translation [20, 21] and language generation tasks [22].

By applying an approach similar to Disco-GANs [20] on a multi-modal setting we will train a network for a cross-modal retrieval task. Specifically, we measure the cross-modal retrieval performance in terms of mAP on the tested dataset (e.g. MSCOCO [23]).

## 3 Proposed Method

Our method, as shown in Figure 1 is composed by two pathways. We are using the dataset MSCOCO and for each image in our training set we have five captions. Our goal is to simultaneously train the network to produce descriptions of images given the images and generate images given a description. This is achieved by utilizing six sub networks in total:

- **Image Encoder:** A *Convolutional Neural Network* is used to map the image to the common representation space
- **Text Decoder:** The text generator of our first GAN is a *Long-Short Term Memory Network* that transforms the representation vector to a word embeddings sequence
- **Text Encoder:** A *bidirectional Long-Short Term Memory Network* that projects a sentence(caption) to the representation domain
- **Image Decoder** The image generator of our second GAN is *Deconvolutional Neural Network* that has as input the representation vector concatenated with a noise vector and outputs an image
- **Image Discriminator:** A *Convolutional Neural Network* that tries to tell generated images apart from the ones present on the training data
- **Text Discriminator:** A *Long-Short Term Memory Network* that determines the probability that its input's lingual sequence is fabricated

### 3.1 Data Preparation

All the images in the MSCOCO dataset are resized to dimensions of 256x256. All words of frequency less than 4 are omitted from the vocabulary, the rest are tokenized. Based on the resulting vocabulary we turn the captions into sequences of GloVe embeddings [24] with dimension of 100, before feeding them to the network.

### 3.2 Networks Pre-training

Before training the architecture on the cross-modal setting, we train the encoder/decoder networks in a auto-encoder fashion, as show in Figure 2. All our schemes share the same color coding. The light orange background denotes that an *encoder*(or *decoder*) has *input*(or *output*) of visual data. The light blue background is used for text inputs and outputs.

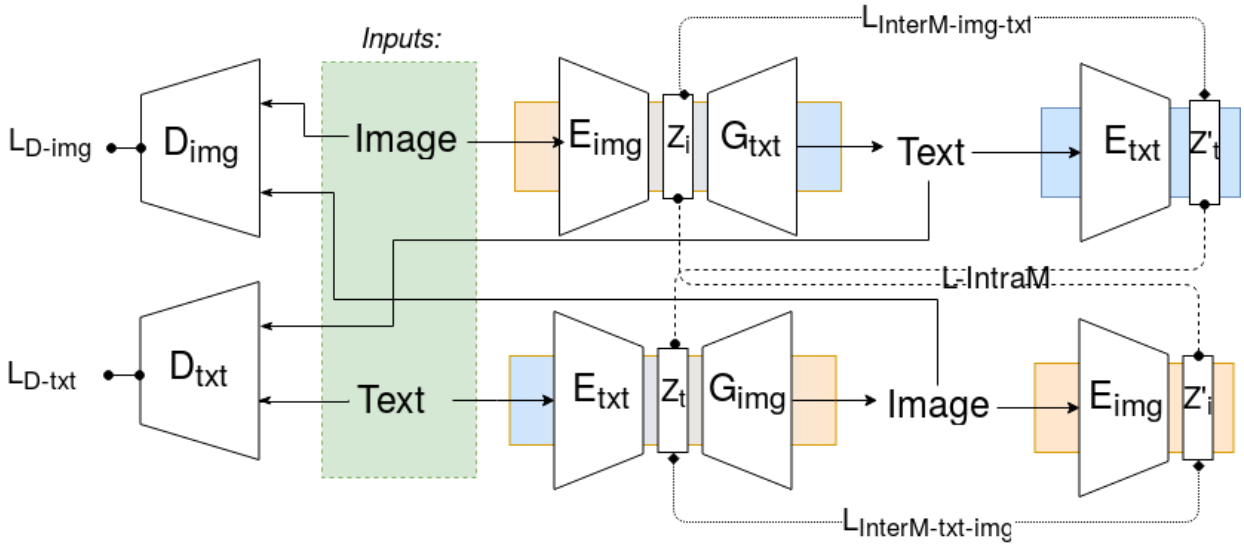


Figure 1: $m^3Gan$ Architecture.

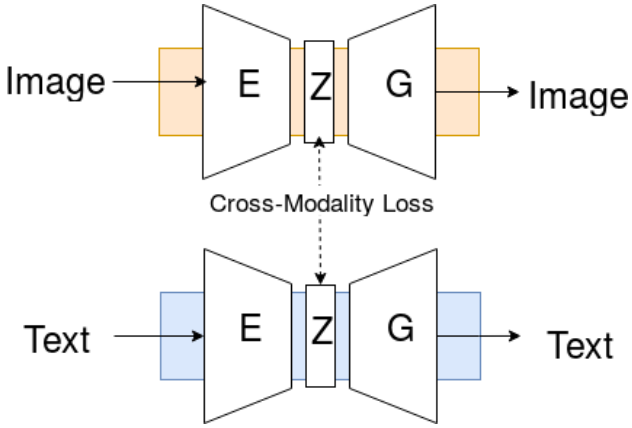


Figure 2: Pre-training of our networks.

The encoder networks map from a unimodal to the multi-modal domain and the decoders vice-versa. Their task has the same nature both in the case they are used as parts of an auto-encoder and as parts of a inter-modal translation network. Yet translating back to the same modality is an easier task with well defined objective functions. For the case of image translation we can use various distance functions such as MSE, cosine distance, and hinge-loss [20]. For the textual reconstruction the cross-entropy between the input and its reconstruction [25].

The two auto-encoders are paired during the training. For each image iteration on the visual network we train on its five captions on the language model. During the process we impose a cross-modal loss between the intermediate representations, denoted as $Z$ in Figure 2. This auxiliary loss can also be MSE, cosine distance or *the squared difference of the dot product of the two representation vectors and one*. An additional parameter $\alpha$ can be used to determine the percentage of the overlapping we want to take into account between the two vectors.[1]

The imposed loss function makes the two vectors represent features that are activated by the pictures and their corresponding captions to share similar positions in the vector. This will prove useful when we will swap the decoder networks of the two encoder-decoder pairs as the background colors denote on figures 2 and 1.

At this point we already have representations whose performance we can gauge on a cross-modal retrieval task.

### 3.3 Our Model

Hereby, we describe the notation and our architecture. When $\boldsymbol{i}$ is used, the referenced variable is in the image domain, in case of $\boldsymbol{t}$ in the lingual domain and we use $\boldsymbol{z}$ for a vector in the representation space.

The *Image Encoder* $\boldsymbol{E_{img}}$ takes an image of dimensions 256x256x3 and transforms it to a vector in the common multi-modal space of dimension 300. This vector is then used as input to the *Text Decoder* (Text Generator) $\boldsymbol{G_{txt}}$ who transforms it on a sequence of

[1] The idea is that the representation hyperplanes of the two modalities are not purely identical, so we only consider that they share a subset of their dimensions. The remaining dimensions are descriptive of features unique to the particular modality.

GloVe embeddings.

$$z_i = \boldsymbol{E_{img}}(i)$$

$$t' = \boldsymbol{G_{txt}} \circ \boldsymbol{E_{img}}(i) = \boldsymbol{G_{txt}}(z_i)$$

In parallel, the *Text Encoder*(Text Generator) $\boldsymbol{E_{img}}$ takes the caption transformed to a sequence of GloVe embeddings as an input. The output is also a vector of 300 elements which is then fed to the *Image Decoder*(Image Generator) $\boldsymbol{G_{txt}}$ concatenated with a vector of noise.

$$z_t = \boldsymbol{E_{txt}}(t)$$

$$i' = \boldsymbol{G_{img}} \circ \boldsymbol{E_{txt}}(t) = \boldsymbol{G_{img}}(z_t)$$

Each of the Generators produces an output that is then fed to the two discriminator networks, the *Image Discriminator* $\boldsymbol{D_{img}}$ and the *Text Discriminator* $\boldsymbol{D_{txt}}$. Using the results of those two networks we can compute our Generative Adversarial Losses and the GAN discriminator losses:

$$\boldsymbol{L_{GANimg}} = -\mathbb{E}_{t\sim P_{txt}}[\boldsymbol{logD_{img}}(\boldsymbol{G_{img}} \circ \boldsymbol{E_{txt}}(t))]$$

$$\boldsymbol{L_{GANtxt}} = -\mathbb{E}_{i\sim P_{img}}[\boldsymbol{logD_{txt}}(\boldsymbol{G_{txt}} \circ \boldsymbol{E_{img}}(i))]$$

$$\boldsymbol{L_{Dimg}} = -\mathbb{E}_{i\sim P_{img}}[\boldsymbol{logD_{txt}}(i)]$$
$$-\mathbb{E}_{t\sim P_{txt}}[\boldsymbol{logD_{img}}(\boldsymbol{1} - \boldsymbol{G_{img}} \circ \boldsymbol{E_{txt}}(t)]$$

$$\boldsymbol{L_{Dtxt}} = -\mathbb{E}_{t\sim P_{txt}}[\boldsymbol{logD_{img}}(t)]$$
$$-\mathbb{E}_{i\sim P_{img}}[\boldsymbol{logD_{txt}}(\boldsymbol{1} - \boldsymbol{G_{txt}} \circ \boldsymbol{E_{img}}(i)]$$

The encoders $\boldsymbol{E_{img}}$ and $\boldsymbol{E_{txt}}$ are utilized again to produce a representation vector given the outputs of the generators. Thus, the inter-modal reconstruction losses are generated.

$$z'_t = \boldsymbol{E_{txt}}(t')$$
$$z'_i = \boldsymbol{E_{img}}(i')$$

$$\boldsymbol{L_{interMtxt-img}} = \boldsymbol{d}(z'_t, z_i)$$
$$\boldsymbol{L_{interMimg-txt}} = \boldsymbol{d}(z'_i, z_t)$$

Our last objective functions measures the intra-modal losses between the representation of an image and the images created from its captions, and the representation of the original captions and the ones created for the corresponding image.

$$\boldsymbol{L_{IntraMtxt}} = \boldsymbol{d}(z'_t, z_t)$$
$$\boldsymbol{L_{IntraMimg}} = \boldsymbol{d}(z'_i, z_i)$$

The summation of the two discriminator losses compose the total discriminator loss:

$$\boldsymbol{L_D} = \boldsymbol{L_{Dimg}} + \boldsymbol{L_{Dtxt}}$$

The loss for the Encoder-Decoder pairs is the following:

$$\boldsymbol{L_{E-G}} = \boldsymbol{L_{GANimg}} + \boldsymbol{L_{interMtxt-img}} + \boldsymbol{L_{IntraMtxt}}$$
$$+\boldsymbol{L_{GANtxt}} + \boldsymbol{L_{interMimg-txt}} + \boldsymbol{L_{IntraMimg}}$$

# 4 Experimental Results

# 5 Conclusion

# References


[1] D. Vernon, *Artificial Cognitive Systems: A primer*. MIT Press, 2014.

[2] J. Hawkins and S. Blakeslee, *On Intelligence*. Macmillan, 2007.

[3] L. W. Barsalou, "Grounded cognition," *Annual Review of Psychology*, vol. 59, no. 1, pp. 617–645, 2008. PMID: 17705682.

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning.," in *ICML* (L. Getoor and T. Scheffer, eds.), pp. 689–696, Omnipress, 2011.

[5] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Baltimore, Maryland), pp. 721–732, Association for Computational Linguistics, June 2014.

[6] V. Vukotic, C. Raymond, and G. Gravier, "Generative adversarial networks for multimodal representation learning in video hyperlinking," *CoRR*, vol. abs/1705.05103, 2017.

[7] Y. Peng, J. Qi, and Y. Yuan, "Cm-gans: Cross-modal generative adversarial networks for common representation learning," *CoRR*, vol. abs/1710.05106, 2017.

[8] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. Tao Shen, "Adversarial cross-modal retrieval," *MM '17*, pp. 154–162, 10 2017.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *ArXiv e-prints*, June 2014.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *CoRR*, vol. abs/1310.4546, 2013.

[11] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *CoRR*, vol. abs/1511.06434, 2015.

[12] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, "Stabilizing training of generative adversarial networks through regularization," *CoRR*, vol. abs/1705.09367, 2017.

[13] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *arXiv*, Jan 2017.

[14] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text-to-image synthesis," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016.

[15] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017.

[16] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan++: Realistic image synthesis with stacked generative adversarial networks," *CoRR*, vol. abs/1710.10916, 2017.

[17] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," *CoRR*, vol. abs/1711.10485, 2017.

[18] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *CoRR*, vol. abs/1502.03044, 2015.

[19] T.-H. Chen, Y.-H. Liao, C.-Y. Chuang, W.-T. Hsu, J. Fu, and M. Sun, "Show, adapt and tell: Adversarial training of cross-domain image captioner," *arXiv preprint arXiv:1705.00930*, 2017.

[20] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," *CoRR*, vol. abs/1703.05192, 2017.

[21] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *CoRR*, vol. abs/1703.00848, 2017.

[22] S. Rajeswar, S. Subramanian, F. Dutil, C. J. Pal, and A. C. Courville, "Adversarial generation of natural language," *CoRR*, vol. abs/1705.10929, 2017.

[23] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[24] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[25] B. Oshri and N. Khandwala, "There and back again: Autoencoders for textual reconstruction," 2016.