Integrating vision and language using generative models: a research proposal

Evangelos Ntavelis
r0692337

December 17, 2017

1 Introduction

There is evidence that the nature of human memory is reconstructive (Vernon 2014 [1]; Hawkins and Blakeslee 2007 [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 (Ngiam et al, 2011 [4]) and visual and text features (Silberer and Lapata 2014 [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 (Vukotic et al 2017).

2 Goals

In the course of this thesis project we aim to test how Generative Adversarial Networks Architectures perform in tasks of lingo-visual representations. Specifically, we want to:

- Gauge the ability of these architectures to map informatively:
  - language to vision and back, (extending Collell et al 2017 [6])
  - both language and vision to a common representation space
- Test jointly trained features versus pre-learned features when training on both modalities.

3 Experimental Architectures

Following the example set by Peng et al 2017 [7] and by Wang et al [8], who seem to follow a similar approach to utilize an adversarial setting [9] for *Cross-Modal Representation Learning*, we aim to experiment with different configurations and building blocks.

In both of the aforementioned 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 inter-modally and intra-modally. Yet the approaches differ in their implementations and also the acmr method uses a label predictor as an additional discriminator network.

I ~~plan to experiment with architectures proposed for image to image transformations and apply them in a Cross-Modal setting. Namely Disco-GANs [11] and UNIT [12] . Initially, the idea would be to imitate the abstract composition of the proposed architectures.~~

~~I~~ want to try two approaches. Firstly, following the above adversarial cross-modal paradigms, I want to try using the embeddings of pre-trained networks as input to these architectures. Alternatively, a more ambitious approach would be to try to train the networks while feeding them with raw image and sentences-describing-the-image data, for instance by using the MSCOCO dataset [12], and acquire the embeddings as a byproduct of the procedure.

In the first approach, the task we would train upon is information retrieval, both Bi-modal and All-modal. Specifically, we want to measure the cross-modal retrieval performance in terms of mAP on the tested datasets (e.g. Wikipedia, Pascal, MSCOCO etc.). In the later, the goal is to train networks simultaneously for $Image \rightarrow Text$ generation [13] and $Text \rightarrow Image$ [14] in an adversarial setting.

I plan to experiment with architectures proposed for image to image transformations and apply them in a Cross-Modal setting. Namely Disco-GANs [11] and UNIT [12] . Initially, the idea would be to imitate the abstract composition of the proposed architectures. For instance, UNIT proposes a combination of Variational Auto-Encoders with GANs, and it would be interesting to test if the usage of VAE instead of Convolutional Auto-Encoders (as proposed by Peng et al [7]) would prove beneficial to our task. The DISCO-GANs paper proposes a coupling of transformations procedures that use generators that take as input an image from one domain and transform it to an other. Drawing upon this idea, we can use a similar architecture with $image \rightarrow text$ and $text \rightarrow image$ generators as building blocks.

4 Challenges

- GANs are computationally demanding to train
- Will need access to GPUs (I don't know if that's a given)

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] G. Collell Talleda, T. Zhang, and M.-F. Moens, "Imagined visual representations as multimodal embeddings," *The Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pp. 4378–4384, 02 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] 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.

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

[13] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, pp. 664–676, Apr. 2017.

[14] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *CoRR*, vol. abs/1605.05396, 2016.