Contrats doctoraux
en Intelligence Artificielle

Projet AHEAD
Artificial Intelligence for Health, Physics, Transportation and Defense

Control of generative models for visual scenes

1 Context

The significant achievements obtained in the recent years by deep learning methods were made possible by a convergence of theoretical advances [LeCun et al., 2015], the availability of massively annotated datasets and growing tensor computation capabilities. These approaches have shown impressive results for several classical visual recognition tasks, natural language and speech processing. Furthermore, they enabled considerable advances in higher level tasks like playing Go [Silver et al., 2016] and supported more novel tasks like answering visual questions [Antol et al., 2015]. For these achievements to grow up to a complete renewal of AI, a few more important advances are required, including more control over models obtained from incomplete or biased data and a better understanding of the internal mechanisms and their failures.

One of the most exciting achievements, often employed to showcase the capabilities of deep learning, is the ability of recent generative models to produce high-resolution photo-realistic images [Karras et al., 2018, Brock et al., 2018, Razavi et al., 2019]. Such methods have many potential applications in content synthesis, dataset completion, image restoration, "deep fakes", etc. While generative methods have a long history in pattern recognition, the recent achievements are a result of novel architectures and learning mechanisms, including Variational Auto-Encoders [Kingma and Welling, 2019], Generative Adversarial Networks [Gui et al., 2020] and Normalizing Flows [Papamakarios et al., 2019, Kobyzev et al., 2019]. Work in the area of generative modelling mainly focused on extending the applicability of the approach to other types of content and on improving the quality of the generated data. However, for most applications, the user should have *significant* control over the generated content. This issue was addressed quite early in the literature and interesting "vector arithmetic" properties were identified for the generators (see e.g. [Radford et al., 2016]), allowing the user to manipulate to some extent a few dataset-specific properties (like adding glasses to a face). The *conditional* generation attempts to provide a rather general solution to this problem by allowing the user to give an input to the generation process, e.g. a semantic sketch for ordinary scenes [Park et al., 2019] or specific climatic properties for satellite images [Requena-Mesa et al., 2018]. Other recent works [Plumerault et al., 2020, Jahanian et al., 2020] propose to find meaningful directions in the latent space of generative models along which one can move to control precisely specific continuous properties of the generated image like the position or scale of the object in the image. While these proposals do provide some control over the generation process, they all fall short in giving the user sufficient direct control over a broad set of properties of the generated images.

2 Subject description

The main goal of this thesis is to define means for a refined control over the generated images, comprising several relatively independent "knobs", each controlling either a continuous, a discrete or a structured variable describing the image.

Depending on the target application (and thus on the nature of the generated images) control may concern:

- the presence and nature of individual entities in the scene (e.g. cyclist, truck),

- their pairwise positional relations (e.g. at the left of, far from),
- visual properties of individual entities (e.g. wearing glasses, wearing helmet),
- visual properties of the scene (e.g. sunny, dark),
- geometrical properties of the camera or the entities (e.g. camera pose, object orientation),
- physical parameters (such as weather conditions or climate) having an impact on the content of the scene or on its visual properties.

To advance towards the ambitious goal mentioned above, one or several of the following directions will be explored:

1. Finding meaningful directions in the latent space of generative models such that moving along one direction allows to control precisely one property of the generated image. This general approach was explored in early work on VAEs or GANs (e.g. [Radford et al., 2016]) with few specific, dataset-dependent properties. It was also explored recently for GANs ([Plumerault et al., 2020, Jahanian et al., 2020]) for specific continuous properties, by attempting to reverse the transformation generating an image. We believe that this approach deserves further study but intend to consider instead Normalizing Flows [Papamakarios et al., 2019] that allow to more explicitly reverse the generation process. Furthermore, the specific control requirements should be also considered during training.

2. In conditional generation [Mirza and Osindero, 2014] the generator has two inputs: one is the latent space where a simple distribution (e.g. multidimensional normal) is sampled and the other is a conditioning vector that allows to provide the control, like a semantic sketch [Park et al., 2019], specific climatic properties [Requena-Mesa et al., 2018], etc. The easy way to obtain joint control information for several variables (including structured ones) is to embed the corresponding data in a fixed-size vector and effective embedding methods are readily available for various types of data (text, semantic graphs, drawings, etc.). While general, this approach does not readily provide disentangled controls, especially when complex data is involved. It is nevertheless possible to pursue in this direction by exploiting the presence of the two inputs: the embedding can be employed for complex variables and optimized in order to reduce entanglement, while the latent space should allow to provide control over complementary variables.

3. It is important to see that one major cause for the difficulties to control generative models is the fact that their training data may well illustrate the application but its distribution is often *not representative* of what one intends to generate. Several variables describing the data are strongly correlated or some values are very under-represented, either because these correlations occur in nature or because data acquisition is (often inherently) incomplete. Finding ways to reduce the mutual dependence among specific variables in the *internal* representations of the network should contribute to an improved control over the generation process.

Solutions that combine several of these approaches may prove to be very effective in practice.

3 Practical interest

Controlled image synthesis with generative models is a rather upstream research field and has the potential to address many problems in computer vision. Applications range from data augmentation, domain adaptation and weakly supervised learning to supporting the explainability of deep learning systems.

Indeed, controlled generation has the ability to produce images to be further used as training samples for supervised algorithms, significantly improving existing approaches like [Bowles et al., 2018, Lee and Seok, 2019]. For example, user-controlled generative models could synthesize images that are difficult to capture in practice – and therefore scarcely represented in datasets – but of high interest for real-world deployment, such as accident-causing situations for autonomous vehicles (nighttime, storm, etc.). Or images that allow to explore (with far more control than typical adversarial examples) cases where the image processing system fails in order to help explaining the failures.

By providing a better understanding of latent representations, the thesis can also contribute to making deep architectures more interpretable. Also, the control of generative models also makes their output easier to understand for human operators as the black-boxes now have "knobs" that steer their outcome. It is a major challenge to make AI systems more transparent and thus improve the *digital trust* a user in these systems.

4 PhD outline

In France, the duration of a PhD program in Computer Science is three years and a natural outline is:

- First 3–4 months: bibliographic work in order to better understand the context of the work, identify opportunities and refine the research directions. This period is also used for "small experiments" with the aim to better handle the tools of CV/ML and reproduce recent contributions to the state-of-the-art.
- The rest of the first year is devoted to a deeper exploration of the research ideas to come up with a first novel contribution that can be published in an international conference. Attending a summer school should also be considered.
- Second year: dedicated to the intense development of research, to be validated by publications in the best conferences of the field (CVPR, ICCV, ECCV, NeurIPS, ICML, ICLR, IJCAI, ECAI, etc.).
- Third year: the final year of the PhD is divided in two. The first 6 months allow to finish the research work, including the most ambitious experiments combining several ideas developed in the previous years, and publish in the best conferences or journals. In the last 6 months the focus is on writing the PhD manuscript (with an outline put forward about 6 months before the end) and on preparing the professional integration (by more actively developing the network, starting job search, etc.).

5 PhD supervision and environment

During the three years of the PhD thesis, the candidate will be both enrolled as a student (SMI doctoral school) and employed by the Conservatoire National des Arts et Métiers (Cnam) in Paris for a fixed-term contract. Starting date is expected to be fall 2020 with a PhD defense and graduation in fall 2023.

This PhD thesis will be supervised by Michel Crucianu[1] and Nicolas Audebert[2] from **CEDRIC**-Cnam and by Hervé Le Borgne[3] from **CEA LIST**.

The Cnam is a higher education institution located in the center of Paris. It comprises research laboratories, multiple teaching faculties (computer science, mathematics, physics, economy, social sciences, etc.) and a museum of technological innovation. The **CEDRIC**[4] is the computer science laboratory of Cnam. It is composed of 8 research teams investigating a broad range of topics such as radio communications, cloud computing, data mining and image processing. The CEDRIC currently has 166 members among which 75 are permanent teacher-researchers and the others are PhD students and post-doctoral fellows. The PhD will take place in the Vertigo team[5] that focuses on complex data, machine learning and representations with a special interest in visual and audio content. Vertigo team consists of 6 permanent researchers and 4 PhD students.

Based in Saclay (Paris region, France), the LIST Institute (**CEA LIST**[6]) is one of CEA Tech three technological research institutes. Dedicated to smart digital systems, its mission is to achieve technological development of excellence for the industrial partners and create value. LIST institute has more than 750 partners and every year more than a 200 partnership activities are being conducted with French and foreign industrial companies on applied research projects in four main domains: Advanced Manufacturing, Embedded systems, Data intelligence, Health and ionizing radiations. Within the institute, the LASTI (laboratory of semantic analysis of texts and images) gathers about 25 persons, from PhD students to permanent researchers, working in the domain of computer vision and natural language analysis.

6 Candidate profile and application

The ideal candidate should have the following qualifications:

- MSc. degree in a relevant field (machine learning, computer vision, applied mathematics...) received in the last 2 years.

[1] http://cedric.cnam.fr/~crucianm/
[2] https://nicolas.audebert.at
[3] https://sites.google.com/view/herveleborgne/home
[4] http://cedric.cnam.fr/
[5] http://cedric.cnam.fr/lab/teams/vertigo-en/
[6] http://www-list.cea.fr/en/

- Previous experience in machine learning and particularly in deep learning.
- Good programming skills in Python are mandatory. Proficiency in a deep learning framework (e.g. PyTorch, Keras, Tensorflow...) is required.
- Good spoken and written communication skills in English. Speaking French is optional but may help for the everyday life during the PhD.
- Knowledge of modern computer vision, natural language processing, signal processing or physics is a plus.
- Being co-author of a submitted or published paper in the field is a plus.

We encourage candidates that do not entirely fit the profile but are highly motivated by the PhD proposal to apply nonetheless. Applications are to be sent to michel.crucianu@cnam.fr, nicolas.audebert@cnam.fr and herve.le-borgne@cea.fr, including a resume and a cover letter, preferably in (quite compressed) PDF.

References

[Antol et al., 2015] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

[Bowles et al., 2018] Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D. A., Valdés Hernández, M., Wardlaw, J., and Rueckert, D. (2018). GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks. *arXiv e-prints*, page arXiv:1810.10863.

[Brock et al., 2018] Brock, A., Donahue, J., and Simonyan, K. (2018). Large scale GAN training for high fidelity natural image synthesis. *CoRR*, abs/1809.11096.

[Gui et al., 2020] Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. (2020). A review on generative adversarial networks: Algorithms, theory, and applications.

[Jahanian et al., 2020] Jahanian, A., Chai, L., and Isola, P. (2020). On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*.

[Karras et al., 2018] Karras, T., Laine, S., and Aila, T. (2018). A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv e-prints*, page arXiv:1812.04948.

[Kingma and Welling, 2019] Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

[Kobyzev et al., 2019] Kobyzev, I., Prince, S. J. D., and Brubaker, M. A. (2019). Normalizing flows: An introduction and review of current methods.

[LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nature*, 521(7553):436–444.

[Lee and Seok, 2019] Lee, M. and Seok, J. (2019). Controllable Generative Adversarial Network. *IEEE Access*, 7:28158–28169.

[Mirza and Osindero, 2014] Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets.

[Papamakarios et al., 2019] Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2019). Normalizing flows for probabilistic modeling and inference.

[Park et al., 2019] Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Plumerault et al., 2020] Plumerault, A., Borgne, H. L., and Hudelot, C. (2020). Controlling generative models with continuous factors of variations. In *International Conference on Learning Representations*.

[Radford et al., 2016] Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations*.

[Razavi et al., 2019] Razavi, A., van den Oord, A., and Vinyals, O. (2019). Generating Diverse High-Fidelity Images with VQ-VAE-2. *arXiv e-prints*, page arXiv:1906.00446.

[Requena-Mesa et al., 2018] Requena-Mesa, C., Reichstein, M., Mahecha, M., Kraft, B., and Denzler, J. (2018). Predicting Landscapes as Seen from Space from Environmental Conditions. In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 1768–1771.

[Silver et al., 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T. P., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489.