More from Less: Learning with Limited Annotated Data in Vision and Language

Despite the impressive results of deep learning models, modern large-scale systems are required to be trained using massive amounts of manually annotated or freely available data on the Internet. But this "data in the wild" (i.e., data scraped from the Internet) is insufficient to learn specific structural patterns of the world, and existing large-scale models still fail on common sense tasks requiring compositional inference. Today systems are trained with this online content, which is bounded by the interaction of only 64.6% [1] of the world's population that has access to the Internet. This establishes a clear limit to the diversity of the available data, impacting specialized information and underrepresented cultures. On top of that, now we have automated systems that have been trained with "data in the wild" and yet are deployed to add textual and visual information to the Internet. This poses a major threat to the reliability of intelligent agents, which may become more biased with arbitrary information systematically distributed. *My research aims to explore ways to train novel techniques to leverage large amounts of information from real-world and synthetically generated data, and create intelligent systems that can benefit from highly diverse images and dense textual descriptions.*

I do research at the intersection of Computer Vision and Natural Language Processing. I am focused on exploring ways to leverage large amounts of data to process, understand and interpret visual and linguistic contexts. My work during my Ph.D. has progressively followed three key research questions: (a) how can we create systems that can learn with limited annotated data and adapt to new tasks and novel criteria [7, 4, 11, 12, 2]? (b) how can we create systems able to encode real-world concepts with granularity in a robust manner [9, 13, 6, 4]? (c) is it possible to create such a system with alternative data, complying with privacy protection principles and avoiding cultural bias [8, 10, 5]? – Given my work's intersection with Computer Vision and Natural Language Processing, I am focused on analyzing and applying algorithms to understand how images and text can interact and model complex patterns, aiming at developing visio-linguistic compositional reasoning systems, in an effective and robust way.

1. Learning with limited annotated data.

Access to annotated data has been critical in achieving human-like performance in a wide range of computer vision and textual understanding tasks. But annotated data is typically limited or expensive to obtain. *How can we create systems that can learn with limited annotated data and adapt to new tasks and novel criteria?* Learning with large amounts of data with limited annotations and learning to generalize to novel data and tasks are two pilar goals in my research. In my work [7], I proposed a pseudo-labeling approach that exploits curriculum learning principles, which is surprisingly robust to out-of-distribution data compared to other methods. The idea is to teach one model to make predictions based on the initial annotated data, and then use this new information to train a new model. With our curriculum approach, the model first learns from easy samples and then progressively move toward harder samples. This work was particularly important because it showed that this label-propagation idea could be viable and competitive against the dominant consistency regularization methods. Recent work combines both pseudo-labeling and consistency regularization methods to achieve state-of-the-art results. I've also explored ways of learning meaningful relationships between images and text to use one as a complement to the other [2], avoiding forgetting when learning from new data [11, 12], which is particularly relevant for applications in which previous data may not be accessible, or when models are removed from public repositories and are no longer available. This is also relevant given that data rehearsal would be prohibitively expensive, unrealistic, or may lead to privacy or ethical issues.

2. Granularity and alignment.

Generalization is the ability to infer and act on new scenarios making assumptions based on prior experience. This experience can be leveraged from diverse types of information [6]. Yet, granularity in data is important. For a computer vision system, distinguishing between a dog and a cat may be easy. However, it becomes more challenging when that system is required to identify a bird with a red chest and yellow crown flying on top of the green bird, among a flock of different bird species flying all over the place in a forest. Even with modern success in training large-scale models with billions of data, these systems have difficulties understanding attributes among objects, actions among agents, and relations among both. *How can we create systems able to encode real-world concepts with granularity in a robust manner?* In [9, 13], we were able to model a complex scene using meaningful representations of the visual and textual descriptions by aligning both modalities. Since we can successfully align these representations, I've also explored data augmentation techniques to expand the underlying knowledge present in the observed data. In [4], we used evolutionary techniques to find interesting ways to interpolate images for data augmentation. We applied a grid-like mask over a set of images and interpolated random sample pairs. I also wanted to investigate if there were better ways to do this interpolation by selecting specific images or classes. Since previous efforts exploited Reinforcement Learning to find augmentation policies in single images, I wanted to explore evolutionary approaches given their nature of being more explicit and easier to trace back all the steps for a particular decision. In addition, I've explored implicit ways of learning meaningful relationships between images and text to use one as a complement to the other. In [3], I explored the Zero-shot Learning setting, which allows a system to leverage textual descriptions to enable this type of generalization by transferring the knowledge from the seen domain (i.e., the available text) to the visual domain (i.e., the unseen image). By training a model to generate new synthetic features, we could enlarge the training set, so that the aligned representations for all seen and unseen classes can be used to train a classifier in a supervised manner. This approach was effective, but the performance gains among all works in this area reached a point beyond which there was no significant progress. Thus, in [2], I conducted a large-scale study and analysis of different zero-shot learning methods, comparing training objectives, model architectures, and feature aligning techniques to benchmark how different visual features impact the performance of models trained for this task. My study reveals non-trivial findings in which the granularity of the data plays a major role when leveraging large-scale models.

3. Synthetic data generation, compositionality and privacy protection.

Adding more data has become the de facto solution to train more functional models. But weakly supervised data is insufficient to learn specific structural patterns of the world, and existing large-scale models still fail on common sense tasks requiring compositional inference. *Is it possible to create such a system with alternative data, complying with privacy protection principles and avoiding cultural bias?* By creating a framework that can guide an agent to generate samples from diverse sources (text, image, audio) to train an unbiased model, we can reduce the need to reiterate over data points that are noisy or unnecessary to re-learn. In this sense, synthetic data is an unlimited resource ready to be explored. From the computer vision point of view, there have been remarkable advances in synthetic 3D environments. In my most recent work [8, 10, 5], I've explored ways to generate realistic synthetic data, which has proven useful in training a model that can perform well under real test data [8] and can be leveraged to teach large models to perform compositional reasoning without compromising their zero-shot capabilities [10, 5]. Taking into account data privacy concerns, I believe that synthetic data is also a reliable source that can mitigate existing ethical issues when training large-scale models.

Conclusion & Future Work.

I plan to continue exploring hyper-realistic synthetic data generation techniques to train large-scale models. As I've been focusing on compositionality aspects of images and text pairs, it's been crucial to include people in the generated images. Unfortunately, digital humans are frequently absent from widely used 3D asset collections. While existing large-scale synthetic datasets often concentrate on the realistic placement of objects within a scene, they usually exclude humans and animals. In my current work, I've been focusing on integrating articulated synthetic humans that can interact with the 3D environment to close the gap between synthetic simulations and the real world, and enable physical interactions. Data quality is a big concern when training large-scale models. Aiding a system to add variations to the visual and textual world and leverage better data would explicitly allow diverse and rich information from which we can model diverse sets of problems.

[1] Digital Around the World — DataReportal – Global Digital Insights — datareportal.com. `https://datareportal.com/global-digital-overview`. [Accessed 11-Jul-2023].

[2] **Paola Cascante-Bonilla**, Leonid Karlinsky, James Seale Smith, Yanjun Qi, and Vicente Ordonez. On the transferability of visual features in generalized zero-shot learning, 2022.

[3] **Paola Cascante-Bonilla**, Yanjun Qi, and Vicente Ordonez. Generalized zero-shot learning via normalizing flows. `https://www.youtube.com/watch?v=xWUWXCPywq8`, October 2021.

[4] **Paola Cascante-Bonilla**, Arshdeep Sekhon, Yanjun Qi, and Vicente Ordonez. Evolving image compositions for feature representation learning. In *British Machine Vision Conference (BMVC)*, November 2021.

[5] **Paola Cascante-Bonilla**, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gül Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, and Leonid Karlinsky. Going beyond nouns with vision & language models using synthetic data, 2023.

[6] **Paola Cascante-Bonilla**, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. Moviescope: Large-scale analysis of movies using multiple modalities, 2019.

[7] **Paola Cascante-Bonilla**, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Self-paced pseudo-labeling for semi-supervised learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[8] **Paola Cascante-Bonilla**, Hui Wu, Letao Wang, Rogerio Feris, and Vicente Ordonez. Simvqa: Exploring simulated environments for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[9] **Paola Cascante-Bonilla**, Xuwang Yin, Vicente Ordonez, and Song Feng. Chat-crowd: A dialog-based platform for visual layout composition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 138–142, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] Sivan Doveh, Assaf Arbelle, Sivan Harary, Amit Alfassy, Roei Herzig, Donghyun Kim, **Paola Cascante-Bonilla**, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, et al. Dense and aligned captions (dac) promote compositional reasoning in vl models. *arXiv preprint arXiv:2305.19595*, 2023.

[11] James Seale Smith, **Paola Cascante-Bonilla**, Assaf Arbelle, Donghyun Kim, Rameswar Panda, David Cox, Diyi Yang, Zsolt Kira, Rogerio Feris, and Leonid Karlinsky. Construct-vl: Data-free continual structured vl concepts learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14994–15004, 2023.

[12] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, **Paola Cascante-Bonilla**, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919, 2023.

[13] Fuwen Tan, **Paola Cascante-Bonilla**, Xiaoxiao Guo, Hui Wu, Song Feng, and Vicente Ordonez. Drill-down: Interactive retrieval of complex scenes using natural language queries. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.