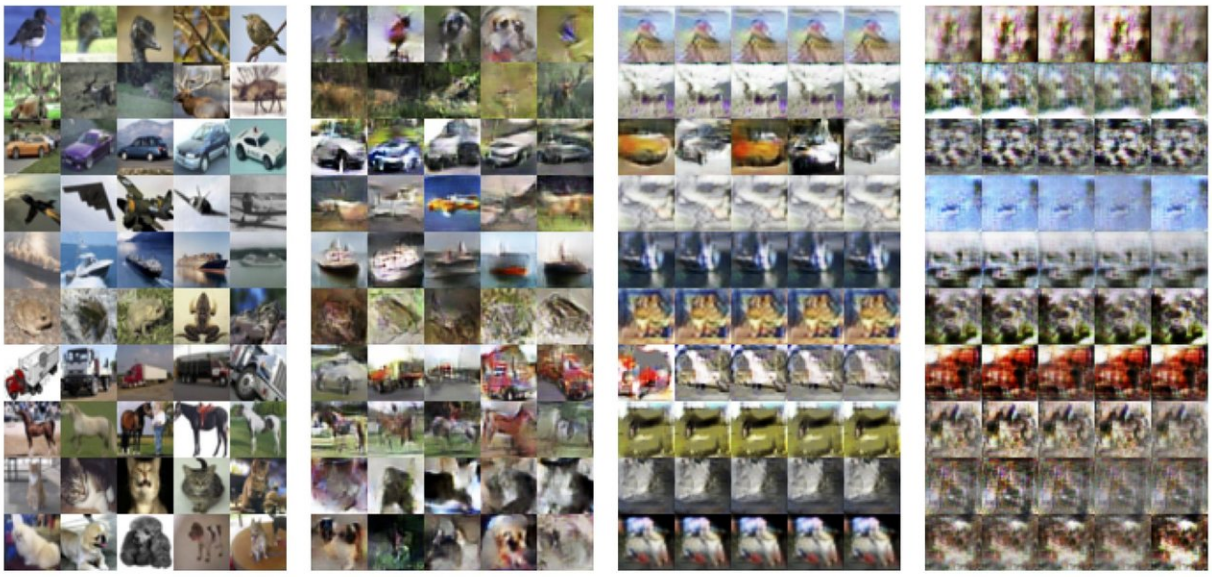
Restoring balance in machine learning datasets

October 11 2018, by Giovanni Mariani



(a) Real image samples (b) *BAGAN* (c) *ACGAN* (d) *Simple GAN*

Five representative samples for each class (row) in the CIFAR-10 dataset. For each class, these samples are obtained with generative models trained after dropping from the training set 40% of the images of that specific class. Credit: IBM

If you want to teach a child what an elephant looks like, you have an infinite number of options. Take a photo from National Geographic, a stuffed animal of Dumbo, or an elephant keychain; show it to the child; and the next time he sees an object which looks like an elephant he will

likely point and say the word.

Teaching AI what an elephant looks like is a bit different. To train a machine learning algorithm, you will likely need thousands of elephant images using different perspectives, such as head, tail, and profile. But then, even after ingesting thousands of photos, if you connect your algorithm to a camera and show it a pink elephant keychain, it likely won't recognize it as an elephant.

This is a form of data bias, and it often negatively affects the accuracy of deep learning classifiers. To fix this bias, using the same example, we would need at least 50-100 images of pink elephants, which could be problematic since pink elephants are "rare".

This is a known challenge in machine learning communities, and whether its pink elephants or road signs, small data sets present big challenges for AI scientists.

Restoring balance for training AI

Since earlier this year, my colleagues and I at IBM Research in Zurich are offering a solution. It's called BAGAN, or balancing generative adversarial networks, and it can generate completely new images, i.e. of pink elephants, to restore balance for training AI.

(a) Real image samples (b) *BAGAN* (c) *ACGAN* (d) *Simple GAN*

Five representative samples generated for the three most represented majority classes in the GT- SRB dataset. Credit: IBM

Seeing is believing

In the paper we report using BAGAN on the German Traffic Sign Recognition Benchmark, as well as on MNIST and CIFAR-10, and when compared against state-of-the-art GAN, the methodology outperforms all of them in terms of variety and quality of the generated images when the training dataset is imbalanced. In turn, this leads to a higher accuracy of final classifiers trained on the augmented dataset.

(a) Real image samples (b) *BAGAN* (c) *ACGAN* (d) *Simple GAN*

Five representative samples generated for the three least represented minority classes in the GT-SRB dataset. Credit: IBM

More information: BAGAN: Data Augmentation with Balancing GAN. Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. arxiv.org/abs/1803.09655

The work was recently published and made open-source. Visit Github today to try it for free github.com/IBM/BAGAN