Compression of deep generative models

Gonçalo Mordido
Ph.D. Student
Hasso Plattner Institute

Motivation

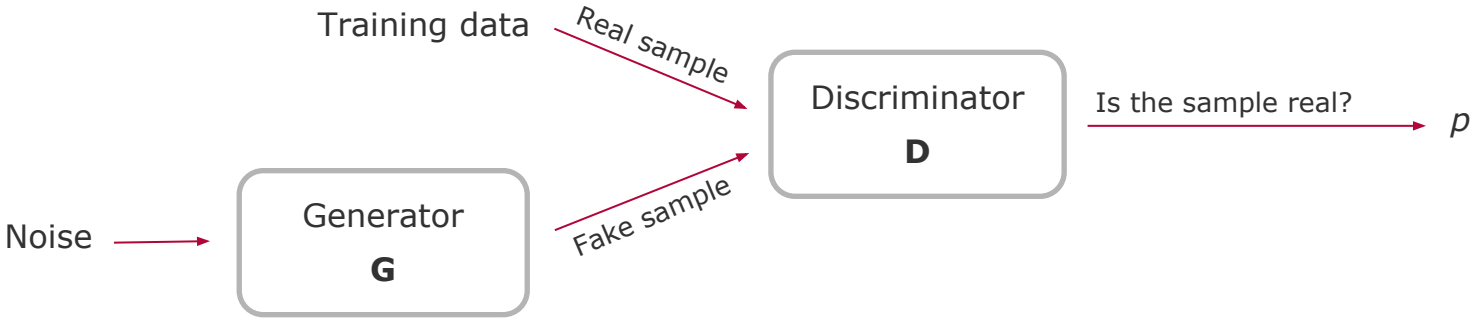
- Generative models are usually **memory and computation intensive**
  - Model size
  - Training and inference cost
- Deployment on **edge devices** is challenging
  - Smartphones, smartwatches, etc
  - Memory and power constrained
- **Compression techniques** reduce the model size and complexity
  - Trade-off between compression levels and model performance

Generative models

- **Learn** the data distribution
- **Generate** fake samples
  - Different but resembling real samples
- Use cases
  - Automatic customer service (e.g. conversational AI)
  - Data augmentation (e.g. dataset improvement)
  - Data manipulation (e.g. style transfer)
  - Data simulation (e.g. self-driving cars)

Generative Adversarial Networks (*Goodfellow et al., 2014*)

- **GANs** for short
- **Generator (G)** that learns the real data distribution to generate fake samples
- **Discriminator (D)** that attributes a probability *p* of confidence of a sample being real (*i.e.* coming from the training data)



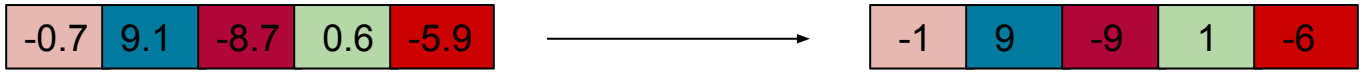
Generative Adversarial Networks (*Goodfellow et al., 2014*)

- Both models are trained together **(minimax game):**
  - G: Increase the probability of D making mistakes
  - D: Classify real samples with greater confidence
- G *slightly changes* the generated data based on D's feedback

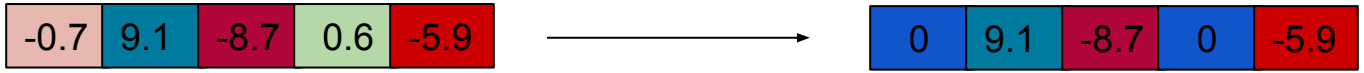
Images from Karras et al., 2019

Quantization and pruning

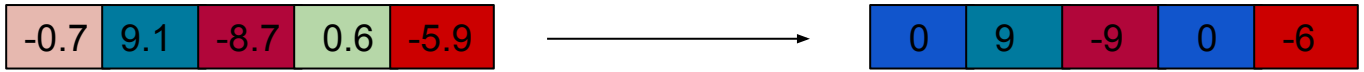
- Quantization
  - Continuous to discrete values (float to integer)
  - Reduce the number of bits



- Pruning
  - Set some values to zero
  - Skip operations



- Quantization and pruning
  - Combined benefits



When to compress?

- Compression during training
  - Generative models tend to suffer from **training instability**
  - Compression techniques may increase such instability
  - May need to compress several models

- Compression after training
  - Compress only the generative model
  - Improve performance (e.g. memory and compute) only at **inference time**

- Train generative model on the cloud

- Compress the pre-trained model and run it on an edge device

How to evaluate the generated samples?

- A generated sample is indistinguishable from a real sample (**quality**)

- Generated samples differ from each other (**diversity**)

Images from Karras et al., 2019

Post-compression evaluation

- Evaluation is usually done with another **pre-trained model**
  - Human evaluation is too expensive
- Are **quality and diversity** affected the same?
  - Requires a separate assessment of quality and diversity
- What about the quality of **individual samples**?
  - The overall generated set may be affected by compression
  - However, same samples may be more affected than others
- Sensitive to different **data domains**
  - E.g. Images, video, text and audio

Conclusion

- Compression may enable running generative models on edge devices
  - Quantization and pruning
- Generative models are often unstable during training
  - Post-training compression
- Evaluation of the compressed models is necessary for real-world applications
  - Quality vs diversity

Thank you!

Goncalo Mordido
Ph.D. Student
Hasso Plattner Institute