Non-Generative Representation Learning

Topics:

- **Strengths/Weaknesses of Unsupervised Learning Methods Covered Thus Far**
    - Main Reading: The slides themselves =)

- **Semi-Supervised Learning**
    - Main Reading: https://arxiv.org/abs/1804.09170
        - Temporal Ensembling for semi-supervised learning: https://arxiv.org/pdf/1610.02242.pdf
        - Virtual Adversarial Training: https://arxiv.org/pdf/1704.03976.pdf
        - Wide Residual Networks: https://arxiv.org/pdf/1605.07146.pdf

- **GPT2**
    - Main Reading: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
    - Related Materials posted during class discussion:
        - The Illustrated GPT-2 http://jalammar.github.io/illustrated-gpt2/
        - GPT-2: 6-Month Follow-Up https://openai.com/blog/gpt-2-6-month-follow-up/
        - GPT-2 Web Demo by HuggingFace https://transformer.huggingface.co/
        - Tool to detect automatically generated text http://gltr.io/

Overview:

- **Autoregressive Models**
    - Examples: PixelRNN, PixelCNN, PixelCNN++, PixelSNAIL
    - Some famous architectures: Transformer, Dilated Conv
    - Negatives: No single layer of learned representation; sampling time is slow; not directly usable for downstream tasks;no interpolations.
- **Flow models**
    - Useful link: https://lilianweng.github.io/lil-log/2018/10/13/flow-based-deep-generative-models.html
    - Drawbacks
        - Input z is the same size of x, which makes model large.
        - Do not allow low-dimensional embedding
        - Need specific initialization
- **Latent Variable models**
    - Well known VAE applications

-

        - Sketch-RNN: https://magenta.tensorflow.org/assets/sketch_rnn_demo/interp.html
        - World Models: https://worldmodels.github.io/
        - beta-VAE: https://docs.google.com/presentation/d/12uZQ_Vbvt3tzQYhWR3BexqOzbZ-8AeT_jZjuuYjPJiY/pub?start=true&loop=true&delayms=30000&slide=id.g1329951dde_0_0

    - Implicit models
        - Generative Adversarial Network (GAN)
            - Original GAN: concepts, generator, discriminator
            - DCGAN: a CNN-based GAN
            - styleGAN: https://github.com/NVlabs/stylegan
            - BigGAN: generate images with high resolution
- **Density (definition)**:
    - Explicit density:
        - Because we need to estimate the data
        - Use for outlier detection
        - Comparing two classes/sample
        - Gives us latent variables
            - able to lower the no. of dimensions
            - control how many we want, often once we can control, we can explain
    - Implicit density:
        - Goal: want to generate data that follow the distribution of given samples, without mapping out the exact density distributions
        - We do this by following a minmax optimization of discriminator and generator in a GAN
        - We don’t need to work out density distributions at all. Let the network figure out the manifold that is similar to the given data.
- **Semi-supervised learning**
    - Insight: how to use unlabeled data to improve the supervised model performance
    - Pi-model
        - Utilize stochastic argumentation on the input x_i
        - For labeled data, the algorithm has 2 parts. (1) Supervised: Train a normal classifier. (2) Unsupervised: Train a data representation without using the label information. If data does not have a label, skip (1).
        - The loss is weighted sum over the cross-entropy of (1) and squared difference of (2)
    - Temporal ensembling
        - A simple and efficient method for training deep neural networks in a semi-supervised setting

-

        - Forms a consensus prediction of the unknown labels using the outputs of the network-in-training on different epochs, and most importantly, under different regularization and input augmentation conditions
    - Virtual adversarial training
        - i.e., adversarial training for unsupervised (i.e., unlabeled data)
        - Miyato and his team applied these ideas to do 'virtual adversarial training' for semi-supervised learning, which is a particularly great fit for models that have to contend with sparsely labeled data. With 'virtual adversarial training', we don't use the labels of our training set but rather the conditional probability that an image will have label X. In other words, when we input an image of a panda and our model predicts that we have 40% panda, 20% bird, 1 % car, and so on, that distribution itself becomes the label during adversarial training. This means that we add noise to the image with the adversarial method but we still tell our classifier that the correct label is: 40% panda, 20% bird, 1 % car, and so on.

          This allows us to do semi-supervised learning. For images with labels, we can follow the previous adversarial example and tell the model that we know the label. And for unlabeled images, we let our model predict the labels (40% panda, 20% bird, 1 % car, and so on) and then we use that to perturb our images. As explained above, our adversarial method tries to makes the model fail the most and in this case the model tries to make perturbations to the image that maximize the divergence between the predicted label and the correct label distribution.

          (From https://medium.com/inside-machine-learning/placeholder-3557ebb3d470)

- **GPT-2**
    - Exploration: Curiosity Help
        - Unsupervised learning: core concepts
        - Neural dictionary: attention is important.
        - Transformer: parallelization over time
    - Path to GPT-2
        - GPT-2: Big improvement across the board
        - Open Domain Question Answering
        - Reading Comprehension
        - GPT-2: Programming with Words
            - E.g. Summarization with "TL;DR:" as a keyword, or translation task with "=" as a keyword.