IFT 6268 Self-Supervised Representation Learning

Michael Noukhovitch

Fall 2020

Notes written from Aaron Courville's lectures.

Contents

1 Introduction

1.1 Motivation

We want good representation learning but

- scaling supervised learning isn't feasible
- we need to learn useful representations unsupervised, but need more info

Self-supervised learning (SSL) recover useful/semantic representations by training models to answer specific questions about the data

- between supervised and unsupervised
- can procedurally generate infinite annotation
- answering the question requires fundamental understanding
- main challenge is choosing a good question that allows to learn *without* extra labels

1.2 Autoencoders

Generative models (AE, VAE) have been historically ineffective compared to supervised

- caveat of LM (BERT predicting input)
- caveat of small data (unsupervised pretraining)

Old school AE

- couldn't learn semantics of MNIST
- even adding depth didn't help

denoising AE (Vincent et al, 2008) reproduces from noisy x

- $x + \text{noise} = \tilde{x}$ used to predict x
- representation robust to noise
- formally corresponds to learning an energy function that has lows at true x

contractive AE (Rifai et al, 2011) adds loss on all information

- autoencoder loss to keep "good" information
- secondary loss on the jacobian of the encoder to throw away all information
- total effect to reduce "bad" information kept
- can use second order methods and might be more stable than dAE

1.3 Transfer Learning

Domain D consists of a feature space $\mathbb{X}$ and marginal po
Given a source domain D_s and learning task T_s, a target domain D_t and learning task T_t.
Transfer learning aims to improve learning a function f to

1. **inductive** same domain, different tasks
2. **transductive** different domain, same task **domain adaptation**
3. **unsupervised** both different

SSL is usually source unlabelled, target labelled
For SSL, the source task T_s

- unsupervised
- extracts semantic information from input
- learns correct invariances

1.4 Image Methods

Rotation prediction: how much each image is rotated

- generate your own label
- but learning the answer gets you some semantic knowledge about images

GANs: is it SSL?

- if you're just creating a generator, no (just generative modelling)
- learning a discriminator augmented with fake data (Salimans et al, 2016), yes

1.5 Contrastive Methods

Mutual information $I(X, Y) = D_{KL}(P_{X,Y}||P_X \otimes P_Y)$

- intersection of marginal entropies
- difficult to compute

MINE (Belghazi et al, 2018) optimize a lower bound to MI

- encode an image with a network
- learn a discriminator with lower bound MI loss

Deep Infomax (Hjelm et al, 2019) try to learn "important" image regions

- use local image patches
- MI between global vector and local patches

CPC (van der Oord, 2018)

- predict feature vectors from context vectors
- single neurons learn semantic meaning

1.6 Iterated Learning

Iterated learning uses learners to teach other learners

- compositional structure of NL may have emerged through teaching language (Kirby et al, 2014)
- may encourage compositional structure in neural models
- may encourage systematic generalization

Self-training similarly learns from its own pseudo-labels

- self-training with noisy students (Xie et al, 2020) iterates on imagenet

Systematic generalization is generalizing through shared rules between training and testing, not shared distribution

- seq2seq models can't regularly do that (Lake and Baroni, 2016)
- maybe SSL can help here