Advanced Machine Learning
Summer 2019
Part 19 – Variational Autoencoders II
10.07.2019

Prof. Dr. Bastian Leibe
RWTH Aachen University, Computer Vision Group
http://www.vision.rwth-aachen.de

Course Outline

• Regression Techniques
  – Linear Regression
  – Regularization (Ridge, Lasso)
  – Kernels (Kernel Ridge Regression)
• Deep Reinforcement Learning
• Probabilistic Graphical Models
  – Bayesian Networks
  – Markov Random Fields
  – Inference (exact & approximate)
  – Latent Variable Models
• Deep Generative Models
  – Generative Adversarial Networks
  – Variational Autoencoders

Topics of This Lecture

• Recap: Variational Autoencoders
  – Autoencoders as Generative Models
  – Intractability
  – Variational Approximation
  – Evidence Lower Bound (ELBO)
• Applying VAEs
  – VAE Training
  – VAE Data Generation

Recap: Autoencoders

L2 Loss function

Reconstructed input data
Features
Input data

• After training
  – Throw away the decoder part
  – Encoder can be used to initialize a supervised model
  – Fine-tune encoder jointly with supervised model
  – Idea used in the 90s and early 2000s to pre-train deeper models

Recap: Variants of Autoencoders

L2 Loss function

Reconstructed input data
Features
Input data

• Regularized Autoencoders
  – Include a regularization term to the loss function: $L(x, g(f(x))) + \Omega(z)$
  – E.g., enforce sparsity by an L1 regularizer $\Omega(z) = \lambda \sum |z_i|$
• **Variational Autoencoders**
  - Rather than the reconstruction loss, minimize $L(x, g(f(\tilde{x})))$ where $\tilde{x}$ is a copy of $x$ that has been corrupted by some noise.
  - Denoising forces $f$ and $g$ to implicitly learn the structure of $p_{data}(x)$.

**Reconstructed input data**

**Features**

**Input data**

**Loss function**

**Encoder**

**Decoder**

- **Denoising Autoencoder (DAE)**
  - Rather than the reconstruction loss, minimize $L(x, g(f(\tilde{x})))$ where $\tilde{x}$ is a copy of $x$ that has been corrupted by some noise.
  - Denoising forces $f$ and $g$ to implicitly learn the structure of $p_{data}(x)$.

**Reconstructed input data**

**Features**

**Input data**

**Loss function**

**Encoder**

**Decoder**

**Reconstruction**

**Reconstructed input data**

**Features**

**Input data**

**Loss function**

**Encoder**

**Decoder**

**VAE Data Generation**

**VAE Training**

Applying VAEs

**Topics of This Lecture**

- **Recap: Variational Autoencoders**
  - Autoencoders as Generative Models
  - Intractability
  - Variational Approximation
  - Evidence Lower Bound (ELBO)
- **Applying VAEs**
  - VAE Training
  - VAE Data Generation

**Reconstruction**

**Variational Autoencoders**

**Reconstructed input data**

**Features**

**Input data**

**Loss function**

**Encoder**

**Decoder**

- **Denoising Autoencoder (DAE)**
  - Rather than the reconstruction loss, minimize $L(x, g(f(\tilde{x})))$ where $\tilde{x}$ is a copy of $x$ that has been corrupted by some noise.
  - Denoising forces $f$ and $g$ to implicitly learn the structure of $p_{data}(x)$.
Applying Variational Autoencoders

• Putting it all together…

  - Maximizing the likelihood lower bound
    \[ \mathbb{E}[\log p_d(x^{(i)} \mid z)] - D_{KL}(q(z^{(i)} \mid x^{(i)}) || p(z^{(i)})) \]
    \[ \mathcal{L}(x^{(i)}, \theta, \phi) \]

  - Let’s look at computing the bound for a given minibatch of input data (forward pass)…

Input data \( x \)

Applying Variational Autoencoders

• Putting it all together…

  - Maximizing the likelihood lower bound
    \[ \mathbb{E}[\log p_d(x^{(i)} \mid z)] - D_{KL}(q(z^{(i)} \mid x^{(i)}) || p(z^{(i)})) \]
    \[ \mathcal{L}(x^{(i)}, \theta, \phi) \]

  Make approximate posterior distribution close to prior

Encoder \( q(z \mid x) \)

Sample \( z \) from \( z \sim \mathcal{N}(\mu_{z^{(i)}}, \Sigma_{z^{(i)}}) \)

Applying Variational Autoencoders

• Putting it all together…

  - Maximizing the likelihood lower bound
    \[ \mathbb{E}[\log p_d(x^{(i)} \mid z)] - D_{KL}(q(z^{(i)} \mid x^{(i)}) || p(z^{(i)})) \]
    \[ \mathcal{L}(x^{(i)}, \theta, \phi) \]

  Make approximate posterior distribution close to prior

Encoder \( q(z \mid x) \)

Sample \( z \) from \( z \sim \mathcal{N}(\mu_{z^{(i)}}, \Sigma_{z^{(i)}}) \)

Applying Variational Autoencoders

• Putting it all together…

  - Maximizing the likelihood lower bound
    \[ \mathbb{E}[\log p_d(x^{(i)} \mid z)] - D_{KL}(q(z^{(i)} \mid x^{(i)}) || p(z^{(i)})) \]
    \[ \mathcal{L}(x^{(i)}, \theta, \phi) \]

  Make approximate posterior distribution close to prior

Encoder \( q(z \mid x) \)

Sample \( z \) from \( z \sim \mathcal{N}(\mu_{z^{(i)}}, \Sigma_{z^{(i)}}) \)

Applying Variational Autoencoders

• Putting it all together…

  - Maximizing the likelihood lower bound
    \[ \mathbb{E}[\log p_d(x^{(i)} \mid z)] - D_{KL}(q(z^{(i)} \mid x^{(i)}) || p(z^{(i)})) \]
    \[ \mathcal{L}(x^{(i)}, \theta, \phi) \]

  Compute this forward pass for every minibatch of input data, then backprop

Encoder \( q(z \mid x) \)

Sample \( z \) from \( z \sim \mathcal{N}(\mu_{z^{(i)}}, \Sigma_{z^{(i)}}) \)
Variational Autoencoders: Generating Data

• Use decoder network
  – Now sample \( z \) from prior
  \[ p(x|z) \]
  \[ q \]
  \[ \mathcal{N} \]
  \[ \mu \]
  \[ \sigma \]
  \[ x \]
  \[ z \]
  \[ \mathcal{N} \]
  \[ 0 \]
  \[ I \]
  \[ \mu \]
  \[ \sigma \]
  \[ x \]
  \[ z \]
  \[ \mathcal{N} \]

D. Kingma, M. Welling, Auto-Encoding Variational Bayes, ICLR 2014

Some More Learned Manifolds

32x32 CIFAR-10
Labeled Faces in The Wild

Variational Autoencoders: Generating Data

• Another example
  – Learning a face manifold

• Comments
  – Diagonal prior on \( z \)
  \[ \Rightarrow \]
  Independent latent variables
  – Different dimensions of \( z \) encode interpretable factors of variation

Degree of smile
Head pose

Summary: Variational Autoencoders

• Idea
  – Probabilistic Spin on traditional autoencoders
  – Intractable density \( \Rightarrow \) derive & optimize a variational lower bound
• Pros
  – Principled approach to generative models
  – Allows inference of \( q_\phi(z|x) \), can be useful feature representation for other tasks
• Cons
  – Only maximizes lower bound of likelihood
  – Samples blurrier and lower quality compared to state-of-the-art (GANs)
• Active area of research
  – More flexible approximations, e.g., GMMs instead of diagonal Gaussian

Combinations

• Attempts at combining the advantages
  – Use learned feature representations in the GAN discriminator as basis for the VAE reconstruction objective
  – Replacing element-wise errors with feature-wise errors to better capture the data distribution

Results

Samples from different generative models
Reconstructions from different autoencoders

– VAE:
– VAE\(\text{-GAN} \):
– VAE/GAN:

VAE/GAN trained together
References

• Variational Auto-Encoders