Paper Introduction:

Context Encoder: feature learning by inpainting[1]

講演者：CAO SHILEI

指導教員：櫻井幸一

科目名：情報学演習

日時：2019年11月11日

場所：工学部第9（総合学習プラザ1F）

Key words: image inpainting, adversarial learning, autoencoder

1. Introduction

Filling missing pixels of an image, also referred as image inpainting, is an important task in computer vision. It has many applications in photo editing, image-based rendering and computational photography, the core challenge of image inpainting lies in synthesizing visually realistic and semantically plausible pixels for the missing regions that are coherent with existing ones.

Rapid progress in deep convolutional neural networks (CNN)[2] and generative adversarial networks (GAN) [3] inspired recent works to formulate inpainting as a conditional image generation problem where high-level recognition and low-level pixel synthesis are formulated into a convolutional encoder-decoder network jointly trained with adversarial networks to encourage the coherency between generated and existing pixels.

2. Related work

2.1 Adversarial Neural Network (GAN)

Adversarial Neural Network (GAN)[3,4] can be considered as a framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game.

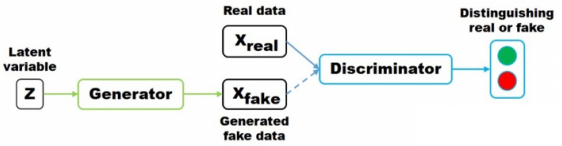

Figure1: the structure of Adversarial Neural Network

The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistiguishable from the genuine articles[5].

In the case where G and D are defined by multilayer perceptrons or neural network layers, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Deal the difficulty of

approximating many intractable probabilistic computations that arise in maximum likelihood estimation .

2.2 Autoencoder

Autoencoders (AE)[6] are neural networks that aims to copy their inputs to their outputs. They work by compressing the input into a latent-space representation, and then reconstructing the output from this representation. This kind of network is composed of two parts: encoder and decoder.

Encoder is the part of the network that compresses the input into a latent-space representation.

Decoder aims to reconstruct the input from the latent space representation.

Using backpropagation, this unsupervised algorithm continuously trains itself by setting the target output values to equal the inputs. This forces the smaller hidden encoding layer to use dimensional reduction to eliminate noise and reconstruct the inputs.

Autoencoders are learned automatically from data examples. It means that it is easy to train specialized instances of the algorithm that will perform well on a specific type of input and that it does not require any new engineering, only the appropriate training data.

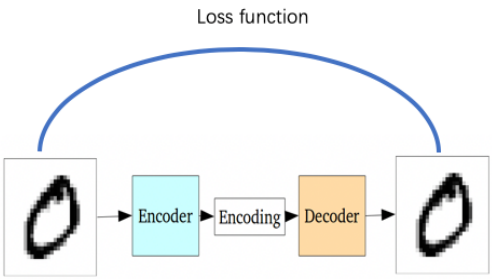

Figure2: the structure of autoencoder

3. The Approach

The paper "Context Encoder: feature learning by inpainting"[1] proposed Context Encoders – a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings. In order to succeed at this task, context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s). In this way, this paper proposed a joint loss, including a standard pixel-wise reconstruction loss and an adversarial loss.

3.1 Framework Overview

The overall architecture is a classic Adversarial Neural Network (GAN)[3] but change the generator as a simple encoder-decoder pipeline. The encoder takes an input image with missing regions and produces a latent feature representation of that image. The decoder takes this feature representation and produces the missing image content.

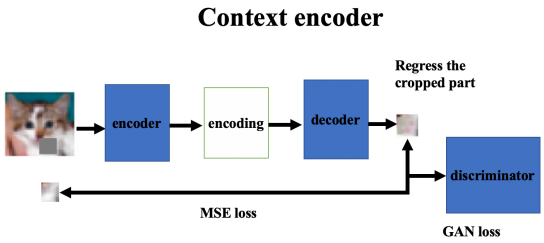

Figure3: the structure of the Context encoder

3.2. The Joint Loss Function

The loss function the paper use is very special since we need to generate the contents of an arbitrary image region conditioned on its surroundings. Thus context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s).

This paper model this behavior by having a decoupled joint loss function to handle both continuity within the context and multiple modes in the output. The reconstruction (L2) loss is responsible for capturing the overall structure of the missing region and coherence with regards to its context, but tends to average together the multiple modes in predictions. The adversarial loss, on the other hand, tries to make prediction look real, and has the effect of picking a particular mode from the distribution.

The Reconstruction Loss

We use a normalized masked L2 distance

as our reconstruction loss function

$$L_{rec}(x) = \left\| x_{gen} - x_{ori} \right\|_2^2 \quad (1)$$

x_{gen} means the generated missing parts of image. x_{ori} means the original masked part of image. It is just comparing pixel to pixel differences.

The Adversarial Loss

Our adversarial loss is based on Generative Adversarial Networks (GAN) [16]. To learn a generative model G of a data distribution, GAN proposes to jointly learn an adversarial discriminative model D to provide loss gradients to the generative model.

$$\min_G \max_D \mathbb{E}_{x \in \mathcal{X}}[\log(D(x))] + \mathbb{E}_{z \in \mathbb{Z}}[\log(1 - D(G(z)))] \quad (2)$$

To customize GANs for this task, one could condition on the given context information. However, conditional GANs don't train easily for context prediction task as the adversarial discriminator D easily exploits the perceptual discontinuity in generated regions and the original context to easily classify predicted versus real samples. We thus use an alternate formulation, by conditioning only the generator (not the discriminator) on context.

$$L_{adv} = \max_D \mathbb{E}_{x \in \mathcal{X}}[\log(D(x))] + [\log(1 - D(F(y)))] \quad (3)$$

y means the input image with the hole masked.

The Joint Loss

The paper defines the overall loss function as

$$L = \lambda_{adv} L_{adv} + \lambda_{rec} L_{rec} \quad (4)$$

In this paper we use $\lambda_{adv} = 0.001$, $\lambda_{rec} = 0.999$.

4. Results

There are the results I simulated the paper and change the data set from ImageNet to Cifar-10.

Original Pre-proceed recovered

Figure4: Result example of Encoder-decoder

We also compare the Semantic Inpainting using different loss function. We can briefly say that Context Encoder with just L2 are well aligned, but not sharp. Using adversarial loss, results are sharp but not coherent. Joint loss alleviate the weaknesses of each of them.

5. Conclusion and Future work

We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods

Unfortunately, this Adversarial algorithm often creates boundary artifacts, distorted structures and blurry textures inconsistent with surrounding areas. Maybe this is likely due to ineffectiveness of convolutional neural networks in modeling long-term correlations between distant contextual information and the hole regions.

Thus, in my future work, I want to optimize the neural network structure and the loss function for improving the training stability and visual quality.

References

[1] DP Kingma, M Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013

[2] Habibi, Aghdam, Hamed. Guide to convolutional neural networks : a practical application to traffic-sign detection and classification. Heravi, Elnaz Jahani. Cham, Switzerland. ISBN 9783319575490. OCLC 987790957, April 2017.

[3] D. Pathak, P. Krahenbuhl, et al. Context encoders: Feature learning by inpainting. arXiv:1604.07379 [cs], April 2016.

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680), 2014

[5]M. Heusel, H. Ramsauer, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. arXiv:1706.08500 [cs, stat], June 2017. arXiv: 1706.08500.

[6] DP Kingma, M Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013