Context Encoders: Feature Learning by Inpainting

2020009507 김나영

목차

- 논문 소개
- 기존 방식과의 차이
- 관련 연구
- Context Encoder
- 결과

논문 소개

- 컨텍스트 인코더 기술: 인페인팅을 통한 특징 학습
  (Context Encoders: Feature Learning by Inpainting)

Context Encoders: Feature Learning by Inpainting

Deepak Pathak Philipp Krähenbühl Jeff Donahue Trevor Darrell Alexei A. Efros
University of California, Berkeley
{pathak,philkr,jdonahue,trevor,efros}@cs.berkeley.edu

Abstract

We present an unsupervised visual feature learning algorithm driven by context-based pixel prediction. By analogy with auto-encoders, we propose 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). When training context encoders, we have experimented with both a standard pixel-wise reconstruction loss, as well as a reconstruction plus an adversarial loss. The latter produces much sharper results because it can better handle multiple modes in the output. We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of our learned features for CNN pre-training on classification, detection, and segmentation tasks. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.

(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

Figure 1: Qualitative illustration of the task. Given an image with a missing region (a), a human artist has no trouble inpainting it (b). Automatic inpainting using our *context encoder* trained with $L2$ reconstruction loss is shown in (c), and using both $L2$ and adversarial losses in (d).

1. Introduction

Our visual world is very diverse, yet highly structured, and humans have an uncanny ability to make sense of this structure. In this work, we explore whether state-of-the-art computer vision algorithms can do the same. Consider the image shown in Figure 1a. Although the center part of the image is missing, most of us can easily imagine its content from the surrounding pixels, without having ever seen that exact scene. Some of us can even draw it, as shown on Figure 1b. This ability comes from the fact that natural images, despite their diversity, are highly structured (e.g. the regular pattern of windows on the facade). We humans are able to understand this structure and make visual predictions even when seeing only parts of the scene. In this paper, we show that it is possible to learn and predict this structure using convolutional neural networks (CNNs), a class of models that have recently shown success across a variety of image understanding tasks.

Given an image with a missing region (e.g., Fig. 1a), we train a convolutional neural network to regress to the missing pixel values (Fig. 1d). We call our model *context encoder*, as it consists of an encoder capturing the context of an image into a compact latent feature representation and a decoder which uses that representation to produce the missing image content. The context encoder is closely related to autoencoders [3, 20], as it shares a similar encoder-decoder architecture. Autoencoders take an input image and try

The code, trained models and more inpainting results are available at the author's project website.

논문 소개

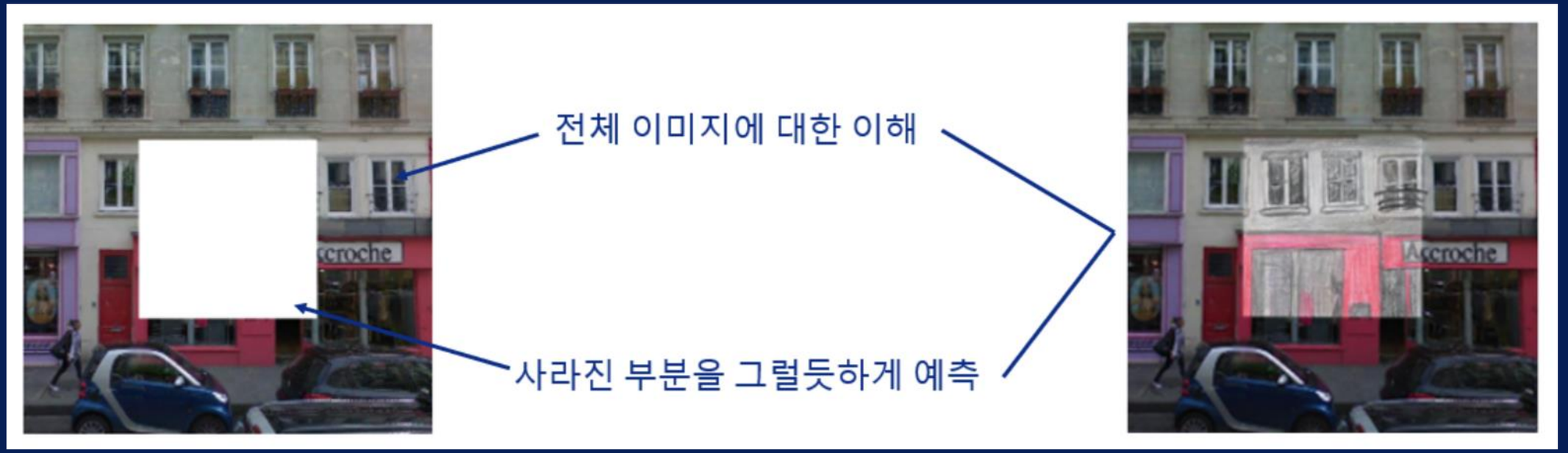
· Image Inpainting 이란?

영상에서 훼손된 부분을 복원하거나 영상 내의 불필요한 문자나 특정 물체를 제거한 후 삭제된 영역을 자연스럽게 채우기 위해 널리 사용되는 기법이다.

논문 소개

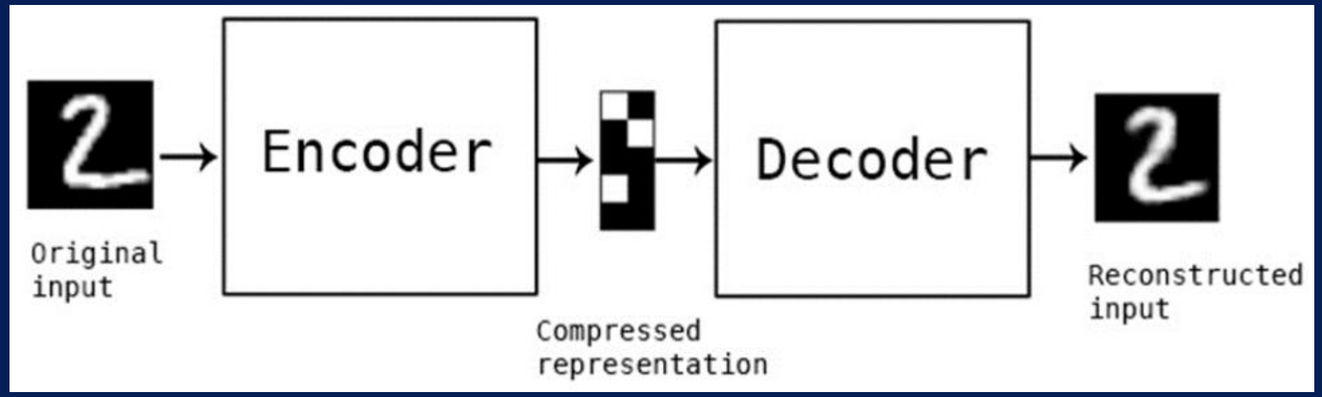
· Context Encoder란?

상황(Context) 기반 픽셀 예측을 이용한 이미지 학습 알고리즘으로 Inpainting을 목적으로 한다.
Auto Encoder와 유사하다.(비지도 학습, Encoder-Decoder 구조)



인코더: 영상의 상황(Context)를 소형의 잠재 표현으로 재구성함.
디코더: 누락된 영상의 컨텐츠(흰 부분)을 생성.

기존 방식과의 차이



· Autoencoder

입력과 동일한 출력을 만드는 것을 목적으로 하는 신경망, 이미지를 더 낮은 차원으로 인코딩하고 다시 원래의 이미지로 디코딩함.

· Denoising Autoencoder

입력할 이미지에 noise를 주고 네트워크 내에서 그 noise를 해결하는(원상태로 만드는) 방식으로 특징을 파악하는 시스템.

기존 방식과의 차이

(c) Context Encoder ($L2$ loss)　　(d) Context Encoder ($L2$ + Adversarial loss)

- **Autoencoder에서 나타나는 문제는 reconstruction loss(재구성 손실)와 adversarial loss(적대적 손실)를 최소화함으로써 제거**

reconstruction loss(재구성 손실): 누락된 영역의 전체 구조를 주변 상황과 관련해서 포착하는데 쓰임 (L2)
adversarial loss(적대적 손실): 누락된 부분을 채우고 일관성을 유지시키는 여러 방법 중 특정 기법을 선택하는 것과 관련이 있다.

- **Context Encoder의 평가는 독립적으로 실시함.**

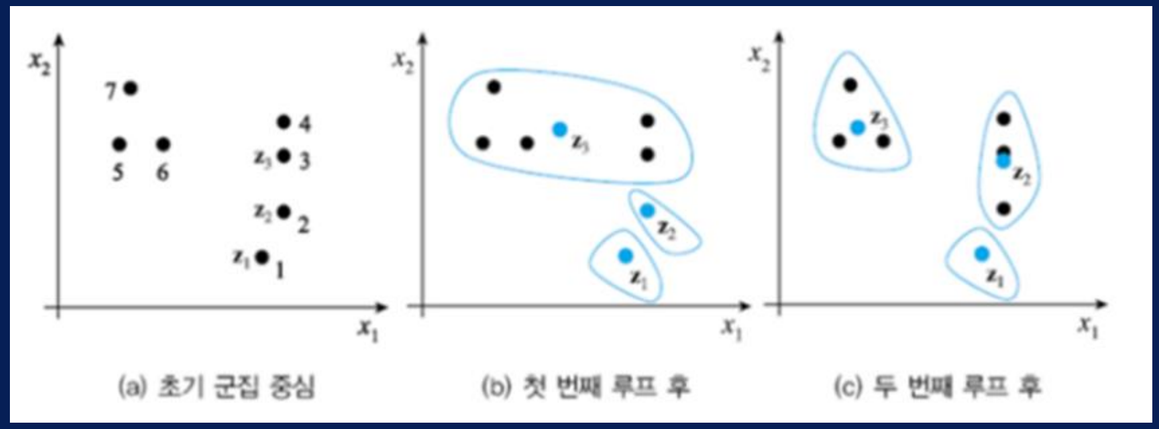
인코더: 영상의 상황(Context)를 소형의 잠재 표현으로 재구성함.
디코더: 누락된 영상의 컨텐츠(흰 부분)을 생성.

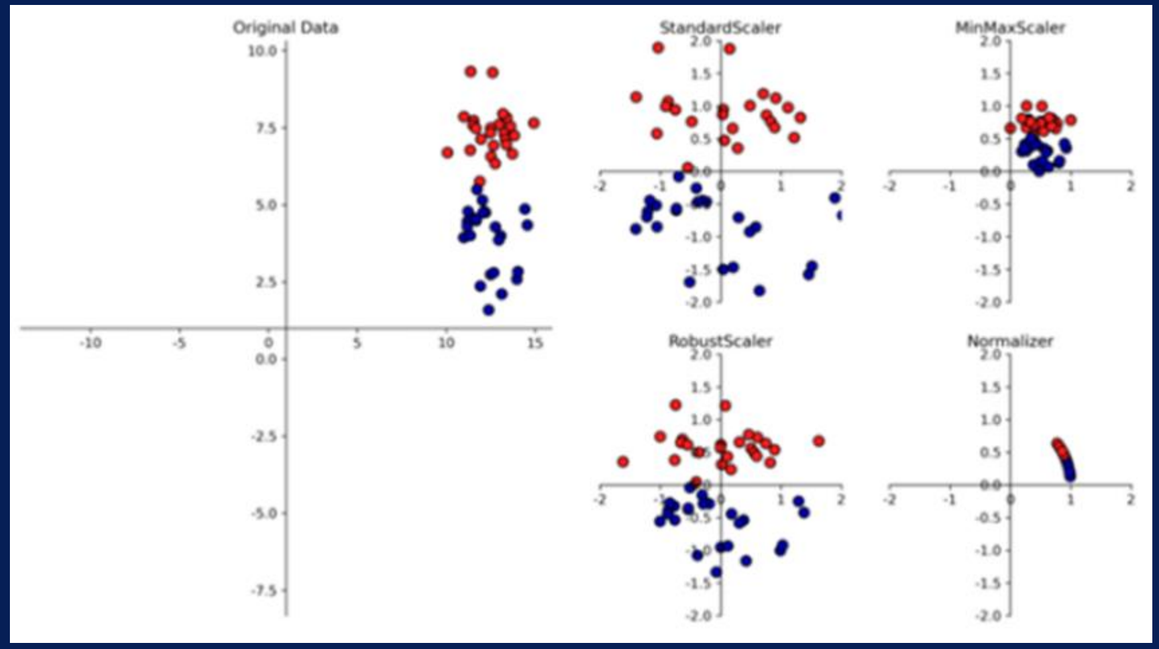
관련 연구

- **비지도 학습 (unsupervised learning)**

정답이 주어지지 않는 학습 전략 (출력 값에 대한 정보없이 학습을 진행하는 것)

ex) 군집화, 데이터 스케일링, 오토 인코더(심층 비지도 학습에서 연구됨)



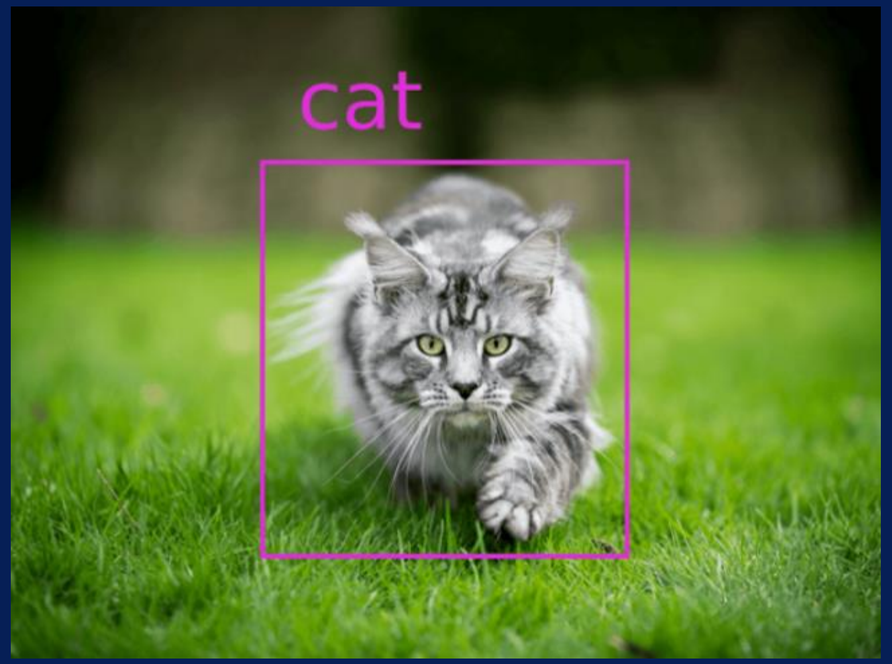


관련 연구

- **약한 지도학습 (weakly supervised learning)**

주어지는 정답에 대한 정보가 제한된 학습.
비지도 학습(unsupervised learning)과 같이 아무 정보도 주어지지 않는 경우와는 다르지만, 일부에 대한 정보만 제공받아 학습하고, 그러한 학습을 통해 제공받지 않은 정보를 예측해내야 한다.

ex) 객체의 클래스 정보만을 제공받고 영상 내의 객체 위치를 예측, 데이터 라벨링

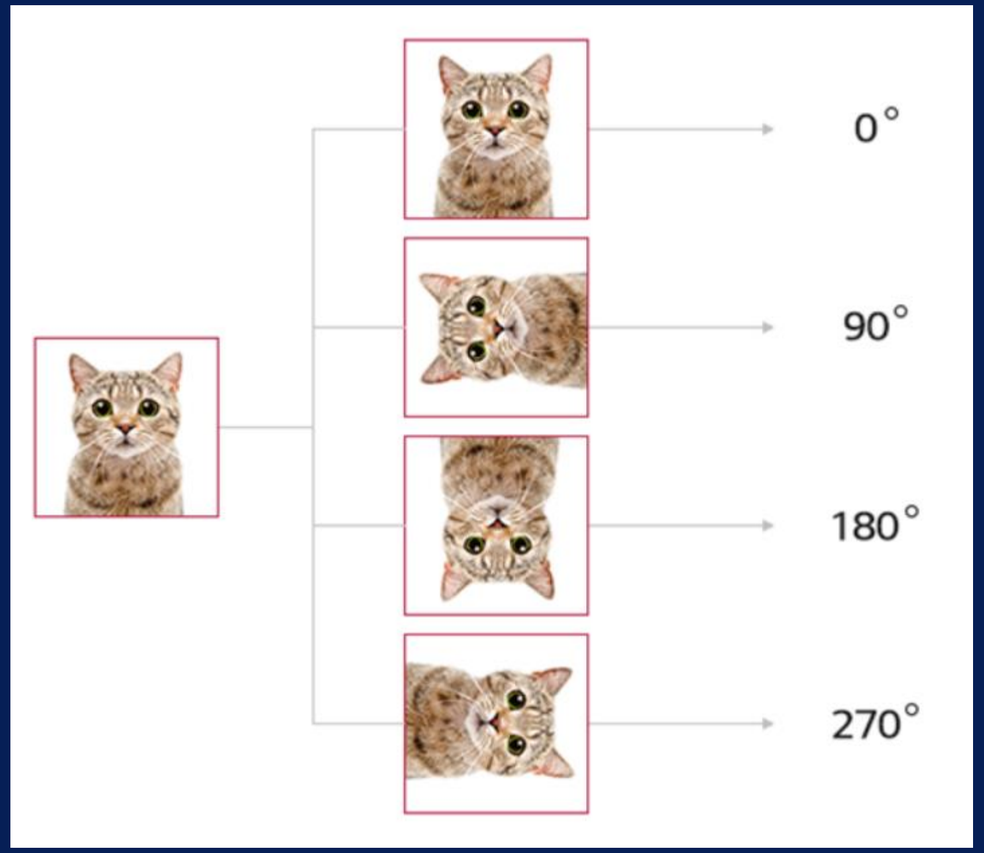


관련 연구

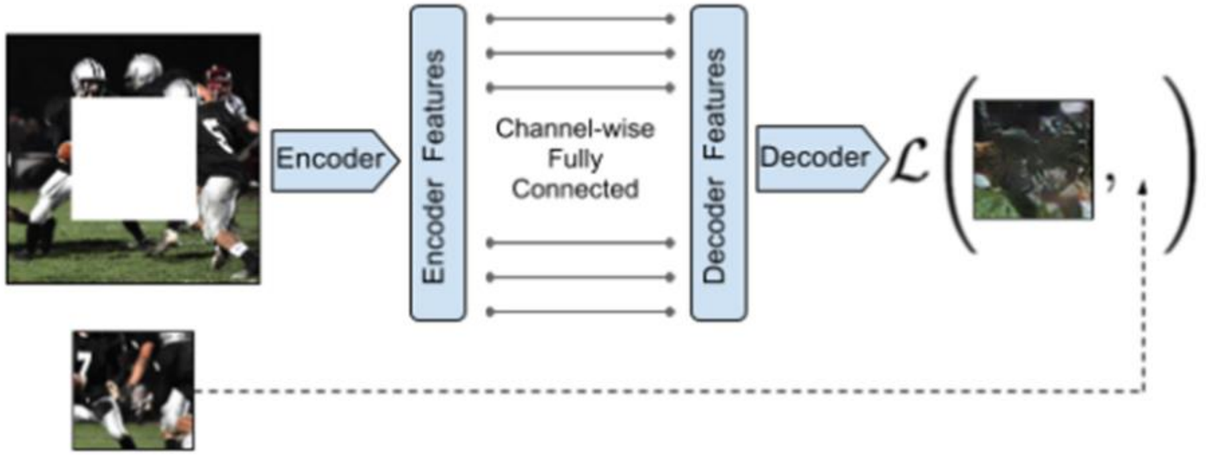
· 자기 지도학습 (Self-supervised learning)

비지도학습의 연구 주제 중 하나, 레이블이 없는 데이터를 이용해서 스스로 학습 후 분류하는 것.

' 사전 학습 ' 이 필요함. 이 사전 학습에 대한 과제를 pretext라고 함.



Context Encoders

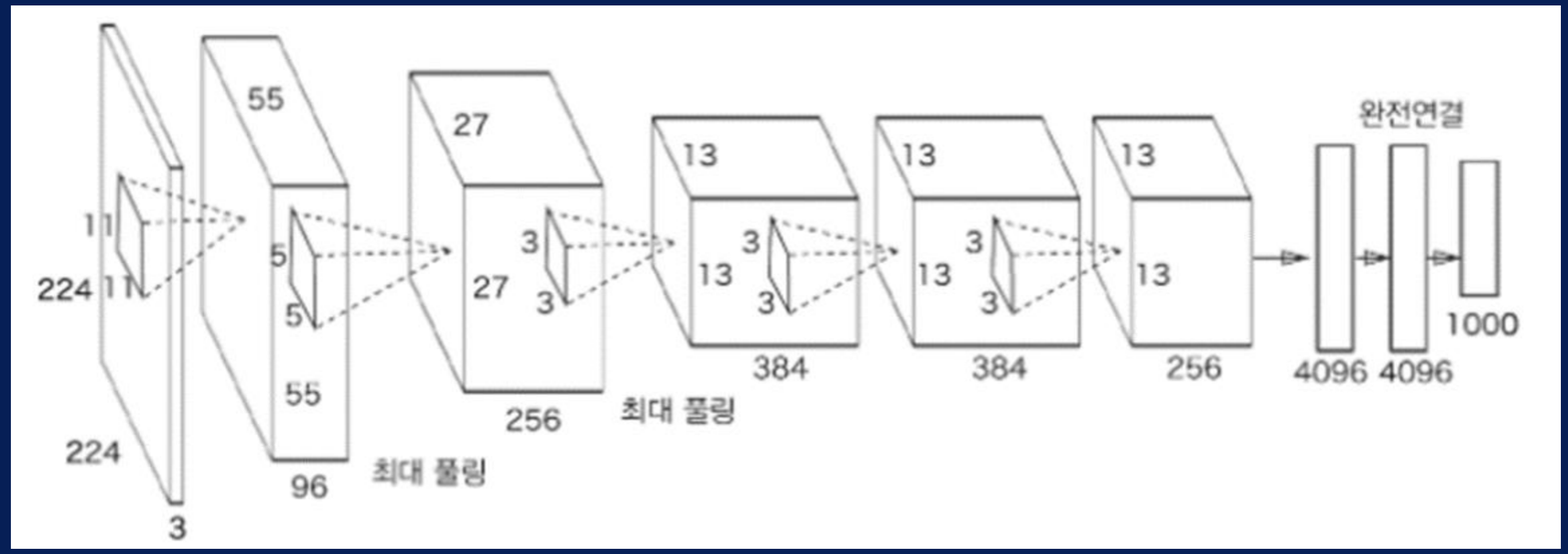


- **Context encoders for image generation**
  (컨텍스트 인코더를 이용한 이미지 생성)

Encoder는 누락된 영상으로부터 잠재 표현을 생성하고 Decoder는 누락된 이미지의 컨텐츠를 생성한다.

이 둘은 Fully Connected Layer로 되어 있고 이렇게 하였을 때 디코더가 전체 이미지 내용을 추론할 수 있다.

Context Encoders

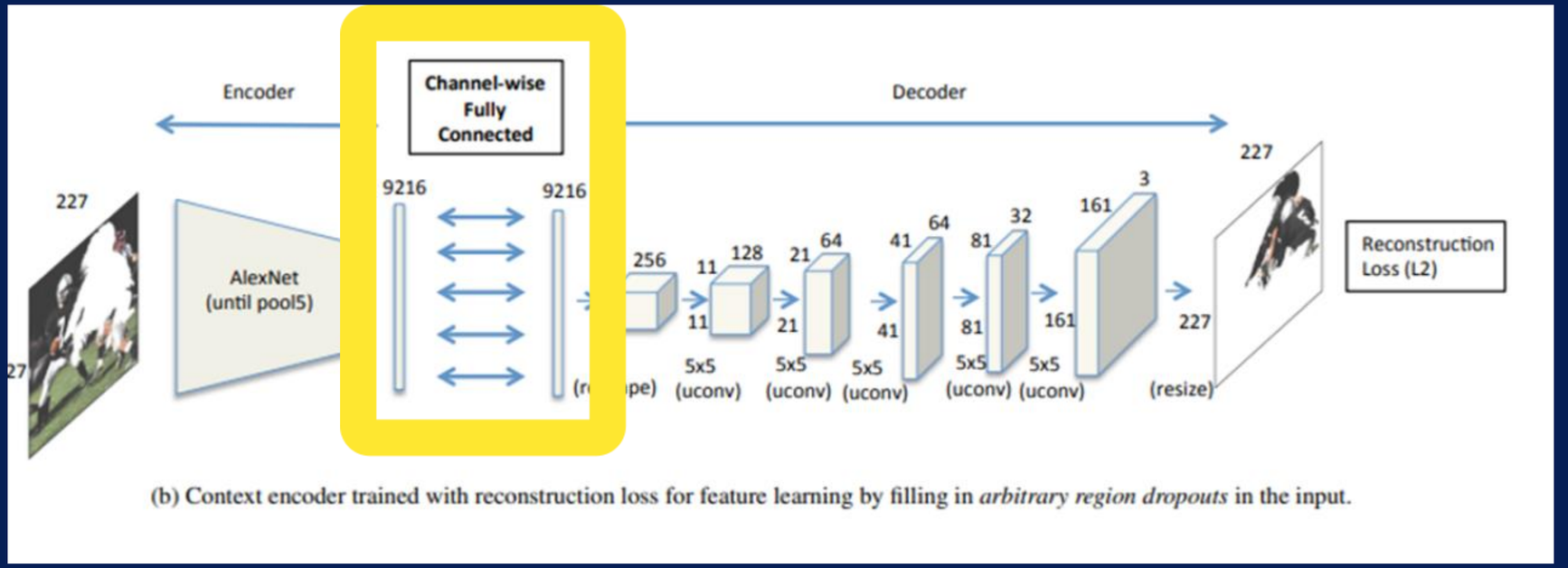


- **Encoder-decoder pipeline**

Encoder

Encoder는 없어진 부분을 갖고 있는 이미지를 latent vector(6x6x256)로 만든다. 구조는 AlexNet에서 파생되었다. AlexnNet구조와 같이 5개의 conv+pooling layer로 이루어진다. 단 여기서는 scratch로 표현될 수 있는 초기 random weights들로 context를 예측하는 것을 학습한다.

Context Encoders



(b) Context encoder trained with reconstruction loss for feature learning by filling in *arbitrary region dropouts* in the input.

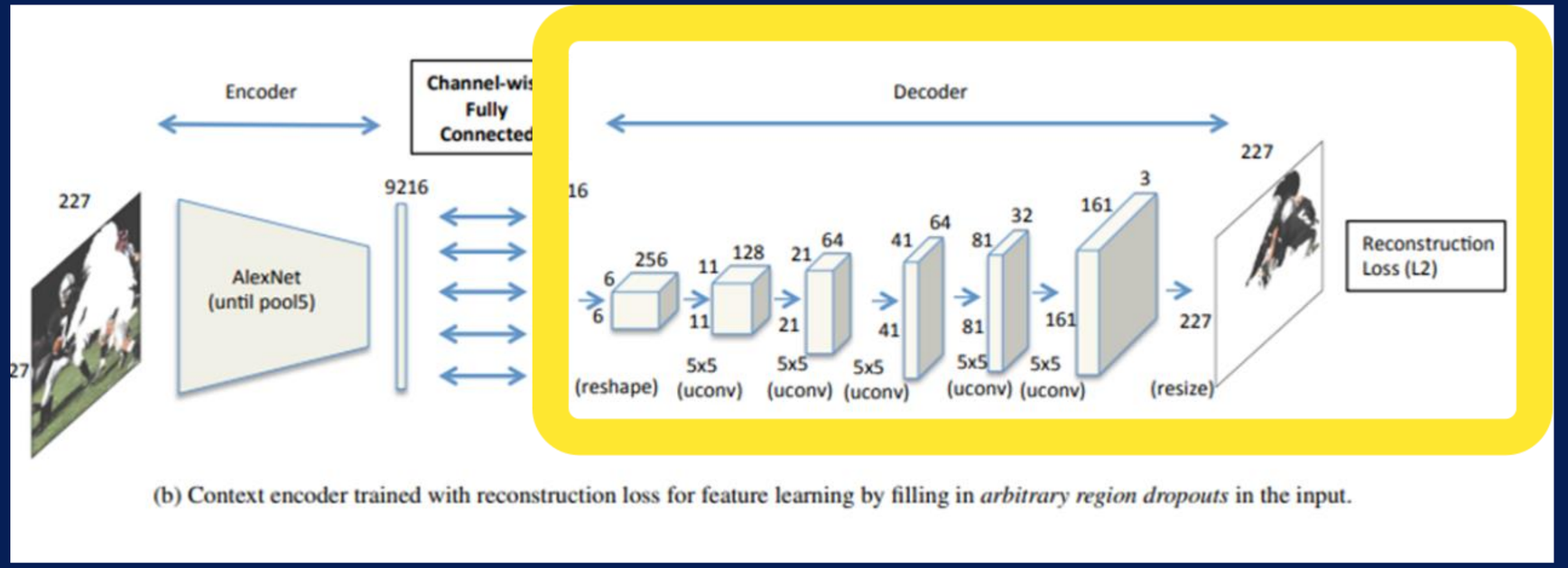
- **Encoder-decoder pipeline**

1. feture map(n*n*m) -> feature map(n*n*m)
2. Feature map마다 활성 함수 적용해서 전달
3. parameter 개수: mn4, 특징맵을 연결하는 paramete가 아님
4. 특징맵 내에서만 정보를 전파함

Channel-wise fully-connected layer

conv+pool 연산을 통해 나온 특징맵(feature map)은 말 그래도 특징만 추출했고 모든 위치를 직접 연결하지 않는다. 따라서 특징맵의 한 모서리에서 다른 모서리로 전파할 수 있는 방법이 없다. 이 모든 전파는 fully connected로 처리한다. 하지만 parameter를 고려해서 채널 방식의 fully connected layer를 구성하여 Decoder에 연결한다.

Context Encoders



(b) Context encoder trained with reconstruction loss for feature learning by filling in *arbitrary region dropouts* in the input.

- **Encoder-decoder pipeline**

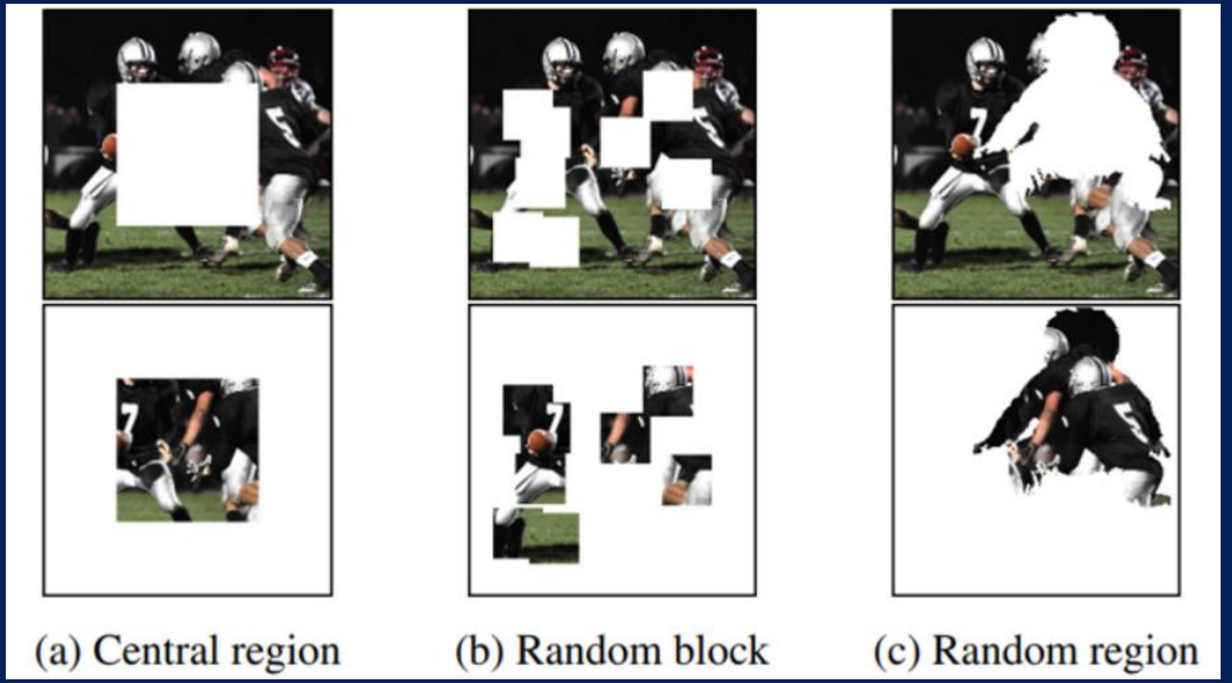
Decoder

Decoder는 전달받은 특징맵을 사용하여 이미지의 픽셀을 생성한다. RELU활성화를 거친 5개의 up-convolution la
이루어져 있다. up-convolution layer는 고해상도 이미지를 생성하는 역할을 한다.

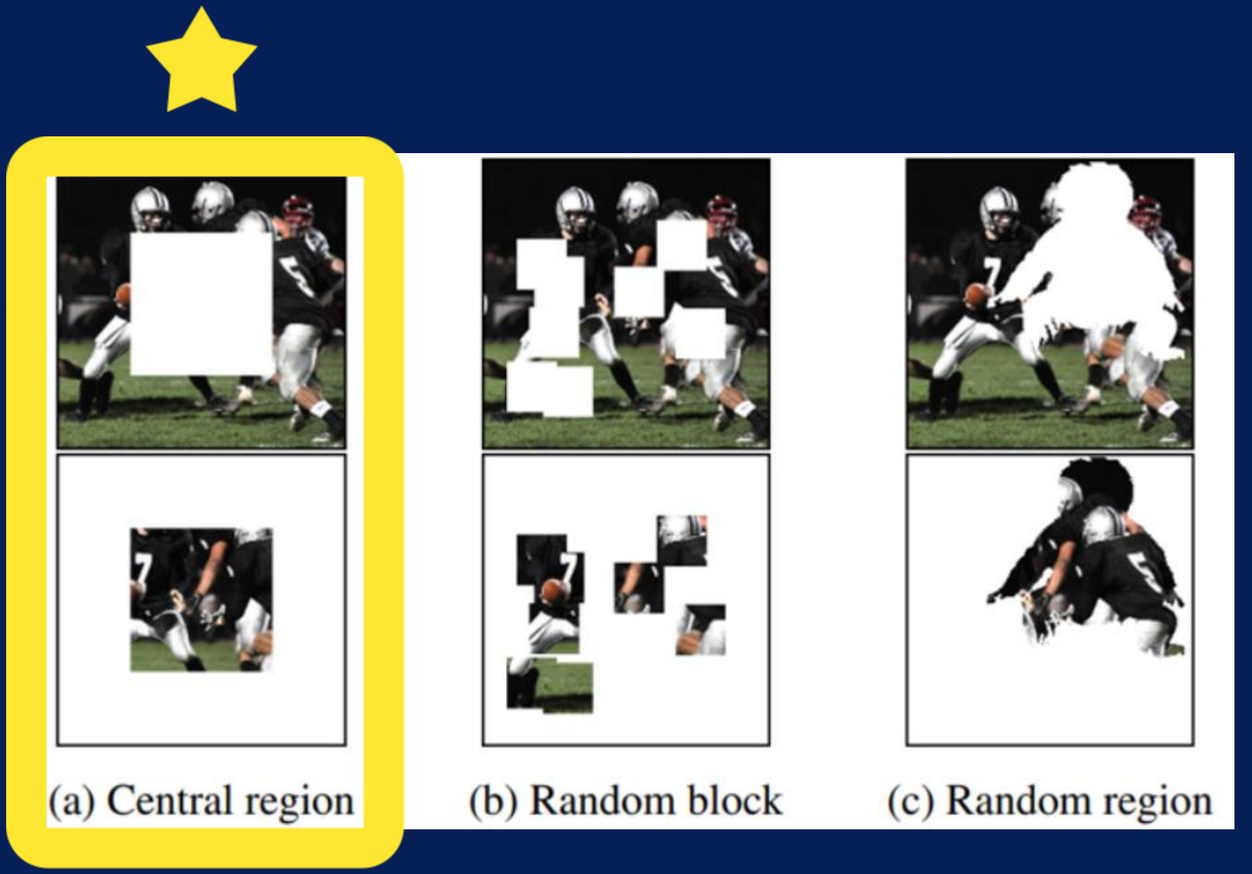
Context Encoder - Region masks

- 세 가지 전략

Context-Encoder의 입력 영상은 다음 중 하나일 수 있다. Paper는 아래의 세 가지 입력에 대한 전략을 제시한다.



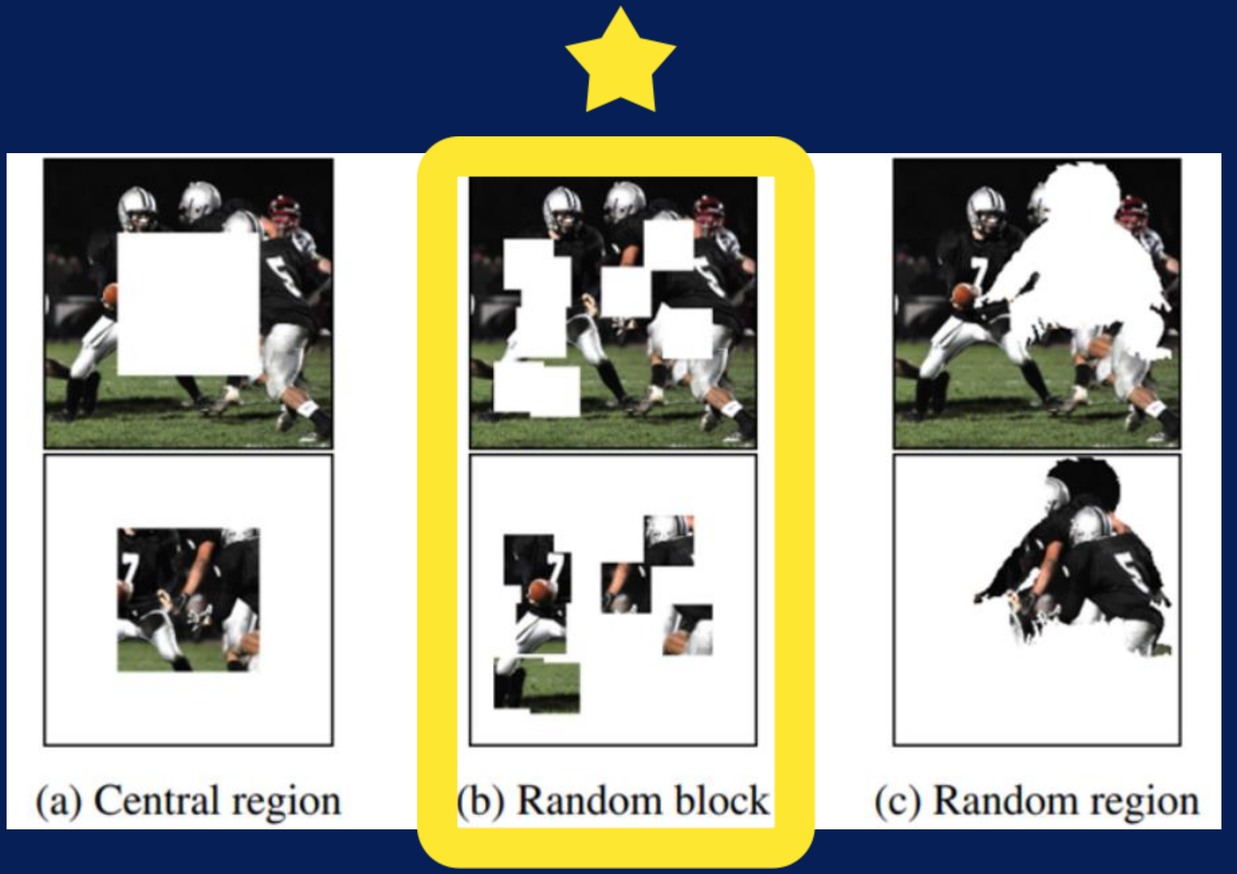
Context Encoder - Region masks



Central Region

- 가장 간단한 모양은 ‘중앙 사각형 패치’
- 인페인팅에는 잘 작동한다.
- central mask의 경계에 고정된 저수준의 특징을 학습한다.
- 제거된 영역에 대응되는 저수준의 이미지 특징을 찾는다.
- 이러한 낮은 수준의 feature는 마스크가 없는 이미지에는 잘 일반화되지 않는 경향이 있다.

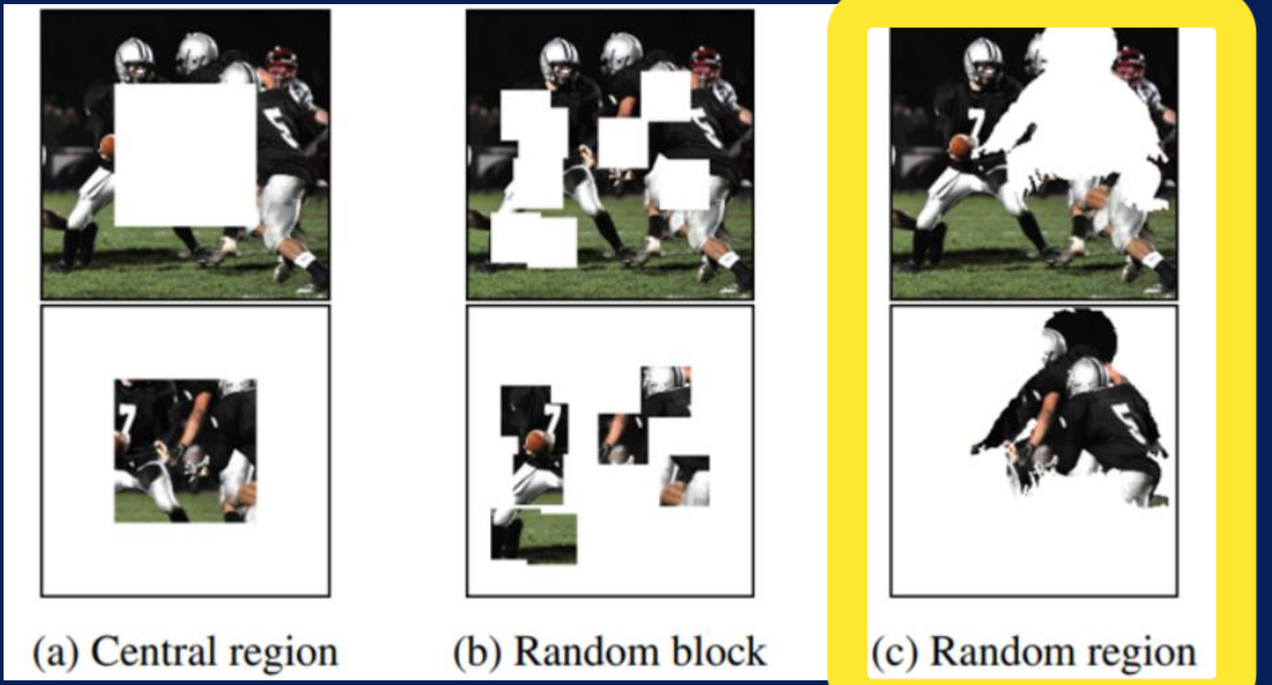
Context Encoder - Region masks



Random Block

- 영상의 최대 1/4을 차지하는 여러 겹치는 사각 마스크가 설정됨.
- random block masking은 sharp한 경계의 특징에 의해서 저수준의 특징밖에 찾지 못한다.

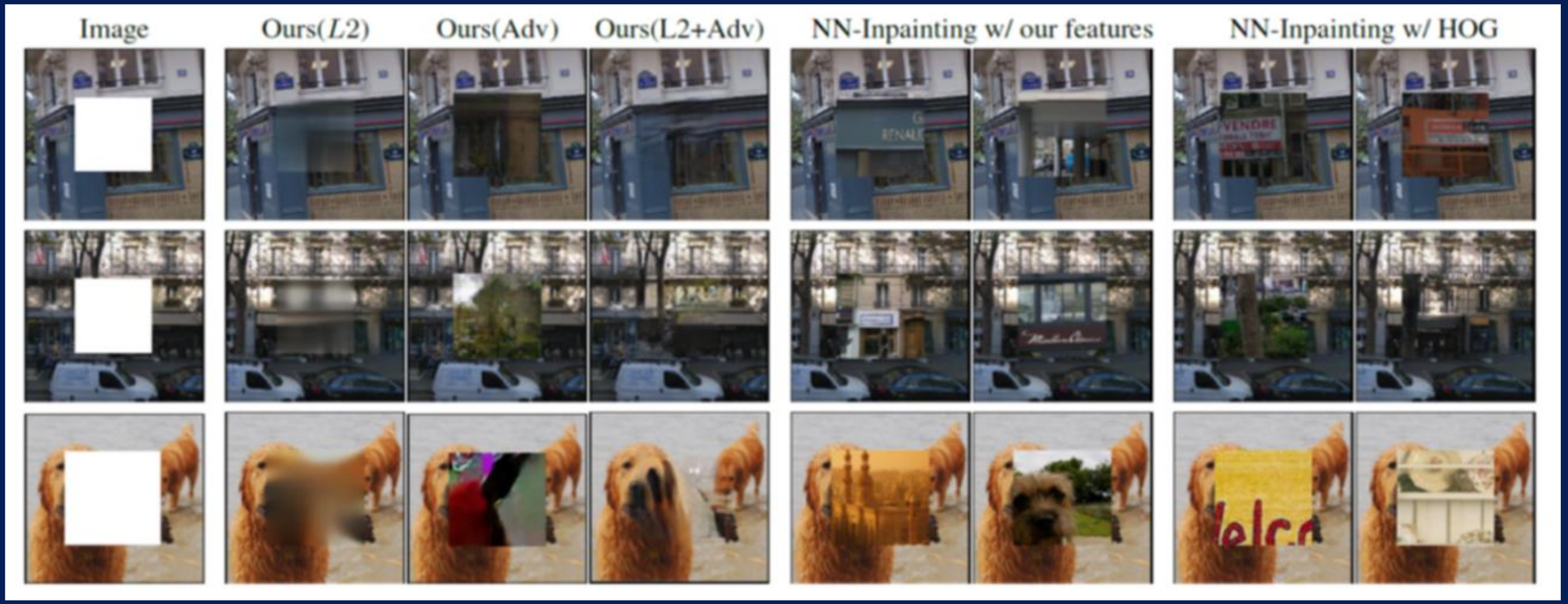
Context Encoder - Region masks



Random Region

- 이미지로부터 임의의 모양이 제거된 형태
- central region과 random block이 유사한 일반적인 특징을 찾아내는 것보다 훨씬 좋은 특징을 잘 찾는다.
- 따라서 random region dropout 은 특징을 찾는 용도로 많이 쓰인다.

결과



- **Semantic Inpainting**

Nearest neighbor inpainting(NN) 기법과 비교하였다.
CE를 통한 방식이 상황에 맞게 Reconstruction이 잘 된 모습을 보인다.

결과

Pretraining Method	Supervision	Pretraining time	Classification	Detection	Segmentation
ImageNet [26]	1000 class labels	3 days	78.2%	56.8%	48.0%
Random Gaussian	initialization	< 1 minute	53.3%	43.4%	19.8%
Autoencoder	-	14 hours	53.8%	41.9%	25.2%
Agrawal *et al.* [1]	egomotion	10 hours	52.9%	41.8%	-
Wang *et al.* [39]	motion	1 week	58.7%	47.4%	-
Doersch *et al.* [7]	relative context	4 weeks	55.3%	46.6%	-
Ours	context	14 hours	56.5%	44.5%	30.0%

- **Transferability**

Segmentation – Randomly initialized network보다 약 10% 성능이 우수하다.
오토인코더보다는 약 5% 성능이 우수하다.
Classification – self-supervised learning을 이용한 방법과 경쟁하는 수준이다.
Detection – 다른 방법들에 비해 크게 개선되지 않았다.

Context Encoders: Feature Learning by Inpainting

Context Encoders: Feature Learning by Inpainting

Deepak Pathak Philipp Krähenbühl Jeff Donahue Trevor Darrell Alexei A. Efros
University of California, Berkeley
{pathak,philkr,jdonahue,trevor,efros}@cs.berkeley.edu

Abstract

We present an unsupervised visual feature learning algorithm driven by context-based pixel prediction. By analogy with auto-encoders, we propose 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). When training context encoders, we have experimented with both a standard pixel-wise reconstruction loss, as well as a reconstruction plus an adversarial loss. The latter produces much sharper results because it can better handle multiple modes in the output. We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of our learned features for CNN pre-training on classification, detection, and segmentation tasks. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.

1. Introduction

Our visual world is very diverse, yet highly structured, and humans have an uncanny ability to make sense of this structure. In this work, we explore whether state-of-the-art computer vision algorithms can do the same. Consider the image shown in Figure 1a. Although the center part of the image is missing, most of us can easily imagine its content from the surrounding pixels, without having ever seen that exact scene. Some of us can even draw it, as shown on Figure 1b. This ability comes from the fact that natural images, despite their diversity, are highly structured (e.g. the regular pattern of windows on the facade). We humans are able to understand this structure and make visual predictions even when seeing only parts of the scene. In this paper, we show

The code, trained models and more inpainting results are available at the author's project website.

(a) Input context (b) Human artist

(c) Context Encoder (*L2* loss) (d) Context Encoder (*L2* + Adversarial loss)

Figure 1: Qualitative illustration of the task. Given an image with a missing region (a), a human artist has no trouble inpainting it (b). Automatic inpainting using our *context encoder* trained with *L2* reconstruction loss is shown in (c), and using both *L2* and adversarial losses in (d).

that it is possible to learn and predict this structure using convolutional neural networks (CNNs), a class of models that have recently shown success across a variety of image understanding tasks.

Given an image with a missing region (e.g., Fig. 1a), we train a convolutional neural network to regress to the missing pixel values (Fig. 1d). We call our model *context encoder*, as it consists of an encoder capturing the context of an image into a compact latent feature representation and a decoder which uses that representation to produce the missing image content. The context encoder is closely related to autoencoders [3, 20], as it shares a similar encoder-decoder architecture. Autoencoders take an input image and try

Context Encoders: Feature Learning by Inpainting

감사합니다.