😀

Stacked Hourglass Networks for Human Pose Estimation

▼ 간단 정리

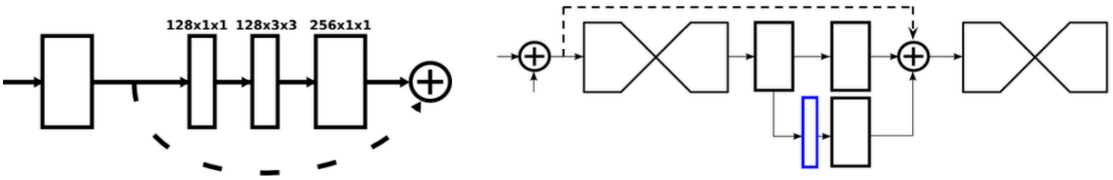


Fig. 4. Left: Residual Module [14] that we use throughout our network. **Right:** Illustration of the intermediate supervision process. The network splits and produces a set of heatmaps (outlined in blue) where a loss can be applied. A 1x1 convolution remaps the heatmaps to match the number of channels of the intermediate features. These are added together along with the features from the preceding hourglass.

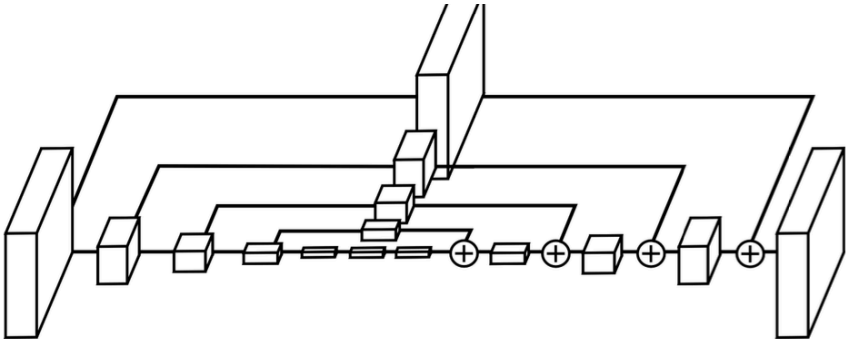


Fig. 3. An illustration of a single "hourglass" module. Each box in the figure corresponds to a residual module as seen in Figure 4. The number of features is consistent across the whole hourglass.

이 논문은 인간 자세 추정을 위해 Stacked Hourglass Network라는 새로운 신경망 구조를 제안합니다. 이 네트워크는 전체 이미지를 여러 스케일에서 처리하는 대칭적인 바텀업과 탑다운 경로를 통해, 고해상도에서 추출한 세부 정보와 저해상도에서 얻은 전역 정보를 효과적으로 결합하여 관절의 위치를 예측합니다. 각 hourglass 모듈은 이미지의 특징을 점차 낮은 해상도로 축소한 후, 다시 업샘플링을 통해 복원하며, 중간 단계에서 heatmap을 생성하고 손실을 적용하는 중간 감독 방식을 채택합니다. 이를 통해 초기 예측의 오차를 보완하고, 반복적인 평가와 정제를 통해 점차 정밀한 예측으로 발전

시킵니다. 또한, 잔차 모듈과 스킵 연결을 활용하여 정보의 효과적인 전달과 학습 안정성을 유지하며, 최종적으로 FLIC과 MPII Human Pose와 같은 표준 데이터셋에서 최첨단 성능을 달성하였습니다. 전체적으로 이 논문은 다중 스케일 정보의 융합과 반복적 추론을 통해 인간 자세 추정 문제의 한계를 극복하는 효과적인 방법론을 제시하고 있습니다.

결국에는 해당 모델은 FPN 모듈이 제안되기 전에 여러 스케일의 이미지에서 특징을 추출하기 위한 모델이였습니다. 단, 이미지의 업샘플링시 단순히 최근접 이웃 방식을 사용하는 등 2025년 에 봐서는 다소 투박한 모델임.

Abstract

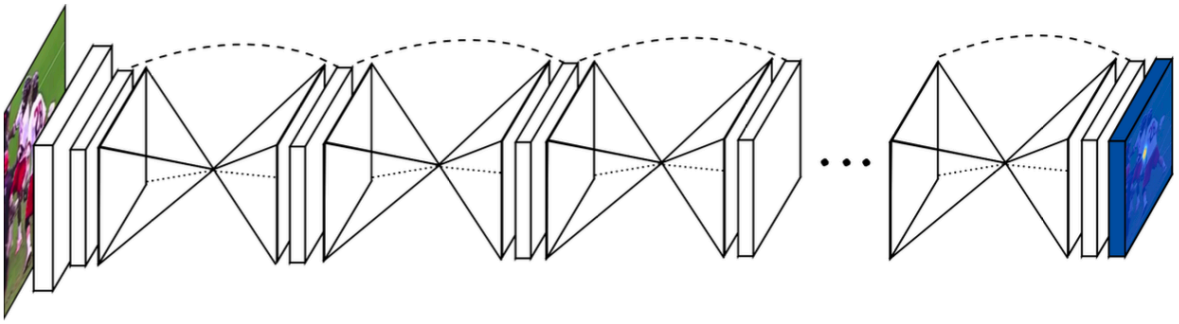


Fig. 1. Our network for pose estimation consists of multiple stacked hourglass modules which allow for repeated bottom-up, top-down inference.

본 논문에서는 인간 자세 추정을 위해, 이미지의 다양한 해상도에서 추출한 로컬 및 글로벌 특징들을 효과적으로 결합하는 새로운 네트워크 구조인 "Stacked Hourglass" 모델을 제안합니다. 이 모델은 입력 이미지를 점진적으로 다운샘플링하여 전역적 컨텍스트를 학습하고, 이후 대칭적인 업샘플링 과정을 통해 세밀한 공간 정보를 복원하며, 중간 감독을 적용하여 반복적으로 예측을 재정제합니다. 이를 통해 관절 간의 복잡한 관계를 정밀하게 파악하여, 최종적으로 높은 정확도의 관절 위치 예측을 가능하게 합니다.

Introduction

사람의 pose를 예측하는 문제는 가림 현상, 변형 등으로 연구자들에게 많은 여려움을 남겼습니다. 이전이의 연구자들은 이러한 문제를 해결하기 위해서 (1) 이미지 특징 (2) 구조적 예

측을 통해서 해결하였습니다. (1)을 통해서 지역적인 좌표값을 예측하고 (2)를 통해서 각 좌표 끼리의 관계성을 학습하였습니다.

하지만 이러한 복합적 구조들은 ConvNet들로 대체가 되기 시작되었고, ConvNet을 활용한 모델들은 이전의 모델들에 비해 드라마틱하게 큰 성능 향상을 보여주었습니다.

본 논문은 이러한 흐름을 따라 ConvNet을 활용하는 "Hourglass Network"를 제안하였고, 해당 모델은 모든 스케일의 이미지의 정보를 추출하고 통합한다고 주장합니다. 하지만 일반적인 ConvNet과 다르게 해당 모델의 경우 업샘플링을 통해서 모래시계 모양을 갖게 됩니다. 또한 여러 Hourglass Network를 순차적으로 연결구조와 중간 감독을 활용하여 높은 성능을 달성하였습니다.

Network Architecture

Hourglass Desing

본 논문에서는 작은 관절들 (예를들어 손가락 관절, 팔의 관절) 은 비교적 저수준 & 높은 해상도의 특징맵을 필요로 하고, 각 관절들 사이의 관계 혹은 자세에 대한 전체적인 패턴들을 파악하기 위해서는 고수준 & 저해상도의 이미지가 필요로 하다고 주장합니다. 이 모두를 학습하기 위해서 본 논문에서는 모든 스케일의 이미지를 학습하는 구조를 제안하고 있습니다.

모든 스케일의 이미지를 다루기 위해서 각 해상도의 이미지를 잘 통합하는 매커니즘을 가져야합니다. 하지만 본 논문에서는 서로 다른 스케일의 이미지를 독립적으로 처리하기 보다 하나의 파이프라인에서 skip-connection을 통한 단일 파이프라인을 제안합니다. 해당 모델의 경우 특징맵은 최소 4 * 4 크기까지 다운샘플링 되게 됩니다.

이미지가 최대 폴링에 도달하게 되면 최근접 이웃 방식을 통해서 특징맵을 업샘플링하게 됩니다 (예를들어 4 * 4 feature map → 8 * 8 feature map) 그리고 이전의 레이어에서 나온 feature map과 element wise하게 더해지게 됩니다 (skip connection). 이러한 방식을 통해서 이미지를 업샘플링 해주게 됩니다. 그리고 최종 출력 단계에서는 최종 feature map 에 1 * 1 conv 를 2개 거치며 다음 레이어로 넘어가며, Hourglass 모듈이후 1 * 1 conv거친 후 관절의 수와 동일한 차원을 갖는 Heat map을 생성하게 됩니다. Heat map을 통해서 Module의 중간중간에서 손실을 구할 수 있게 됩니다. 이러한 방식을 Intermediate supervision 이라고 합니다. 본 논문에서는 Intermediate supervision 방식을 활용하여 높은 수준의 정확도를 달성 할 수 있었다고 주장합니다.

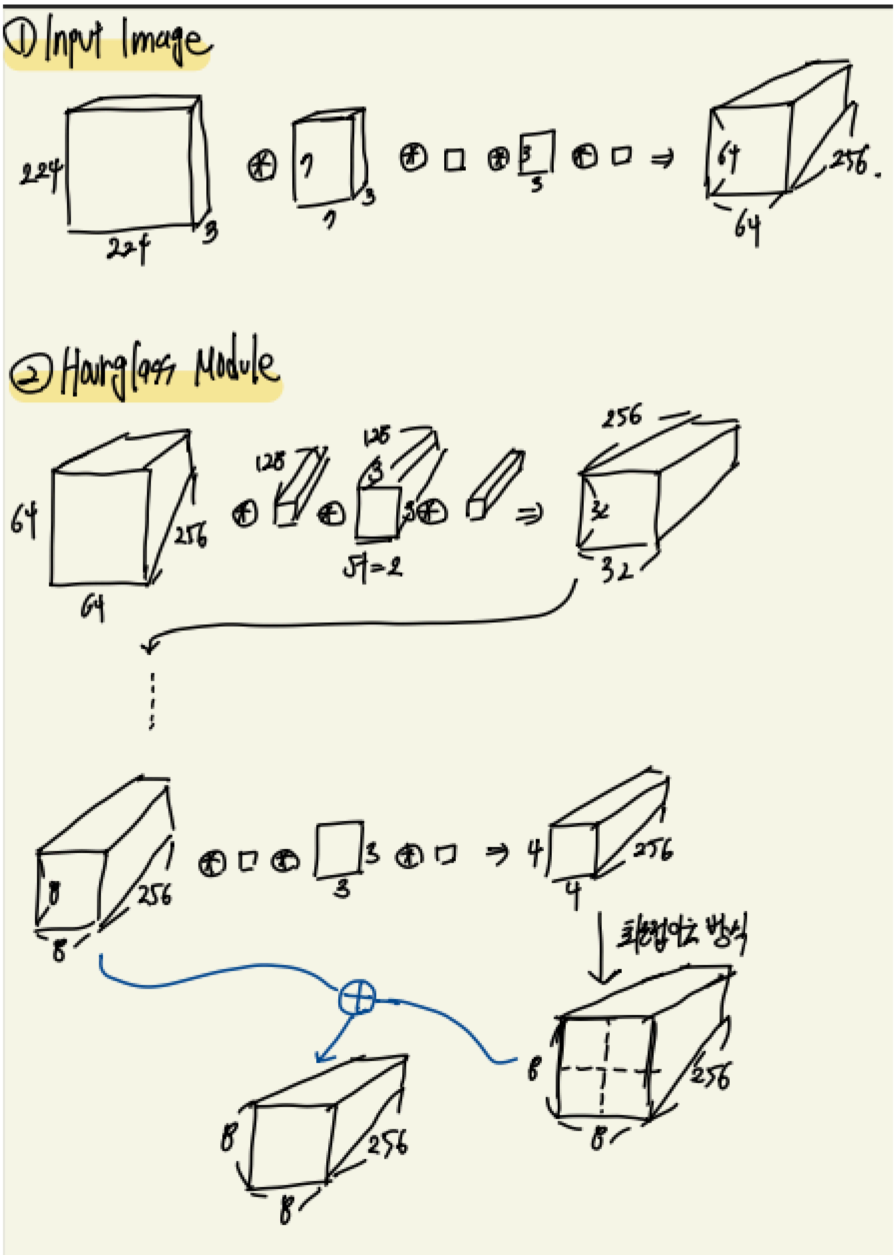
Layer Implementation

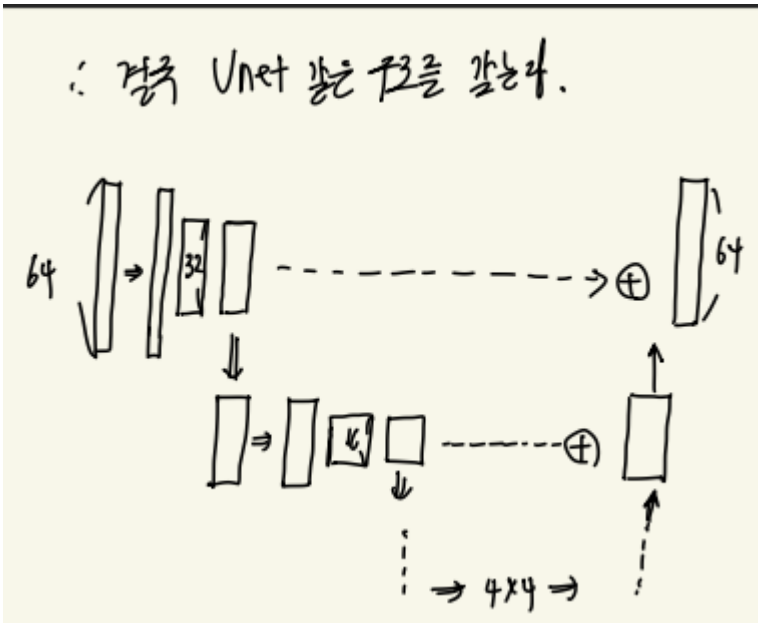
최근 1 * 1 conv와 Residual module의 효율성 연구가 진행되면서 본 논문에서도 이러한 기법들을 적용하였습니다. 최종적으로는 residual module을 사용하였고, 모든 filter size를 3 * 3 크기로 고정하였습니다. 추가로 당시 GPU 부족 문제로 224 * 224 이미지 원본을 그대로 활용하기 보다 7 * 7커널과 Residual Module을 거쳐서 64 * 64 크기의 이미지를 활용하였다고 합니다. 그래서 최종적으로 입력 이미지는 (64 X 64 X 256) 크기를 갖게 됩니다.

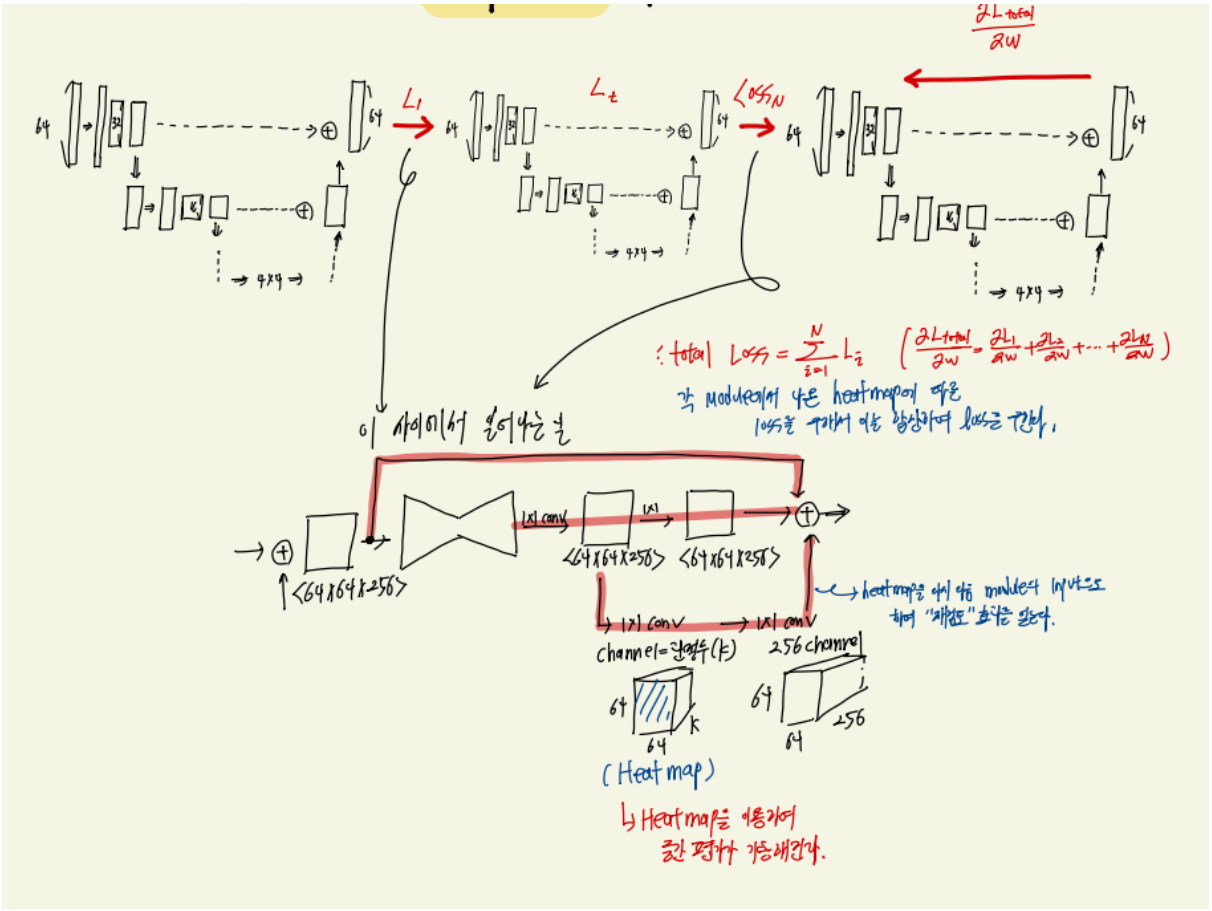
Stacked Hourglass with Intermediate Supervision

본 논문에서는 위에서 설명한 Hourglass module을 여러개 쌓는 구조를 활용하였습니다. 반복적인 구조를 통해 추론을 하게 되면 이전의 결과가 다음의 입력으로 들어가게 되며 추가로 이전의 결과를 재검토하는 효과를 얻을 수 있다고 설명합니다. 그리고 중간 마다 Heat map을 생성하여 Intermediate supervision을 통해 지역적 정보 그리고 글로벌 정보를 모두 학습한다고 주장합니다. 본 논문에서는 최종적으로 8개의 Hourglass 모듈을 사용하였다고 합니다.

직접 그린 파이프라인


① Input Image
② Hourglass Module


∴ 결국 Unet 같은 구조를 갖는다.
64
32
64
4x4


∂L total / ∂W
L1
Lt
Loss N
∴ total Loss = Σ Li
각 module에서 나온 heatmap의 작은 loss를 구해서 이를 합산하여 loss를 구한다.
이 사이에서 일어나는 일
1x1 conv
<64x64x256>
<64x64x256>
<64x64x256>
heatmap을 이용 다음 module의 input으로 하여 "재귀적" 효과를 얻는다.
channel=관절수(k)
256 channel
64
64
k
256
(Heat map)
↳ Heatmap을 이용하여 최종 평가 가능해진다.

간단 정리

Stacked Hourglass 논문은 Hourglass 형태의 다운-업 샘플링 과정을 통해서 여러 스케일의 이미지 특징을 한번에 처리하였습니다. 이는 이후 Feature pyramid와 비슷한 방식으로 FPN이 알려지기전 굉장히 효과적인 아이디어로 보입니다. 추가로 중간 손실을 사용하여 stacked 된 모듈마다 이를 재검토하는 방식을 통해서 다양한 특징과 패턴을 학습할 수 있었습니다.

나의 생각

중간 중간 손실 함수를 사용해서 감독을 통해서 초반에 방향성을 잡는다는 아이디어는 좋지만 역전파는 최종적으로 마지막 결과가 도출될 때 흘러 들어간다는 것이 다소 효과적이지 않다고 생각을 한다. 이에 각 모듈이 끝날때 마다 손실을 구해서 바로 역전파를 통해서 결과가 나오고 & 역전파가 흐르고 이러한 파이프라인을 사용하게 되면 조금더 효과적일것이라고 생각이든다.

그리고 이미지 사이즈를 224 * 224 사이즈에서 64 * 64로 줄였는데, 성능 차이가 거의 없었다는 것은 이해하기는 조금 어려운것 같다. 만일 GPU가 주어진다면 한번 고행상도 이미지를 통해 실험을 진행해보고싶다.