Light-Weight Facial Landmark Prediction Challenge

Yun-Xiang Yang
雲翔楊
B10730022
Department of Computer Science & Information Engineering
b10730022@mail.ntust.edu.tw

Ming-Sheng Huang
明勝黃
B10730026
Department of Electrical Engineering
b10730026@mail.ntust.edu.tw

Ruei-Fu Lee
睿莆李
R10521802
Department of Civil Engineering
r10521802@g.ntu.edu.tw

https://github.com/EonianCoda/2d_landmark_detection/tree/main

1 Methodology and Model Architecture

我們參考了[2]，使用Stacked Hourglass Network[1]來解決facial landmark detection任務。如Fig 1，該模型的特色是利用重複地放大、縮小特徵圖的方式來提取不同空間尺度的特徵，並且透過輸出heatmap的方式來獲取臉部特徵位置。此外，為了符合輕量化的限制，我們減少了模型中hourglass network的數量、深度與特徵維度，詳細的實驗與比較於Sec.4 中。

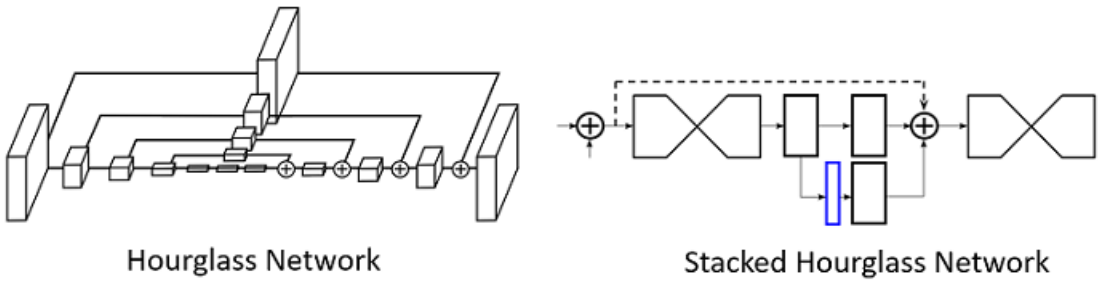

Figure 1: Model Architecture，左圖是一個深度為4的hourglass network，而右圖則是整個模型的架構圖，圖中沙漏狀的部件代表的便是hourglass network。

1.1 Attention Block

為了更進一步提升模型的能力，我們加入了attention block，我們測試了兩種著名的attention block：Squeeze and excitation networks(SE layer)[5]及Coordinate Attention block(CA layer)[6]，最後選定採用CA layer。加入此改動後，可以使模型更加專注重要的特徵，並省略不重要的特徵。此外，由於attention block可能會導致某些特徵被丟棄，因此，我們選擇在hourglass network中upsampling layer前加入attention block(Fig 2)，在經過upsampling layer後，將會進行一次特徵融合，這樣不但可以使重要的特徵得到增強，且不重要的特徵也不至於被丟棄。

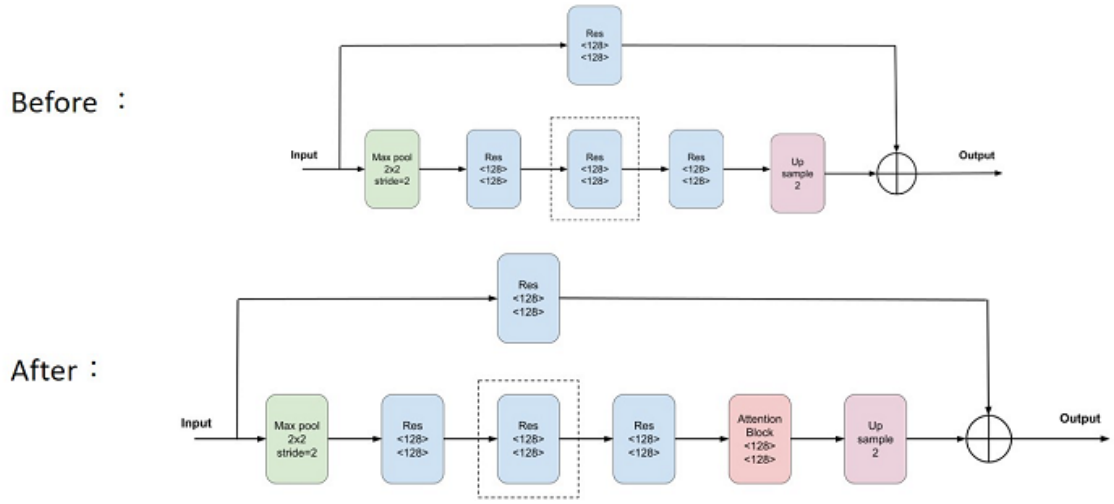

Figure 2: 加入Attention Block，上圖為原始的hourglass network，下圖則是加入attention block後的hourglass network

1.2 Coordinate convolution

此外，我們參考了[4]，將模型的第一層卷積層，以coordinate convolution layer(CoordConv)[6](Fig 3)代替，加入這個修改後，模型的性能有著大幅的提升，我們的猜想是因為臉部特徵辨識任務與空間資訊具有高度相關，因此加入coordinate convolution後，可以使模型能夠更好地對空間資訊進行處理。

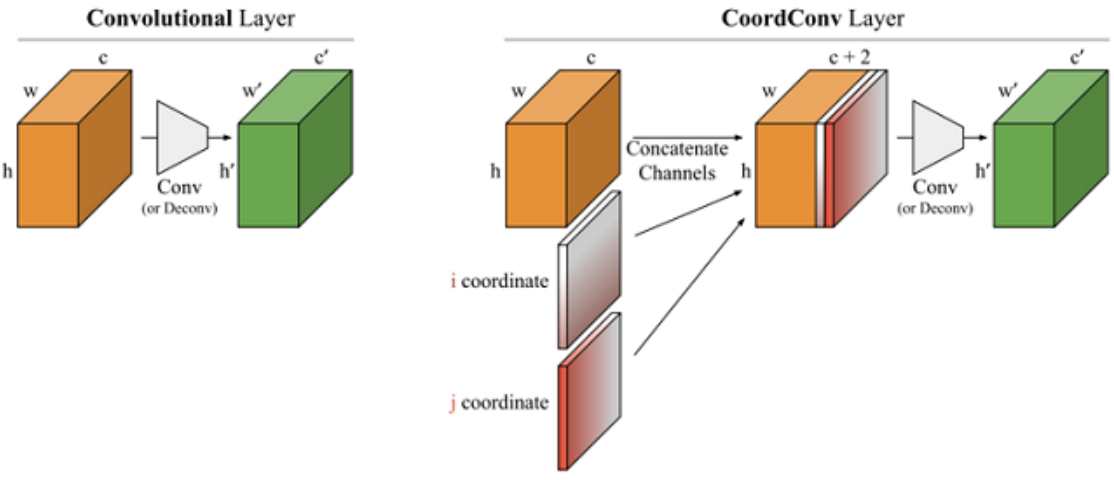

Figure 3: CoordConv

1.3 Loss function

對於heatmap regression問題，常使用Mean squared error(MSE)，然而對於我們這個任務來說，直接採用MSE loss會產生一個問題，即背景像素點過多，會過度關注背景。為了解決這個問題，我們針對每個像素點進行加權，為前景與靠近前景之像素給定較高的權重，防止過度關注背景像素，具體的作法為先針對原始的heatmap H進行dilation運算，再將其結果作為weight map M ，$\bigotimes$表element-wise運算，為原始loss進行加權

$$M = \begin{cases} 1 & \text{where } H^d \geq 0.2 \\ 0 & \text{otherwise} \end{cases}$$

$$MSE_{weighted}(H, \hat{H}) = MSE(H, \hat{H}) \bigotimes (W \cdot M + 1)$$

2 Implementation Details

2.1 Landmark to Heatmap

為了訓練模型，我們需要先將原始68個landmark座標轉為68張heatmap，由於輸出的heatmap的尺寸僅有原圖的1/4，首先要做的便是對原始座標進行轉換，我們採用的方案是直接對座標除以4，並進行四捨五入，接著再根據計算出來的座標點為中心，以$7 \times 7, \sigma = 1.75$的gaussian distribution填入對應的值。此外，為了避免heatmap轉換回座標時產生的精度誤差，我們採用了經過處理的gaussian distribution，我們稱之為Gaussian distribution with offset，概念是透過轉換後的座標與原始座標之間的偏移量對gaussiand distribution進行加權，步驟如下：

1. 根據偏移量計算ratio kernel
2. 根據ratio kernel對原始gaussian kernel進行卷積運算
3. 進行normalize，將gaussian kernel中心數值定義為1

2.2 Augmentation

我們採用了6種不同的Augmentation：horizontal flip, rotation(-30°, 30°), colorjitter, padding, gaussian blur, erasing，其中rotation與colorjitter對我們實驗的影響最大。

2.3 Hyperparameter Choices

在scheduler方面採用了RMSprop, learning rate為1e-4，前2k step為warm up，在80k step時調整為5e-5，120k step時調整為2e-5，共訓練20epoch。此外，加入了L2 weight decay，比例為1e-6。

3 Experiments

我們對於Model Architecture的實驗數據如Table 2，可以看見在Number of Stack = 2, Depth of Hourglass Network = 4, Number of Channels = 128 時，所得到的Test NME Loss 最佳。

Number of Stack	Depth of Hourglass Network	Number of Channels	Model Size	Test NME Loss
1	2	256	14.0	2.7770
1	4	128	6.12	2.4217
2	4	128	11.8	2.2929
2	5	128	14.2	2.3211
3	3	128	14.0	2.6709
4	2	128	13.8	2.7012

Table 1: Experiment for basic structure

在此基礎上加入Attentation block 並觀察其Test NME Loss(Tabel 2)，可以看見使用了CA Layer和Coordinate Conv的效果最為顯著。

Extra srtucture	Model Size	Test NME Loss
None	11.8	2.2929
+ SE layer	12.1	2.2347
+ CA layer	12.2	2.3044
+ Coordinate Convolution	11.8	2.2121
+ SE layer and Coordinate Convolution	12.1	2.2211
+ CA layer and Coordinate Convolution	12.2	2.1242

Table 2: Ablation Study for extra structure

在我們進行Landmark的Heatmap轉換時，由於精度問題，會有些許偏移，這些偏移會造成我們的預測產生巨大的誤差，所以我們計算了並導正了誤差，Table 3左半邊為使用Ground Truth直接轉換後測試的NME Loss可以很明顯的看出使用前後的Test NME Loss 相差巨大，右半邊為同一批資料進行預測的NME Loss，使用前後也有明顯提升。

With Ground Truth		With Prediction	
Use kernel with offset	Test NME Loss	Use kernel with offset	Test NME Loss
	0.5127		2.3382
◯	0.0076	◯	2.2929

Table 3: Gaussian distribution with offset

我們也觀察到原資料集以及現實生活中其實會有很多時候臉會被遮擋住，所以我們測試了在臉上新增了一些遮擋，可以看到在新增遮擋後(Table 5)，可以更好的預測到那些被遮擋的部位(Fig 4)。

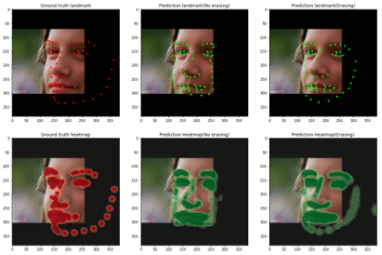

Figure 4: Visualization：Ground truth(左)，No Erasing(中)，Add Erasing(右)

Add erasing	Test NME Loss
	2.0481
◯	2.0219

Figure 5: Erasing Comparison

最後是關於Loss，如Table 4在使用了Weighted Loss之後，對於NMELoss也有顯著提升。

Loss	Test NME Loss
MSE	2.2929
Weighted MSE	2.2037

Table 4: weighted L2

4 Conclusion

- 透過加入attention block及coordinate convolution，在僅提升模型大小0.3 % 的情況下，大幅提升模型的性能
- 提出了gaussian kernel with offset，用以改善進行座標轉換時會產生偏差的現象
- 發現Erasing Augumentation可以有效改善人臉的遮擋問題

References

[1] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In ECCV, 2016.

[2] Bulat, Adrian, and Georgios Tzimiropoulos. "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[3] Feng, Zhen-Hua, et al. "Wing loss for robust facial landmark localisation with convolutional neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[4] Wang, Xinyao, Liefeng Bo, and Li Fuxin. "Adaptive wing loss for robust face alignment via heatmap regression." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[5] Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. CVPR, 2018.

[6] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. arXiv preprint arXiv:2103.02907, 2021.

[7] Liu, Rosanne, et al. "An intriguing failing of convolutional neural networks and the coordconv solution." Advances in neural information processing systems 31 (2018).

[8] Guo, Xiaojie, et al. "PFLD: A practical facial landmark detector." arXiv preprint arXiv:1902.10859 (2019).

[9] Wu, Wayne, et al. "Look at boundary: A boundary-aware face alignment algorithm." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.