UNIVERSITY of HOUSTON

Deep Imbalanced Attribute Classification using Visual Attention Aggregation

Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris

Computational Biomedicine Lab, University of Houston

ECCV 2018 European Conference on Computer Vision
8 – 14 September 2018 | Munich, Germany

Introduction

Problem Statement: Recognize the visual attributes of humans in images

Desirable Characteristics of the Solution

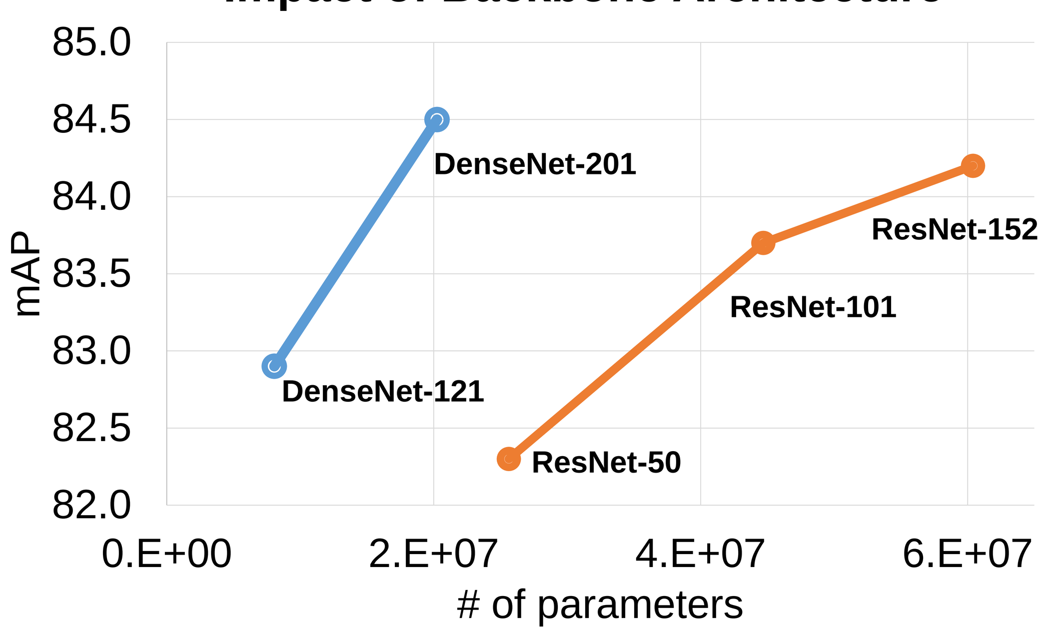
- Account for class imbalance during learning
- Account for the large prediction variance originating from the attention masks
- Keep the architecture as simple as possible

Contributions

- A weighted-variant of the focal loss that handles class imbalance at a class and at an instance level
- A loss that penalizes attribute predictions with high prediction variance in a weakly-supervised setup

Ablation Studies on WIDER

Impact of Backbone Architecture



Impact of individual proposed components

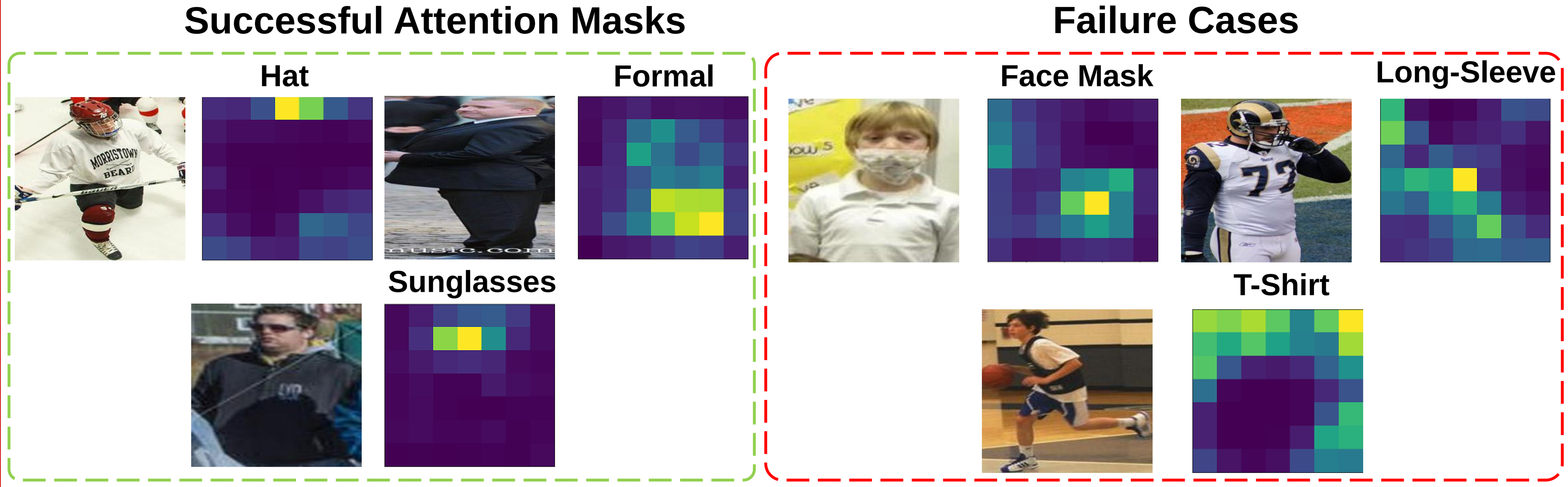
Primary Network	L_w	Attention	L_a	Multi-scale	mAP
ResNet-101					83.7
ResNet-101	✔				84.4
ResNet-101	✔	✔			85.0
ResNet-101	✔	✔	✔		85.7
ResNet-101	✔	✔		✔	85.9
ResNet-101	✔	✔	✔	✔	86.4

Quantitative Results

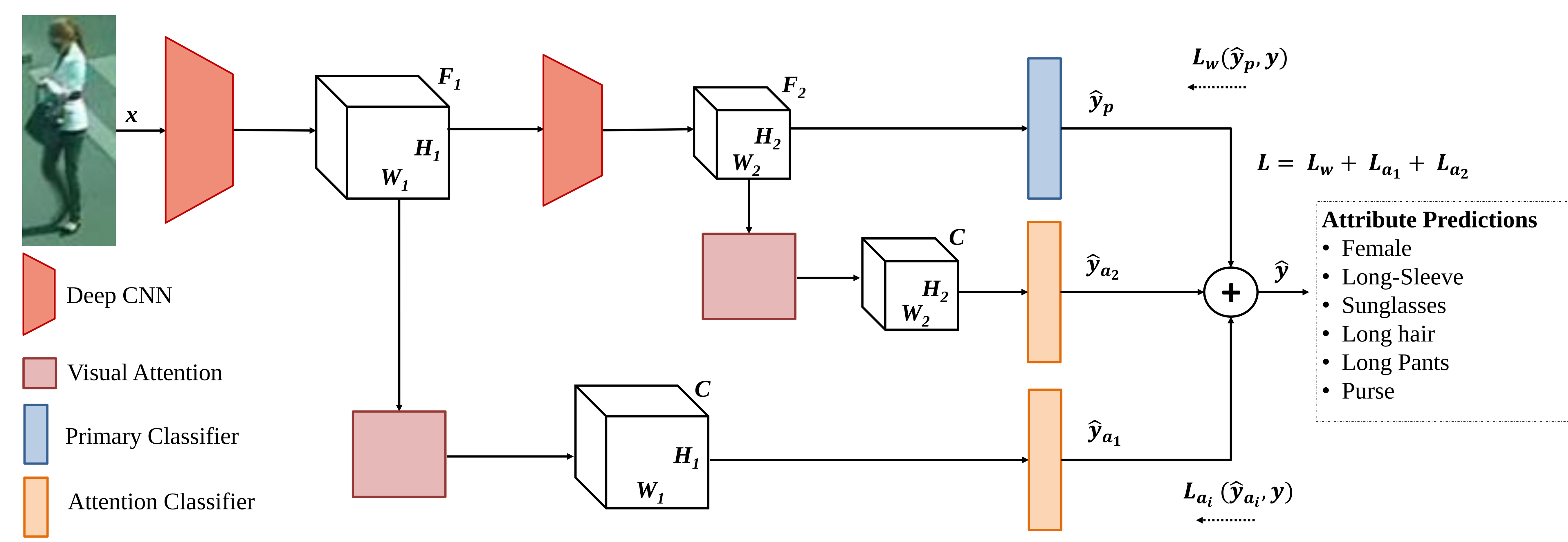
Method	Male	Long hair	Sunglasses	Hat	T-shirt	Long sleeve	Formal	Shorts	Jeans	Long Pants	Skirt	Face Mask	Logo	Plaid	mAP
Imbalance Ratio	1	3	18	3	4	1	13	6	11	2	9	28	3	18	
RCNN	94	81	60	91	76	94	78	89	68	96	80	72	87	55	80.0
R*CNN	94	82	62	91	76	95	79	89	68	96	80	73	87	56	80.5
DHC	94	82	64	92	78	95	80	90	69	96	81	76	88	55	81.3
VeSPA	-	-	-	-	-	-	-	-	-	-	-	-	-	-	82.4
CAM	95	85	71	**94**	78	96	81	89	75	96	81	73	88	60	82.9
ResNet-101	94	85	69	91	80	96	83	91	78	95	82	74	89	65	83.7
ResNet-101 + MTL + CRL	94	86	71	91	81	96	83	92	79	96	84	76	90	66	84.7
SRN	95	87	70	92	82	95	84	92	80	96	84	76	90	66	85.0
Ours	**96**	**88**	**74**	93	**83**	**97**	**85**	**93**	**81**	**97**	**85**	**78**	**90**	**68**	**86.4**

Qualitative Results

Successful Attention Masks | **Failure Cases**



Method



Weighted Focal Loss: Handles class imbalance at a class and at an instance-level

$$L_w = -w_c \sum_{c=1}^{C} \left[\left(1 - \sigma(\hat{y}_p^c)\right)^{\gamma} \log \sigma(\hat{y}_p^c)\ y^c + \sigma(\hat{y}_p^c)^{\gamma} \log\left(1 - \sigma(\hat{y}_p^c)\right)(1 - y^c) \right]$$

Attention Loss: Accounts for the weak supervision of the attention heatmaps and penalizes samples with high prediction variance:

1. Collect history (H) of predictions $p_H(y_s|x_s)$ for sample x_s and compute its standard deviation:

$$\widehat{std}_s(H) = \sqrt{\widehat{var}\left(p_{H^{t-1}}(y_s|x_s)\right) + \frac{\widehat{var}\left(p_{H^{t-1}}(y_s|x_s)\right)^2}{|H_s^{t-1}| - 1}}$$

2. Compute the loss for the predictions originating from the attention masks:

$$L_{a_i}(\hat{y}_{a_i}, y) = \left(1 + \widehat{std}(H)\right) L_b(\hat{y}_{a_i}, y)$$

3. Compute total loss for the primary network and the attention modules: $L = L_w + \sum_{i=1}^{M} L_{a_i}$

Sources of Error

- Resizing rectangular images of pedestrians to a fixed square-size resolution distorts the original image
- Several images in the PETA dataset have very low resolution which complicates attribute recognition
- The annotations contain a third unspecified/uncertain class, which is used as negative during training in the literature, that dilutes the learning process

Key Takeaways

- Handling class imbalance and focusing on hard misclassified positive samples can improve the performance
- Penalizing samples with high prediction variance can be beneficial for weakly-supervised applications

UNIVERSITY of HOUSTON | CBL

Changing the way ~~people~~ look at ~~computers~~
computers people

Acknowledgements: This work has been funded in part by the UH Hugh Roy and Lillie Cranz Cullen Endowment Fund