BURNED: Efficient and Accurate Burn Prognosis Using Deep Learning

Orry Despo, CERC, UAR

Motivation

Early excision of burns saves peoples live. Like most treatments, this relies on having accurate early-stage burn depth diagnosis. Unfortunately, besides the experts, there exists a lack of accuracy among the burn community in terms of early diagnosis. Can we create an automated visual system which detects burn severity and spatial outline, and thus scale expert level care to millions of burn victims worldwide?

Problem Definition

Given | Predict

Given a raw digital image of the burned area, predict the burn severity of each pixel (represented by different colors).

Dataset

- A novel dataset called BURNED was created for this task. 650 images of pre-48 hour burns were obtained form Valley Medical Center. They were combined with 200 images manually curated from Google.
- 6 plastic surgeons from Stanford collaborated to segment and label these images using an adapted annotation tool[1]. Each image was labeled by three different plastic surgeons.

Figure 2: The adapted annotation tool's interface for a labeler. The mask (shaded in blue) has already been created by another plastic surgeon.

- This is the largest dataset of its kind.

	Superficial	Partial Thickness	Full Thickness	Un-debrided
Pixels	660K	3.8M	1.1M	313K
Images	98	564	163	86

Figure 3: Breakdown of the number of pixels and images corresponding to each burn depth category. Only a subset of the 850 images was used for the analysis.

Technical Approach

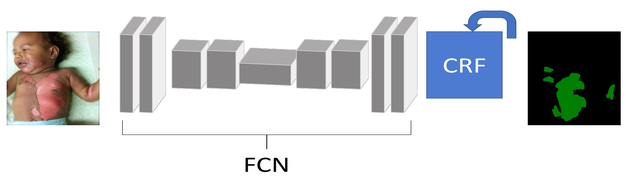


Figure 4: The model takes in the raw 2d image which is then passed through a fully convolutional network[2]. The specific network is FCN-8, which is a reformulation of VGG-16. Instead of a final fully connected layer, the network upsamples the downsampled prediction back to the original input size. A conditional random field formulated as a RNN layer is attached at the end, which acts to smooth the pixel label assignments[3]. The whole network is end-to-end trainable.

- We used the pre-trained FCN-8 with CRF attached. This was pre-trained using the Pascal VOC 2012 dataset[3]. We then fine-tuned using an overall learning rate of $5e^{-}5$.
- Each image was resized to 250x250 and training was done using the recommended batch-size of 1[3].

Metrics

Pixel Accuracy (PA)

$$\frac{1}{n_{cl}} \sum_{i=1}^{n_{cl}} \frac{n_{ii}}{t_i}$$

Mean IOU

$$\frac{1}{n_{cl}} \sum_{i=1}^{n_{cl}} \frac{n_{ii}}{t_i + \sum_{j=1}^{n_{cl}} n_{ji} - n_{ii}}$$

Figure 5: n_{cl} represents the number of classes (not including the background class). t_i represents the total number of pixels for class i. n_{ij} is the number of pixels of class i predicted to be class j. These are common metrics for semantic segmentation[2]. The combination of these two methods allow us to quantify whether we over or under predict a certain class.

Burn/No Burn

- Attempting to discriminate burned skin from the rest of the image, we achieved a mean IOU of .67 and accuracy of .85, which is a great job considering the small dataset. The 20% boost in IOU from data augmentation indicates the need for a more extensive dataset to help us from over-predicting.
- The three most common errors are: predicting slightly over the boundary, picking up on less severe burns not segmented, and struggling to differentiate non-clear skin from burnt skin.

	PA	IOU
FCN - No CRF	.82	.54
FCN - CRF	.85	.56
FCN - CRF, Aug	.85	.67
Pascals	N/A	.75

Figure 6: We see the breakdown in metrics for the test set. The FCN with CRF layer and data augmentation clearly does the best. The pascals represents the models performance on the PASCAL VOC 2012 dataset, one of the gold standards for semantic segmentation[3].

Truth | Predicted

Figure 7: An example of us slightly over-predicting the bounds. Longer training and more data should help this case.

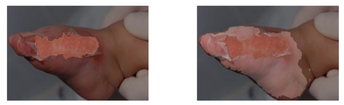


Figure 8: The algorithm seemingly over-predicts, but actually catches less severe burns which were not labeled. More refined dataset collection techniques are needed here or weak/semi-supervised methods.

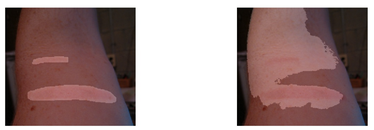


Figure 9: The algorithm struggles to differentiate between burnt skin and non-clear skin (slightly red and freckled). Adding in a dermatology corpus of various skin conditions is needed.

Multi Burn

- Extending to predicting the 4 burn depths causes a substantial drop in both metrics. The main culprit appears to be the substantial class imbalance seen in Figure 2.
- Methods to combat this were to upsample by augmenting images that didn't contain partial thickness burns and weighting the predictions before the softmax layer using a 1x1 conv. Upsampling was mildly successful while the weighting was disastrous.

	PA	IOU
FCN - CRF	.60	.37
Upsampled	.57	.39
Fixed Weighted	.33	.19
Learned Weighted	.36	.24

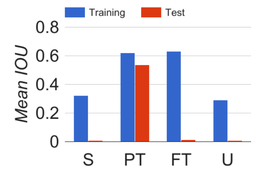
(a) Performance on the test set when we expand to multiple burn depths. Fixed weighted refers to keing the 1x1 conv layer static while learned refers to allowing this layer to be updated.

Discussion

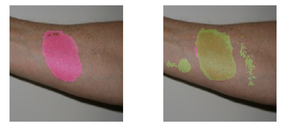
Predicted

	B	S	PT	FT	U
B	650K	105K	458K	50K	4K
S	13K	11K	54K	2K	0
PT	78K	23K	607K	21K	0
FT	72K	2K	88K	44K	0
U	35K	0	26K	2K	0

(a) Confusion matrix of results on the test set. We see that most of our wrong predictions are predicting no burn or over-predicting partial thickness burns.



(b) We see all of the accuracy is coming from predicting partial thickness burns further indicating that the mass prevalence of PT burns is hindering our training.



A clear example where the true burn is a full thickness burn (left), but we classify it as partial thickness (right).

Future Work

Dataset: The dataset needs to continue to be developed in terms of size and variance. Actively seeking out images corresponding to non partial thickness burns as well as those of non-clear skin is needed. If the distribution of burns is similar to what we've seen (in that the majority of images have PT burns), developing an efficient augmentation strategy to emphasize areas of images with non PT burns is critical as only sampling images without PT leads us to only sample from the tails of the distribution. Lastly, we need to have a more fine-grained labeling strategy.
Modeling: Given the in-feasibility of having perfectly labeled data, augment the current technique with weak/semi-supervised methods. To account for the class imbalance, weight logits after the softmax instead of before the softmax.
Metrics: Semantic segmentation metrics, such as IOU, need to be converted to clinically relevant classification metrics, such as PPV and AUC.

References

1. Russell, B.C., Torralba, A., Murphy, K. P., and Freeman W.T. "LabelMe: a database and web-based tool for image annotation." In: *International Journal of Computer Vision* 77 (2008), pp. 157-173.
2. Long, J., Shelhamer, E., and Darrell T. Fully Convolutional Networks for Semantic Segmentation. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3431-3440.
3. Zheng, S. et al. Conditional Random Fields as Recurrent Neural Networks. In: *The IEEE International Conference on Computer Vision (ICCV)* (2015), pp. 1529-1537.