[NSFW] Deep Visual Learning of Reddit Images

Tyler Chase
tchase56@stanford.edu

Rolland He
rhe@stanford.edu

Kareem Hegazy
khegazy@stanford.edu

Introduction

Reddit is an online social news aggregation and internet forum. We apply deep learning to the tasks of image classification by subreddit and automatic not-safe-for-work (NSFW) tagging of images. These tasks have varying applications ranging from subreddit posting suggestions for an image to automatic NSFW tagging of a post. We also investigate improvements using multitask learning and visualize characteristic features in specific subreddits. Unlike ImageNet, our categories are built from communities where features are often shared, requiring learning of more subreddit specific features. To our knowledge, there hasn't been work done applying deep learning to Reddit images.

Problem

For our project, we focus on learning visual elements of images in subreddits that primarily contain images:

1. We aim to classify an image to a relevant subreddit and visualize unique features leading to this classification
2. Some images on Reddit contain adult content (i.e. porn) and are labeled as NSFW. We try to build a model that is able to differentiate between NSFW and non-NSFW images.

To accomplish these tasks, we implement well-known CNN architectures, using classification accuracy and a F1 score as the performance metrics for all tasks

Dataset

- Our dataset comes from the Reddit Submission Corpus*, which contains data for all reddit submissions from 2008 to 2016
- We consider 20 hand-selected subreddits that exclusively contain photo submissions
- We took roughly the top 75 image submissions per month for each of the subreddits from 2015 and 2016
- Input data (X): RGB pixels for all images
- Output data (y): subreddit label, nsfw indicator
- The data is fairly evenly distributed between subreddits (3.6-5.4%), but contains only 8.5% NSFW images
- Train: 80%, Val: 10%, Test: 10%

* http://files.pushshift.io

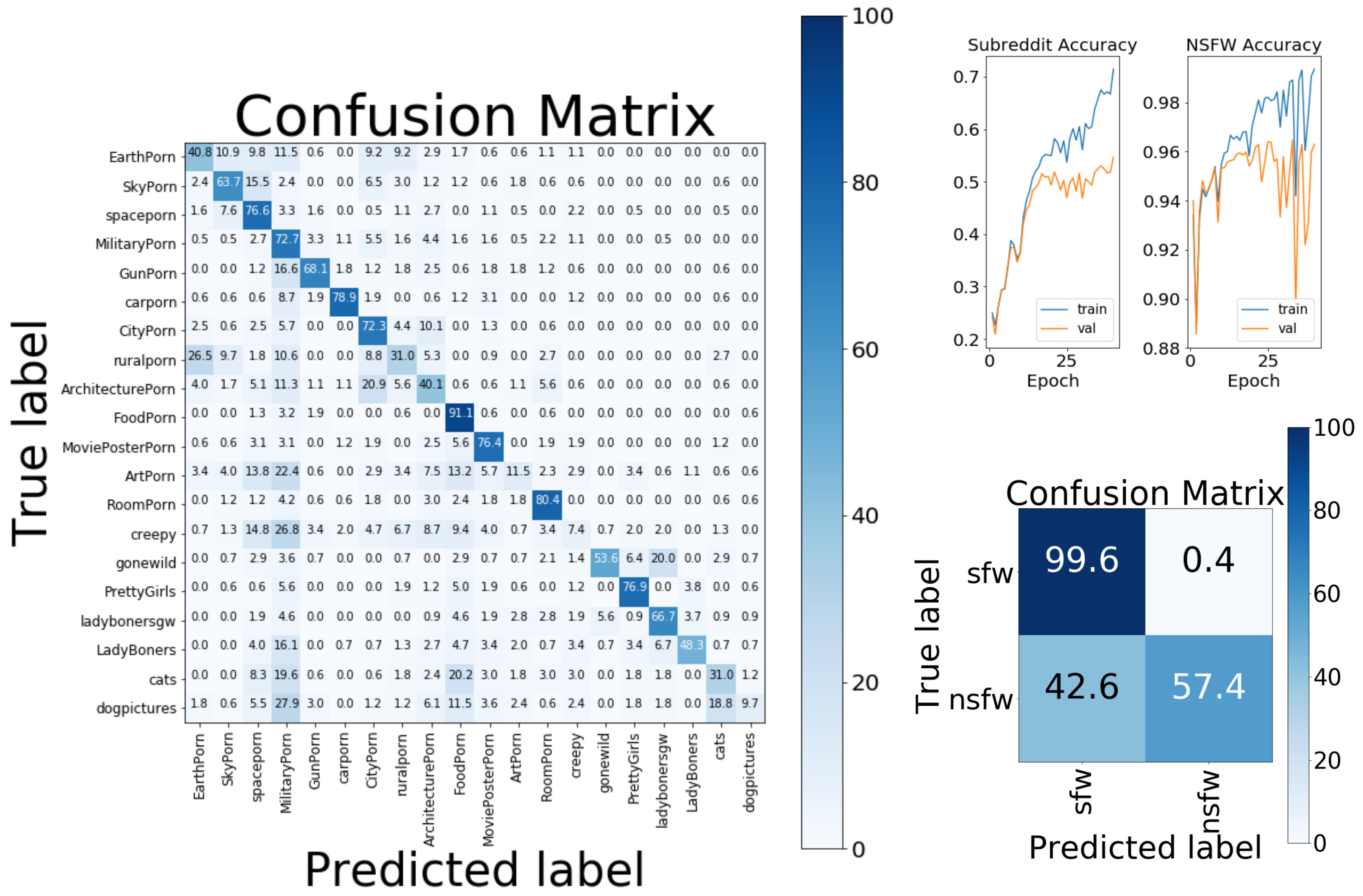
Methodology

- We apply 3 different architectures: AlexNet[3], GoogleNet[4], and ResNet[2], with a few modifications
- Multitask learning[1] was used as a method to improve the performance of each individual task
- Each architecture was applied to 3 separate tasks: standalone subreddit classification, standalone NSFW classification, and multi-task learning combining the 2

Experiments and Results

		AlexNet	GoogleNet	ResNet
Train	Sub	0.485	0.572	0.737
	Multi-Sub	0.532	0.722	0.715
	NSFW	0.915	0.967	0.960
	Multi-NSFW	0.946	0.969	0.994
Val	Sub	0.418	0.500	0.518
	Multi-Sub	0.446	0.505	0.547
	NSFW	0.920	0.951	0.934
	Multi-NSFW	0.938	0.947	0.963
Test	Sub	0.414	0.486	0.520
	Multi-Sub	0.450	0.517	0.551
	NSFW	0.909	0.943	0.931
	Multi-NSFW	0.931	0.942	0.957

Table 1: Results of our 3 architectures trained on each of the 3 tasks for train, val, and test dataset.



Confusion Matrix

True label \ Predicted label	EarthPorn	SkyPorn	spaceporn	MilitaryPorn	GunPorn	carporn	CityPorn	ruralporn	ArchitecturePorn	FoodPorn	MoviePosterPorn	ArtPorn	RoomPorn	creepy	gonewild	PrettyGirls	ladybonersgw	LadyBoners	cats	dogpictures
EarthPorn	40.8	10.9	9.8	11.5	0.6	0.0	9.2	9.2	2.9	1.7	0.6	0.6	1.1	1.1	0.0	0.0	0.0	0.0	0.0	0.0
SkyPorn	2.4	63.7	15.5	2.4	0.0	0.0	6.5	3.0	1.2	1.2	0.6	1.8	0.6	0.6	0.0	0.0	0.0	0.0	0.6	0.0
spaceporn	1.6	7.6	76.6	3.3	1.6	0.0	0.5	1.1	2.7	0.0	1.1	0.5	0.0	2.2	0.0	0.5	0.0	0.0	0.5	0.0
MilitaryPorn	0.5	0.5	2.7	72.7	3.3	1.1	5.5	1.6	4.4	1.6	1.6	0.5	2.2	1.1	0.0	0.0	0.5	0.0	0.0	0.0
GunPorn	0.0	0.0	1.2	16.6	68.1	1.8	1.2	1.8	2.5	0.6	1.8	1.8	1.2	0.6	0.0	0.0	0.0	0.0	0.6	0.0
carporn	0.6	0.6	0.6	8.7	1.9	78.9	1.9	0.0	0.6	1.2	3.1	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.6	0.0
CityPorn	2.5	0.6	2.5	5.7	0.0	0.0	72.3	4.4	10.1	0.0	1.3	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ruralporn	26.5	9.7	1.8	10.6	0.0	0.0	8.8	31.0	5.3	0.0	0.9	0.0	2.7	0.0	0.0	0.0	0.0	0.0	2.7	0.0
ArchitecturePorn	4.0	1.7	5.1	11.3	1.1	1.1	20.9	5.6	40.1	0.6	0.6	1.1	5.6	0.6	0.0	0.0	0.0	0.0	0.6	0.0
FoodPorn	0.0	0.0	1.3	3.2	1.9	0.0	0.0	0.6	0.0	91.1	0.6	0.0	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.6
MoviePosterPorn	0.6	0.6	3.1	3.1	0.0	1.2	1.9	0.0	2.5	5.6	76.4	0.0	1.9	1.9	0.0	0.0	0.0	0.0	1.2	0.0
ArtPorn	3.4	4.0	13.8	22.4	0.6	0.0	2.9	3.4	7.5	13.2	5.7	11.5	2.3	2.9	0.0	3.4	0.6	1.1	0.6	0.6
RoomPorn	0.0	1.2	1.2	4.2	0.6	0.6	1.8	0.0	3.0	2.4	1.8	1.8	80.4	0.0	0.0	0.0	0.0	0.0	0.6	0.6
creepy	0.7	1.3	14.8	26.8	3.4	2.0	4.7	6.7	8.7	9.4	4.0	0.7	3.4	7.4	0.7	2.0	2.0	0.0	1.3	0.0
gonewild	0.0	0.7	2.9	3.6	0.7	0.0	0.0	0.7	0.0	2.9	0.7	0.7	2.1	1.4	53.6	6.4	20.0	0.0	2.9	0.7
PrettyGirls	0.0	0.6	0.6	5.6	0.0	0.0	0.0	1.9	1.2	5.0	1.9	0.6	0.0	1.2	0.0	76.9	0.0	3.8	0.0	0.6
ladybonersgw	0.0	0.0	1.9	4.6	0.0	0.0	0.0	0.0	0.9	4.6	1.9	2.8	2.8	1.9	5.6	0.9	66.7	3.7	0.9	0.9
LadyBoners	0.0	0.0	4.0	16.1	0.0	0.7	0.7	1.3	2.7	4.7	3.4	2.0	0.7	3.4	0.7	3.4	6.7	48.3	0.7	0.7
cats	0.0	0.0	8.3	19.6	0.6	0.0	0.6	1.8	2.4	20.2	3.0	1.8	3.0	3.0	0.0	1.8	1.8	0.0	31.0	1.2
dogpictures	1.8	0.6	5.5	27.9	3.0	0.0	1.2	1.2	6.1	11.5	3.6	2.4	0.6	2.4	0.0	1.8	1.8	0.0	18.8	9.7

Confusion Matrix

True label \ Predicted label	sfw	nsfw
sfw	99.6	0.4
nsfw	42.6	57.4

Fig. 1: Confusion matrices and accuracy plots for our subreddit and nsfw classification models

Conclusion and Future Work

From our results, we make the following observations:

- Multitask learning provides noticeable improvements to both classification tasks. This is expected given that some subreddits contain mostly NSFW images.
- Relative performance of the models are consistent with their performance on ImageNet
- We get varied results for our saliency maps – some maps are able to pinpoint unique features in the images, while other maps seem a bit random (select images below)
- NSFW classification achieves decent results, being able to classify a large majority of SFW and NSFW images (F1 score of 0.815). These results are good, but not quite good enough for automatic filtering of NSFW images.

Future ideas to extend this project:

- Spend more time training/tuning (with more GPU resources), such as hyperparameter optimization
- Simplify architectures/regularize to reduce overfitting
- Attempt to apply NSFW visualization to auto blurring

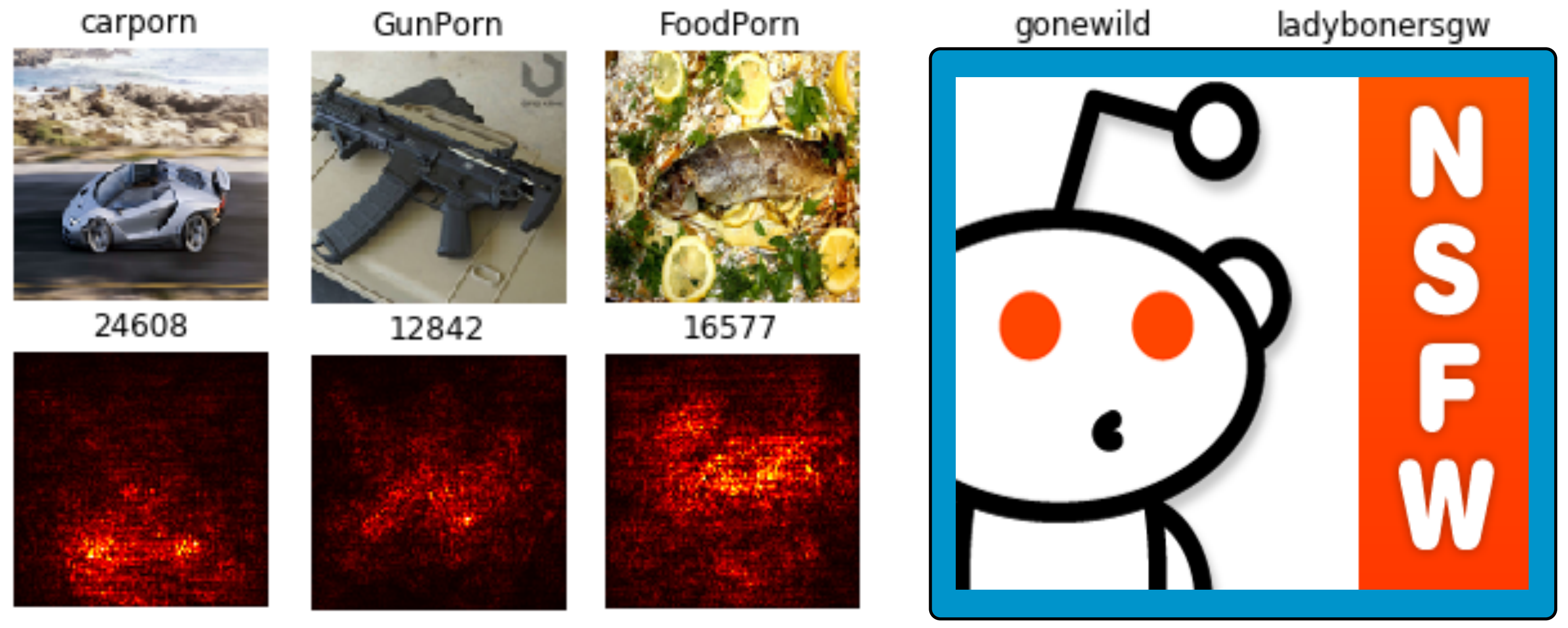


Fig. 2: Saliency maps for select images

Acknowledgements

We would like to thank Timnit, our project mentor; Justin Johnson, Serena Yeung, and Fei-Fei Li for teaching the CS 231N class; Google Cloud Platform for providing us with GPU resources

References

1. Caruana, Rich. "Multitask learning." *Learning to learn*. Springer US, 1998. 95-133.
2. He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
3. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.
4. Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.