How Machine Perception Relates to Human Perception: Visual Saliency and Distance in a Frame-by-Frame Semantic Segmentation Task for Highly/Fully Automated Driving

Nico Herbig
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus
Nico.Herbig@dfki.de

Frederik Wiehr
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus
Frederik.Wiehr@dfki.de

Atanas Poibrenski
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus
Atanas.Poibrenski@dfki.de

Janis Sprenger
German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus
Janis.Sprenger@dfki.de

Christian Müller
Head of Competence Center for Autonomous Driving, DFKI
Saarland Informatics Campus
Christian.Mueller@dfki.de

ABSTRACT

In this paper, we investigate the link between machine perception and human perception for highly/fully automated driving. We compare the classification results of a camera-based frame-by-frame semantic segmentation model (Machine) with a well-established visual saliency model (Human) on the Cityscapes dataset. The results show that Machine classifies foreground objects better if they are more salient, indicating a similarity with the human visual system. For background objects, the accuracy drops when the saliency increases, giving evidence for the assumption that Machine has an implicit concept of saliency.

CCS CONCEPTS

• **Computing methodologies** → *Image segmentation*; *Interest point and salient region detections*;

KEYWORDS

Semantic Segmentation, Saliency, Automated Driving

ACM Reference Format:
Nico Herbig, Frederik Wiehr, Atanas Poibrenski, Janis Sprenger, and Christian Müller. 2018. How Machine Perception Relates to Human Perception: Visual Saliency and Distance in a Frame-by-Frame Semantic Segmentation Task for Highly/Fully Automated Driving. In *SEFAIAS'18: SEFAIAS'18:IEEE/ACM 1st International Workshop on Software Engineering for AI in Autonomous Systems, May 28, 2018, Gothenburg, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3194085.3194092

SEFAIAS'18, May 28, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5739-5/18/05…$15.00
https://doi.org/10.1145/3194085.3194092

1 INTRODUCTION

Frame-by-frame semantic segmentation is one of the standard tasks in the context of deep learning for environment perception in highly/fully automated driving. Given a single frame of a video source, the task is to determine the semantic target class of each pixel within this frame. Depending on the dataset, the set of target classes differs. It typically contains: street, vehicle, pedestrian, building, sidewalk, vegetation, sky, etc. Despite the fact that between-frame information is neglected and pixel-accurate classification does not necessarily reflect the requirements of the application, the task can be considered a good test bed for the underlying algorithms, because it is well-defined and relatively simple.

In this paper, we relate frame-by-frame semantic segmentation, Machine, to a well-known saliency model motivated by human perception, Human. In particular, we show the dependencies between Machine and Human, gaining new insights on how to interpret the segmentation results.

It is important to note that we do not simplify or even modify the underlying semantic segmentation task. Training and classification procedures remain the same. Visual saliency is added afterwards as an additional means for interpreting the results.

In the remainder of this paper, we use autonomous driving and automated driving as equivalent terms. Moreover, highly automated driving is associated with Level 3 autonomous driving, while fully automated driving relates to Levels 4 and 5. Please refer to [7] for an introduction to the level model of autonomous driving.

2 THE ROLE OF SITUATION AWARENESS AND VISUAL SALIENCY IN LEVEL 3 TAKE-OVER SCENARIOS

2.1 Take-Over Situations

In Level 3 (and as a fall-back also in Level 4) autonomous driving, so-called take-over situations are expected to occur, in which the driver is requested to take over control of the vehicle within a relatively short period of time. It is beyond the scope of this article to discuss

(a) Original **(b) Saliency**

Figure 1: An example saliency map of an image from the Cityscapes dataset [2].

the exact nature of such a take-over situation or the minimal take-over time. The only fact important to consider here is that the driver needs to fulfill a few mental and physical steps between, say, reading a novel (or whatever tasks she or he is pursuing) and successfully steering the vehicle in a presumably difficult driving situation (otherwise we would assume the car would not have to initiate take-over in the first place). Those steps or phases are (A) a mental arousal phase, (B) a situation awareness phase, and (C) a physical preparation phase. Mental arousal refers to the cognitive activity that is necessary to perceive the take-over requests and understand their meaning. Physical preparation covers all activities related to getting ready to take action (including storing away keyboards, adjusting the seat position, etc.) Phase B is of interest in this paper – the situation awareness phase.

2.2 Situation Awareness

Among all psychological approaches to situation awareness, Endsley's model is probably one of the best-known [4]. At its core, the model is described as: perceiving the elements in the environment, understanding their meaning, and predicting their status in the future. It is embedded in a loop of decision-making and acting, in which the latter affects the status of the elements, thus looping back to the beginning. Intrinsic factors (such as capacity, skills, and attitudes) as well as extrinsic factors (such as complexity of the scene) influence the loop.

2.3 Visual Saliency

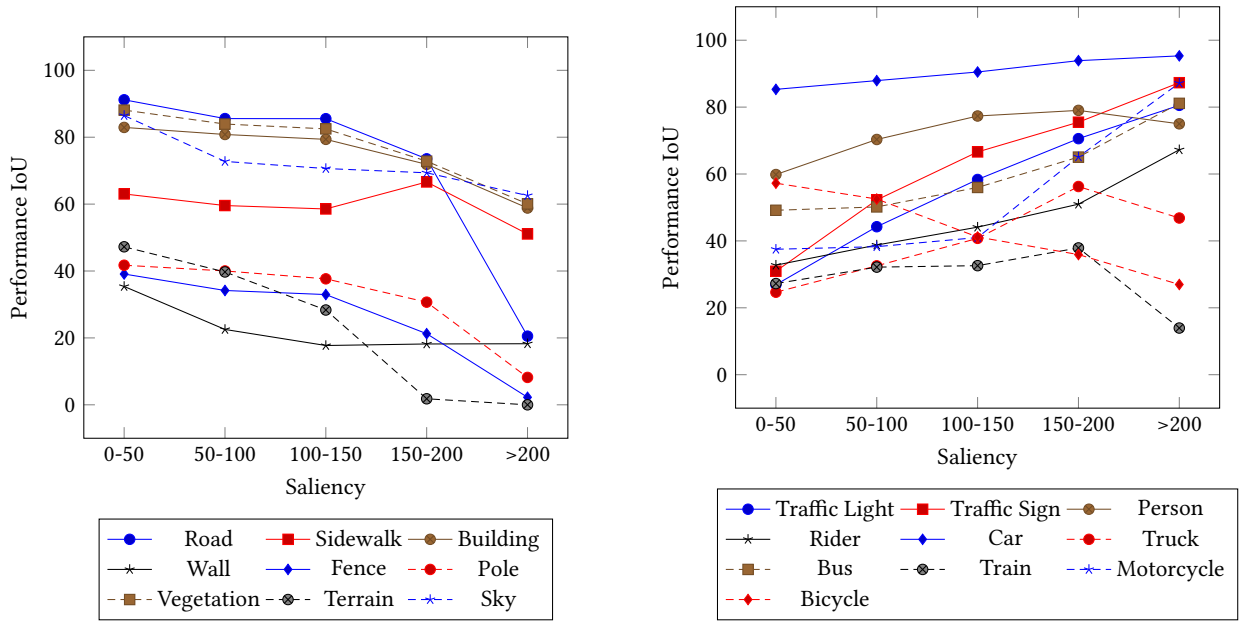
How well (and especially how fast) a human perceives the elements of the environment depends on their saliency. There are two factors influencing visual attention. The first one constitutes the bottom-up features of the visual environment such as color or orientation of objects. For example, a red scarf worn on top of a white shirt automatically stands out and attracts our attention. The second factor is task-dependent, a top-down component that is guided by the human's current task, cognitive abilities and experience. A saliency map is a topographical 2D map that combines the information from different feature maps to assess the global level of conspicuity of any location in a scene. Brighter parts within saliency maps are more salient than darker ones and represent locations that are statistically viewed more frequently by humans. These automatically generated maps are frequently used for image compression or image cropping, where less salient areas can be subject to stronger compressions or can be cropped without too much loss of information.

In this paper, we use the well established Itti-Koch saliency model [6]. While this model is not targeted towards the driving domain, it assembles the bottom-up visual attention reasonably well and has been used for a large variety of applications. The algorithm starts by computing Gaussian pyramids of the input image at nine spatial scales. For each scale, an intensity feature, red, green, blue, and yellow color features, and orientation features (by using Gabor filters) are computed. These are then combined using so-called center-surround differences of the intensity features at different scales, of the orientation features at different scales and different angles, as well as combining the complementary color features for (red, green) and (blue, yellow). The resulting feature maps are normalized and combined, yielding separate conspicuity maps for intensity, color and orientation. In a final step, these three maps are added up using equal weights, resulting in a single saliency map covering all considered bottom-up features (cf. Figure 1). For more information, please consult the original paper.

In the future, this model could be combined with a task-dependent top-down approach to better suit the driving domain. This could include adding road-specific features or using weight modulation to adapt the importance of the different features included in Itti-Koch to the driving domain.

3 THE ROLE OF FRAME-BY-FRAME SEMANTIC SEGMENTATION IN MACHINE PERCEPTION FOR HAD/FAD

Frame-by-frame semantic segmentation is one of the standard tasks in the context of deep learning for environment perception in highly/fully automated driving. Given a single frame of a video source, the task is to determine the semantic target class of each pixel within this frame. The whole image is densely separated into different segments and simultaneously each segment is assigned a semantic class, thus using conceptual differences of objects for the segmentation task. Since semantic segmentation is directly using the sensor data (e.g. RGB-camera) its main task in the context of autonomous driving is the generation of more abstract representations for high-level processes, for instance object recognition and scene understanding.



(a) Performance for Background Classes Relating to Saliency

(b) Performance for Foreground Classes Relating to Saliency

Figure 2: Performance by Distance Group Relating to Saliency

There are multiple challenges published, consisting of a training and an evaluation set of densely labeled 2D RGB images displaying real street scenes viewed from the perspective of a driving car. The label information assigns each pixel to a semantic class (e.g. street, sidewalk, person, car, etc.) thus providing enough data to train and evaluate a neural network. In this work we are using the Cityscapes dataset [2] which provides ground-truth labels for semantic segmentation. This enables us to analyze the results of semantic segmentation in regard to visual saliency and thus compares a simplified version of human visual perception with a neural network for segmentation.

3.1 State of the Art Research

The fully-convolutional network [9] is considered the stepping-stone in deep learning for frame-by-frame semantic segmentation. All of the state-of-the-art methods developed afterwards are based on its simple principle. The main breakthrough was to remove fully-connected layers (in a typical convolutional network architecture) and replace them with fully-convolutional ones, allowing for the output of spatial maps instead of classification scores.

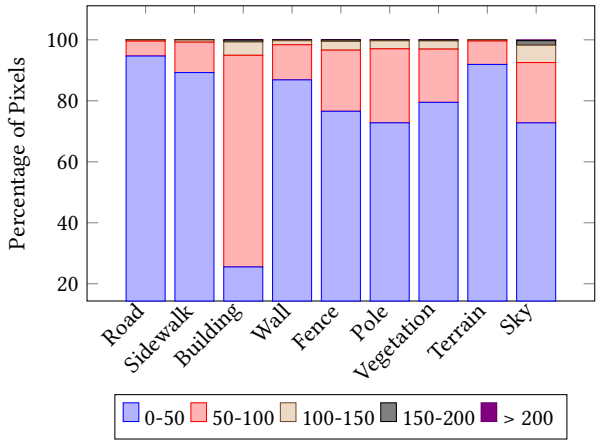
In order for semantic segmentation to work well, it should gather knowledge from multiple spatial resolutions. It is important to have local fine-grained information in order to achieve good pixel-accuracy. On the other hand, global information is crucial as well in order to resolve ambiguities that occur locally. *Multi-scale prediction* is a method that deals with global information and is used in several state-of-the-art segmentation networks. Some of the current top performers in semantic segmentation are PSPNet [11] and Deeplab v3 [1]. They both integrate multi-scale prediction and are based on the ResNet architecture [5]. PSPNet uses a fully-convolutional ResNet to extract features from the input image and then a pyramid parsing module is applied to gather information from different scales. After that, everything is upsampled and concatenated in one feature representation, which now has both local and global information. Deeplab v3 uses spatial pyramid pooling in a similar fashion to PSPNet in order to capture context at several ranges.

Another way to gather context information is by increasing the receptive field of the convolutional filters without losing any resolution. This is done with *dilated convolutions* which are a generalization of Kronecker-factored convolutional filters [12]. The dilated VGG16 network [10] uses these dilated convolutions and is the model used for frame-by-frame segmentation in this paper.

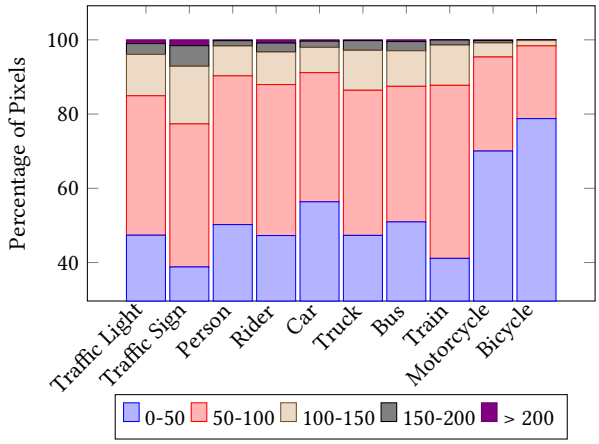
Our human-based saliency model computed at different spatial image scales can provide useful insights on such frame-by-frame segmentation models in terms of how well these models are gathering both local and global image information.

4 THE DEPENDENCIES BETWEEN VISUAL SALIENCY AND DISTANCE TO SEMANTIC SEGMENTATION

The convolutional network serves as a proxy for computer vision; hence, we are not interested in the actual performance of the model, but in the relation of computer vision and human visual perception. Since the segmentation network should identify object distinct features better if they are salient, we hypothesize that there is a relation between saliency and segmentation results: the higher the saliency for a particular pixel is, the better the classification performance of the segmentation model will be. We designed and executed an experiment to verify this hypothesis and to further analyze the behavior of the network.



(a) Percentage of Pixels per Semantic Class and Saliency Group for Background Classes



(b) Percentage of Pixels per Semantic Class and Saliency Group for Foreground Classes

Figure 3: Percentages of Pixels for Background and Foreground Classes

4.1 Experiment Description

We trained a dilated VGG16 model [10], that was initialized on the pre-trained weights from ImageNet [3], on the training set of the Cityscapes dataset [2] (2975 images). The training was done with the standard mini-batch stochastic gradient descent with momentum. Since the full images are too big to fit into GPU memory, a random crop of 628x628 is used. The mini-batch size is set to 12, the learning rate is 0.0001 and the momentum is 0.99. The training of the model was done using the Caffe framework [8] and a Tesla P100 GPU (16GB). The VGG16 model was evaluated on the same set and all pixels were grouped into 5 saliency groups. We report the intersection over union (IoU; see Equation 1) for each semantic class and saliency group.

$$IoU = \frac{\text{true positive}}{\text{true positive} + \text{false positive} + \text{false negative}} \quad (1)$$

This performance measure was chosen because it takes into account both the false alarms and the missed values of each class, giving us more information on what is happening compared to a simple accuracy measure. It has also become a standard for measuring the performance of semantic segmentation algorithms.

In addition, we calculate the saliency map on the evaluation set of Cityscapes (500 images) based on the Itti-Koch model [6] (cf. Subsection 2.3). As stated above, this model only considers bottom-up saliency and is not specific to the driving domain. However, as it is widely used and resembles the human visual system, we consider it a good starting point for relating segmentation performance to visual saliency. Thus, we get a saliency value between 0 and 255 for every single pixel, which we can relate to its IoU.

4.2 Results

The result of the experiment can be seen in Figures 2a and 2b. We are not interested in the actual magnitude of performance accuracy, but in the relation and general trend in regard to saliency. The amount of pixels belonging to each class drops with an increasing saliency, thus the results for a high saliency have to be observed with caution. The percentage of pixels for each class belonging to the different saliency groups are shown in Figures 3a and 3b.

The results are separated into typical background classes (e.g. road, building, vegetation) and typical foreground classes (e.g. traffic light, person, car). An analysis of variance (ANOVA) was conducted on the factors saliency and distance group (background, foreground) on the mean intersection over union for each distance group. The descriptive statistics can be seen in Figure 4. The assumption of sphericity was violated; hence, we applied Greenhouse-Geisser correction, but we report the uncorrected degrees of freedom, in order to increase the readability. There was no main effect for saliency ($F(4, 14) = 1.068$, $p = 0.343$) and no main effect of distance group ($F(1,17) = 0.214$, $p = 0.65$) but there was a significant interaction of saliency and distance group ($F(4,14) = 18.178$, $p < 0.001$).

4.3 Discussion

As expected, the main effect of distance group was not significant, meaning there is no statistical difference between the segmentation performance of background and foreground objects over all saliency groups. The main effect of saliency is not significant either, implying there is no statistical difference between the saliency groups over all semantic classes (independent of foreground and background). However, the interaction is significant, thus showing with statistical certainty that the performance of the segmentation model for background classes decreases with increasing saliency and the performance for foreground classes increases with increasing saliency.

In the case of foreground objects, this means being more distinct from the background and surrounding objects (being more salient) is beneficial for the segmentation. Hence, for the foreground classes, our results support the assumption that similar to the human visual system, the segmentation algorithm can identify salient pixels more easily.

Even more interesting are the results for background classes, where the performance drops with increasing saliency. A large

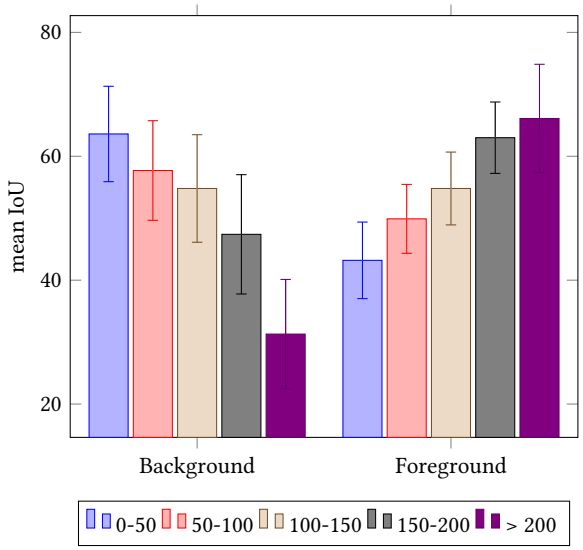


Figure 4: Mean intersection over union (IoU) for all background and foreground classes depending on the saliency. Error bars denote one standard error (SE).

proportion of pixels belonging to these background classes have a very low saliency, whereas foreground objects tend to have a larger portion of more salient pixels. A more in-depth analysis of the confusion matrices shows that background classes are not only confused with foreground classes (e.g. fence being confused with pole), but also with other background classes (e.g. sidewalk with street and vice versa).

The results lead us to the interpretation that the model has an implicit concept of saliency. By definition, foreground objects are more salient than background objects. Thus less salient features should be attributed to background classes and hence increase the performance, whereas more salient features should be attributed to foreground classes and hence increase the performance. Our results support this interpretation with statistical certainty.

However, there are two cases where this interpretation cannot be applied. First of all, the performance of the bicycle class drops, even though we consider it to be a foreground object. Based on our results there is no explanation for this behavior, and it will require further experiments. The second case is the performance drop for the train class, which can be mostly attributed to a confusion with buildings. This is an indication that less salient features of trains and buildings seem to be very similar.

One should note that this experiment is subject to the following limitations: (i) we consider only the VGG16 architecture for semantic segmentation, (ii) and only the Itti-Koch model [6] for saliency estimation, and do not analyze the transferability of the results. However, since VGG16 is a standard model which is referred to by many state-of-the-art segmentation models, and many saliency models are extensions to the one used here, we believe this to be a good choice for our initial analysis. Nevertheless, one should consider more sophisticated models in the future.

5 CONCLUSION

To conclude, there seems to be a relation between object saliency and the classification performance of a segmentation model. Apart from a few exceptions, foreground objects seem to be classified more accurately if they are more salient, while the opposite holds for background objects. This leads to four questions which we want to address in the future: (1) Can results of similar or even better interpretability be produced with a different saliency model, e.g. one focusing more strongly on the driving domain? (2) Does the classification accuracy improve when the saliency of an image is provided additionally to the image itself? The hypothesis here would be that it can learn faster based on the provided prior knowledge. (3) How well do saliency models (potentially incorporating top-down features) capture dangerous traffic situations? We expect e.g. a deer or a child running across the street to be easily detected by even simple saliency models. Could this information, potentially in combination with improved segmentation accuracy (see (2)), be used to develop better perception systems that allow analysis even of complex traffic situations? (4) Can we use saliency estimation to quantify the uncertainty of a semantic segmentation model?

6 ACKNOWLEDGEMENTS

This research was funded in part by the German Federal Ministry of Education and Research under grant number 01IS12050 (project DriveSense) and 01/W17003 (project REACT). The responsibility for this publication lies with the authors.

REFERENCES

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *CoRR* abs/1706.05587 (2017). arXiv:1706.05587 http://arxiv.org/abs/1706.05587

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3213–3223. https://doi.org/10.1109/CVPR.2016.350

[3] J. Deng, W. Dong, R. Socher, L. J. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[4] Mica R. Endsley. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors* 37, 1 (1995), 32–64. https://doi.org/10.1518/001872095779049543 arXiv:https://doi.org/10.1518/001872095779049543

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[6] L. Itti, C. Koch, and E. Niebur. 1998. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (Nov 1998), 1254–1259. https://doi.org/10.1109/34.730558

[7] SAE International Standard J3016. 2014. Taxonomy and Definitions for Terms related to On-Road Motor Vehicle Automated Driving Systems. (2014).

[8] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *CoRR* abs/1408.5093 (2014). arXiv:1408.5093 http://arxiv.org/abs/1408.5093

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2014. Fully Convolutional Networks for Semantic Segmentation. *CoRR* abs/1411.4038 (2014). arXiv:1411.4038 http://arxiv.org/abs/1411.4038

[10] Fisher Yu and Vladlen Koltun. 2015. Multi-Scale Context Aggregation by Dilated Convolutions. *CoRR* abs/1511.07122 (2015). arXiv:1511.07122 http://arxiv.org/abs/1511.07122

[11] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2016. Pyramid Scene Parsing Network. *CoRR* abs/1612.01105 (2016). arXiv:1612.01105 http://arxiv.org/abs/1612.01105

[12] Shuchang Zhou, Jia-Nan Wu, Yuxin Wu, and Xinyu Zhou. 2015. Exploiting Local Structures with the Kronecker Layer in Convolutional Networks. *CoRR* abs/1512.09194 (2015). arXiv:1512.09194 http://arxiv.org/abs/1512.09194