# Towards More Meaningful Deep Representations with Neural Networks

Gustav Grund Pihlgren

Luleå University of Technology, Aurorum 1, 97187 Luleå, Sweden
gustav.pihlgren@ltu.se
https://www.ltu.se/staff/g/gusgru-1.180588?l=en

## 1 Research Direction

The aim of this project is to improve the meaningfulness of the features and deep representations generated by neural networks. Additionally the research also aims towards provide methods for understanding deep representations.

This topic have already seen many major breakthroughs in different areas of machine learning. Examples of such breakthroughs are word embeddings, autoencoders, generative adversarial networks, and internal world models. Research into how to make machine learning generate even more useful representations and how to use these could be greatly beneficial to all these common machine learning components. As many of these components are already used in real world applications an improvement that is beneficial to them will also have immediate real world effects.

## 2 Research Questions

This project invites for many different interesting research questions. Here ordered in the order of which the project will probably examine them or similar research questions.

- How does loss functions affect the encodings created by autoencoders?
- How can a model be trained to learn importance of features in an unsupervised fashion?
- How can an encoder be trained to learn representations that generalizes to previously unseen features?
- How can an agent learn to create an internal representation of the world in an unsupervised manner?
- How can an agent be incentivized to improve its representation of the world?

## 3 Current Work

Currently work is being done on the first of the research question mentioned. The use of perceptual loss to create more meaningful features in autoencoders is being investigated. Normally when training autoencoders for images the error is calculated by comparing each pixel in the target to the corresponding pixel in the output. The error is calculated by taking the average pixel-wise loss which is usually calculated with mean-square error or cross-entropy loss.

Pixel-wise loss suffers from being fundamentally different from how a person would view the similarity of two images. For example an image of white and black stripes that has been shifted one pixel to the side would remain a very similar to the original in the eyes of a human but the element-wise error would be very high as the target would have black pixels where the output has white and vice versa. Rather than using pixel-wise loss one can use perceptual loss.

Perceptual loss on images makes use of another network trained to solve some other image problems (like classification). Rather than comparing the target directly to the output one compares features extracted from some layer of the second network. Since perceptual loss does not suffer from the problems caused by pixel-wise loss calculation the expectations is that it will improve the encoding of some features. Especially features where a small shift would cause a relatively large increase in element-wise loss.

Preliminary experiments have been done with trying to encode images taken from the OpenAI Gym problem LunarLander-v2 [2]. The problem is a game where the player attempts to land a lunar lander on the surface without crashing. So far two types of autoencoders have been used; standard and variational. After training the encodings a linear perceptron is trained to predict the position of the lander from the encodings. Results show that prediction from encodings trained with perceptual loss is better than those trained with pixel-wise loss with both types of autoencoders.

## 4 Future Work

The project's current focus is in the realm of autoencoders for the creation of deep representations. Future work will try to generalize the knowledge gained to other models that generate deep representations. Specifically in the realm of world modelling in reinforcement learning. *Hindsight Experience Replay* [1] and *On learning to think* [3] are two papers which present interesting findings in that direction that the project aims to build upon.

## References

1. Andrychowicz, M., Wolski, F., Ray, A., Schneider, J., Fong, R., Welinder, P., McGrew, B., Tobin, J., Abbeel, O.P., Zaremba, W.: Hindsight experience replay. In: Advances in Neural Information Processing Systems. pp. 5048–5058 (2017)
2. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: Openai gym (2016)
3. Schmidhuber, J.: On learning to think: Algorithmic information theory for novel combinations of reinforcement learning controllers and recurrent neural world models. arXiv preprint arXiv:1511.09249 (2015)