The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Rating Super-Resolution Microscopy Images with Deep Learning

Louis-Émile Robitaille,* **Audrey Durand, Marc-André Gardner, Christian Gagné**
LVSN, Université Laval, Canada

Paul De Koninck
Flavie Lavoie-Cardinal
CERVO, Université Laval, Canada

In order to improve their understanding, cellular mechanisms need to be observed at the nanoscale, which is allowed by optical super-resolution microscopy. Among these techniques, STimulated Emission Depletion (STED) microscopy (Hell and Wichmann 1994; Klar et al. 2000; Willig et al. 2006)[1] brings a 10 fold improvement of the imaging resolution over conventional optical microscopy, allowing to observe molecular structures and protein complexes of living cells in action (Sahl, Hell, and Jakobs 2017). Super-resolution microscopes are highly specialized devices, significantly more complex to use than conventional optical microscopes, hence reducing their accessibility. Moreover, the overall quality of the obtained images can vary a lot depending on the imaging parameters or the biological structure of interest. It is therefore very difficult to evaluate the quality of such images for non-expert users, making it a challenge when tuning imaging parameters toward good images.

In this work, we tackle the problem of learning to evaluate the quality of STED images. This could allow not only to support non-experts in their measurements, but also constitute a step toward a fully automated imaging system. We address this problem using deep learning. To this end, a brand new dataset was built, which is used to train the network and assess its performance. We then conduct a user study to evaluate the capability of the network for fooling an expert in front of other experts. We also evaluate the capability of the network to generalize its quality prediction to STED images of a different protein.

Problem Statement

Let $\mathcal{I}$ denote the space of possible STED images. We aim at learning the quality function $f : \mathcal{I} \mapsto [0, 1]$ that takes as input an image and outputs a quality score. This corresponds to a standard regression problem.

The quality score of an image incorporates several features such as the resolution of the observed structures, the signal-to-noise ratio (SNR), the deterioration of the fluorophores (photobleaching) and structure (photoxicity) due to the imaging process, or the observability of specific structures. The quality score given to an image by an expert is therefore some sort of (unknown) tradeoff between several objectives.

Proposed Approach

We propose to learn the quality function from an expert using a CNN (Convolutional Neural Network). More specifically, we consider a network made of 6 convolutional layers and 2 fully connected layers. An ELU activation (Exponential Linear Unit) is used after each convolutional and fully connected unit. Max pooling (kernel 2x2, stride 1) is added after each convolutional unit. Batch normalization is applied to all the layers except the first one. The output is driven by a nonlinear activation (sigmoid) to retrieve a quality score between 0 and 1.

A brand new dataset has been built for the task at hand. It contains 1140 grayscale images of 224×224 pixels of 20 nm. Each image was obtained by the imaging of the protein actin. Different qualities of images were produced by changing the acquisition parameters. Images have been labeled by an expert, where each label corresponds to a quality score in $[0, 1]$. A 80/10/10 split of the randomly shuffled dataset is used for training, validation, and testing respectively. The training set was doubled using data augmentation, resulting in 1,824 training images. MSE was used as loss function and we reached $\sim 12\%$ RMSE.

User Study

If the MSE is a an acceptable loss function for training, it is not very informative for assessing the ability of the neural network to fully replace a human expert in the learning loop. Indeed, if experts appear to be very noisy, the MSE might be high while the system might be performing quite well. It is therefore interesting to compare the system predictions *against* the expert, from the perspective of another expert.

To this extent, we developed a web-based application that would sequentially present an expert with STED images and two scores: the target and the network prediction, in random order. The expert would either pick the most relevant score, mark both scores as *equivalent*, or discard the image. The last case means that an error occured at the time of the labeling since that neither the prediction nor the target appears

*1065, avenue de la Médecine, Québec, QC, Canada,+1 (581) 308-3921, louis-emile.robitaille.1@ulaval.ca

[1] Stefan W. Hell was awarded a Nobel prize in 2014 for this revolutionary microscopy technique.

Table 1: Confusion (%) given the target quality (%).

	Actin		Tubulin	
Target	Network	Random	Network	Random
0 - 20	**63±25**	26±15	**58±25**	4±9
20 - 40	**61±29**	46±22	**77±20**	44±23
40 - 60	**76±10**	**76±17**	**80±12**	44±15
60 - 80	**84±11**	70±9	79±20	**85±13**
80 - 100	**68±20**	48±37	–	–

Table 2: Domination (%) given the target quality (%).

	Actin		Tubulin	
Target	Network	Random	Network	Random
0 - 20	**28±21**	10±9	**33±23**	2±4
20 - 40	**32±22**	23±11	**51±26**	16±14
40 - 60	**48±14**	**48±15**	**57±18**	20±8
60 - 80	**40±12**	30±7	**61±18**	46±13
80 - 100	**61±24**	25±20	–	–

to be good to the expert tester. Similar user studies for assessing the capability of a system to produce realistic results from a human perspective have been used previously (Gardner et al. 2017).

Measuring Performance

Let $\tilde{N}$ denote the size of the *effective* test set, that is the number of images that were not discarded by the tester. Let T, P, and E respectively denote the number of images where the tester picked the target, the prediction, and marked them as equivalent. We introduce two performance measures:

$$\mathcal{C} = 1 - \frac{\left|(2P + E) - \tilde{N}\right|}{\tilde{N}} \quad \text{and} \tag{1}$$

$$\mathcal{D} = \frac{P}{T + P}. \tag{2}$$

Confusion (Eq. 1) indicates whether network predictions can be confused with true targets by other experts, and **domination** (Eq. 2) indicates how much the network predictions are beating the labeling expert.

Benchmark

The user study is performed on two datasets. The **Actin** dataset contains 103 images of the actin protein on fixed hippocampal neurons. More specifically, these images are drawn from the 10% test split taken from the initial data. The **Tubulin** dataset contains 94 images of a different protein: the tubulin of cytoskelettal.

The experiment is performed by 11 experts. Each expert performs the experiment for both datasets. None of these experts were involved in the gathering of the data. The neural network is compared against a random system that predicts a score by sampling it uniformly from the training labels, therefore predicting based on the training data distribution.

Results

Tables 1 and 2 show these performance measures per bin of quality scores, averaged over the 11 testers, with one standard deviation. More specifically, the performance measures are calculated *for each tester* and their scores are then averaged. We observe that the proposed network approach beats the random baseline in almost all target quality bins, on both datasets, and regarding both measures. Note that not enough data are currently available in order to make these results statistically significant and further experiments would be required to this extent. We also observe that both algorithms have troubles when it comes to predict a low score, a possible explanation could come from the data distribution. However, our network outperforms the random strategy, hence exhibiting a capacity of generalization. More details can be found in the supplementary materials.

Surprisingly, the network obtains a high domination performance on the tubulin protein dataset. Recall that images of this particular protein have never been seen by the network. In other words, the network predicts scores that often appear to be even better than the true targets, from the eye of a tester. This is a very interesting situation that raises questions regarding the noise inherent to human labelling as well as the human perception of subtle concepts such as *quality*.

Discussion and openings

The obtained results raise several questions. For example, how can predictions of the proposed network become better than actual scores given by a human expert? Could we use the resulting network to help understanding the quality scoring process by a human expert? More specifically, given that the quality function is driven by the appearance of specific structures, could the resulting network be able to detect these structures? This application is a first step toward the automatization of analysis and optimization tasks for neuroscientists working with high-end microscopy settings. In fact, the system has been fully deployed on a STED setting and is currently being used in a control loop for optimizing imaging parameters. These tools have the potential to help users in taking full advantage of these systems, which could facilitate the adoption of this powerful technique.

References

Gardner, M.-A.; Sunkavalli, K.; Yumer, E.; Shen, X.; Gambaretto, E.; Gagné, C.; and Lalonde, J.-F. 2017. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (SIGGRAPH Asia)* 9(4).

Hell, S. W., and Wichmann, J. 1994. Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics letters* 19(11):780–782.

Klar, T. A.; Jakobs, S.; Dyba, M.; Egner, A.; and Hell, S. W. 2000. Fluorescence microscopy with diffraction resolution barrier broken by stimulated emission. *Proceedings of the National Academy of Sciences* 97(15):8206–8210.

Sahl, S. J.; Hell, S. W.; and Jakobs, S. 2017. Fluorescence nanoscopy in cell biology. *Nature reviews. Molecular cell biology*.

Willig, K. I.; Kellner, R.; Medda, R.; Hein, B.; Jakobs, S.; and Hell, S. W. 2006. Nanoscale resolution in gfp-based microscopy. *Nature Methods* 3(9):721–723.