Capstone Proposal - Udacity MLND

Eric G. Cavalcanti

(Dated: July 29, 2018)

I. DOMAIN BACKGROUND

The aim of this project is to use machine learning techniques to estimate photon counts from the output of a transition-edge sensor (TES) detector illuminated by laser light. Transition-edge sensors provide, among other applications, unprecedented sensitivity for photon-number resolved optical detection, with applications in quantum information and computation technologies [1, 2]. The capability of accurately and efficiently resolving photon numbers with TES detectors has the potential to open up higher-dimensional applications in optical quantum information processing.

TES detectors provide the highest efficiencies among photon-number resolving detectors [3]. One of the drawbacks of these sensors however are the slow response times and thus small detection rates. One of the motivations for this project is to explore whether machine learning techniques could provide insights on the structure and classification of TES traces, that could potentially translate into the possibility of performing hardware-based classification of photon numbers and accelerate detection rates.

Further details on the TES detector used for this project can be found in an introductory set of notes [4], available in https://github.com/egcavalcanti/machine-learning, under `projects/capstone/photon_detector/TES_introductory_material.pdf`.

II. PROBLEM STATEMENT

This project will utilise data produced by Geoff Gillett from the University of Queensland's Quantum Technology Laboratory. This data consists of a collection of output signals ("traces") produced by a TES illuminated by a laser source. As the energy of a pulse is proportional to the number of photons, photon number in these traces has been estimated via the area of the signal (and thus its energy). The distribution of pulse areas for each photon number value follows a slightly skewed Gaussian distribution. For low photon numbers, an estimate of photon numbers via pulse area appears to be quite accurate as the clusters are reasonably well separated. For higher photon numbers however, there is considerable overlap between the clusters, and thus an estimate by area alone seems to be insufficient to fully resolve photon counts. The goal of the present analysis is to explore whether unsupervised learning techniques could provide an improvement on these estimates.

As the photon source produces optical coherent states, it is expected that the photon numbers follow a Poissonian distribution, which can therefore serve as an independent benchmark to compare the technique produced as part of this work with the direct estimate via trace areas.

III. DATASETS AND INPUTS

The TES data to be utilised in this work was produced by Geoff Gillett at the University of Queensland QT Lab, and permission has been given by that author to use the data in this work, and for its publication on Github. The directions to download the dataset are given in the README.md file found in https://github.com/egcavalcanti/machine-learning, under `projects/capstone/photon_detector`.

The file `TES_dataset.npz` contains several numpy arrays:

- `trace`: a numpy record array of 3×10^5 captured traces.
- `raw_hist`: a histogram created from 3×10^7 area measurements taken with the same optical input as the traces.
- `hist_x`: the trace area ($\sim$energy) values at the center of the histogram bins.
- `smooth_hist`: `raw_hist` smoothed using a gaussian filter.
- `peaks`: a sequence of slices that divide the histogram data into 12 distinct peaks, based on the mid point between maxima.
- `max_i`: the indices of the first 12 peaks in the smoothed histogram.
- `std`: the std deviations of the first 12 peaks estimated from the FWHM.

More details on the dataset and an initial analysis showing the Gaussian distributions for each photon number, as well as a plot of traces classified via area measurements is shown in the notebook `TES_dataset.ipynb` (also exported to `TES_dataset.html`) contained in the same directory as the dataset.

IV. SOLUTION STATEMENT

A solution to the problem presented in this project would be a classification of the traces in the dataset into clusters corresponding to pulse photon numbers. Ideally, this classification would be closer to a Poissonian distribution than the benchmark classification based on trace areas, indicating that it is closer to the expected photon-number distribution of the coherent laser source. However, due to the limited size of the dataset used here, a classification that is close to but inferior to that benchmark could be considered a successful result for the purposes of this exploratory project, and merit further exploration with a larger dataset.

V. BENCHMARK MODEL

The benchmark model for this analysis is the classification based on areas, displayed in the notebook `TES_dataset.ipynb`.

VI. EVALUATION METRICS

Unfortunately, there is no independent estimate of the photon number in a particular pulse apart from the pulse area. This makes it impossible to determine the "true" number of photons detected by any given pulse. As mentioned above, however, an independent ground-truth evaluation metric is given by the fact that laser sources produce coherent states, which follow a Poissonian distribution of photon numbers. That is, for a coherent state the probability of detecting n photons in a given time interval is given by:

$$P(n) = e^{-\langle n \rangle} \frac{\langle n \rangle^n}{n!} \ ,$$

where $\langle n \rangle$ is the expected number of photons in that interval. Photon number estimates that are closer to a Poissonian distribution would therefore be considered more accurate than those further from it.

VII. PROJECT DESIGN

The project will proceed by employing unsupervised learning techniques to the `trace` record in the dataset. Firstly, I will separate the original data into a training and test

subsets. As the area distributions for each photon number is approximately Gaussian, a reasonable expectation is that a Gaussian Mixture model would give good results. Therefore the starting point of the analysis will be to produce a GMM classification of the training dataset. Potential difficulties are expected to arise from the fact that for higher photon numbers ($n > 12$), the clusters are small, very mixed and deviate from a Gaussian distribution. I will thus consider excluding data corresponding to clusters above 12 photons, if feasible. This would also allow for a more direct comparison with the original area classification, which is limited to $n \leq 12$.

As each trace contains a time series of 2048 intensity values, this is a high-dimensional clustering problem, and it would likely benefit from dimensionality reduction. I will therefore apply dimensionality reduction techniques such as principal component analysis in an attempt to determine a lower-dimensional description of the dataset that captures the clusters sufficiently well. It is very likely, of course, that one of these dimensions will correspond to the area of the pulse, but the hope is that the analysis will show that there are other relevant dimensions along which to separate the pulses, and thus provide a more accurate cluster classification.

Finally, I will compare the cluster classification obtained via the above techniques with the original area-based estimation, against the expected Poissonian distribution on the test dataset, as a final evaluation.

---

[1] D. H. Smith, G. Gillett, M. P. de Almeida, C. Branciard, A. Fedrizzi, T. J. Weinhold, A. Lita, B. Calkins, T. Gerrits, H. M. Wiseman, S. W. Nam, and A. G. White, Nature Communications **3**, 625 (2012), arXiv:1111.0829.

[2] B. G. Christensen, K. T. McCusker, J. B. Altepeter, B. Calkins, T. Gerrits, A. E. Lita, A. Miller, L. K. Shalm, Y. Zhang, S. W. Nam, N. Brunner, C. C. W. Lim, N. Gisin, and P. G. Kwiat, Physical Review Letters **111**, 130406 (2013), arXiv:1306.5772.

[3] M. D. Eisaman, J. Fan, A. Migdall, and S. V. Polyakov, Review of Scientific Instruments **82**, 071101 (2011), https://doi.org/10.1063/1.3610677.

[4] G. Gillett, *TES Introductory material* (2018).