TOWARDS QUANTUM INSPIRED CONVOLUTION NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Convolution Neural Networks (CNNs), rooted by the pioneer work of Rumelhart et al. (1986); LeCun (1985); Krizhevsky et al. (2012), and summarized in LeCun et al. (2015), have been shown to be very useful in a variety of fields. The state-of-the art CNN machines such as image rest net He et al. (2016) are described by real value inputs and kernel convolutions followed by the local and non-linear rectified linear outputs. Understanding the role of these layers, the accuracy and limitations of them, as well as making them more efficient (fewer parameters) are all ongoing research questions.

Inspired in quantum theory, we propose the use of complex value kernel functions, followed by the local non-linear absolute (modulus) operator square. We argue that an advantage of quantum inspired complex kernels is robustness to realistic unpredictable scenarios (such as clutter noise, data deformations). We study a concrete problem of shape detection and show that when multiple overlapping shapes are deformed and/or clutter noise is added, a convolution layer with quantum inspired complex kernels outperforms the statistical/classical kernel counterpart and a "Bayesian shape estimator" . The superior performance is due to the quantum phenomena of interference, not present in classical CNNs.

1 INTRODUCTION

The convolution process in machine learning maybe summarized as follows. Given an input $f^{L-1}(x) \geq 0$ to a convolution layer L, it produces an output

$$g^L(y) = \int K(y-x) f^{L-1}(x)\,\mathrm{d}x$$

From $g^L(y)$ a local and non-linear function is applied, $f^L(y) = f(g^L(y))$, e.g., $f = ReLu$ (rectified linear units) or $f = |.|$, the magnitude operator. This output is then the input to the next convolution layer (L+1) or simply the output of the whole process. We can also write a discrete form of these convolutions, as it is implemented in computers. We write $g_i^L = \sum_j w_{ij} f_j^{L-1}$, where the continuous variables y, x becomes the integers i, j respectively, the kernel function $K(y-x) \to w_{ij}$ becomes the weights of the CNN and the integral over $\mathrm{d}x$ becomes the sum over j.

These kernels are learned from data so that an error (or optimization criteria) is minimized. The kernels used today a real value functions. We show how our understanding of the optimization criteria "dictate" the construction of the quantum inspired complex value kernel. In order to concentrate and study our proposal of quantum inspired kernels, we simplify the problem as much as possible hoping to identify the crux of the limitation of current use of real value kernels.

We place known shapes in an image, at any location, and in the presence of deformation and clutter noise. These shapes may have been learned by a CNN. Our main focus is on the feedforward performance, when new inputs are presented. Due to this focus, we are able to construct a Bayesian a posteriori probability model to the problem, which is based on real value prior and likelihood models, and compare it to the quantum inspired kernel method.

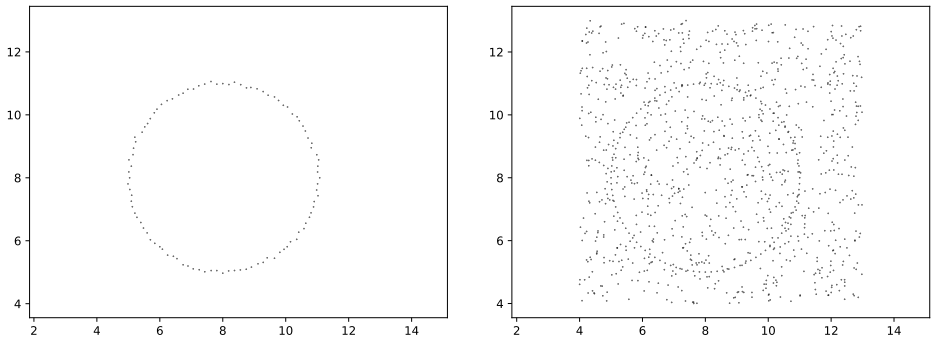
The main advantage of the quantum inspired method over existing methods is its high resistance to deviations from the model, such as data deformation, multiple objects (shapes) overlapping, clutter noise. The main new factor is the quantum interference phenomenon Feynman & Hibbs (2012);

Feynman (1971), and we argue it is a desired phenomena for building convolution networks. It can be carried out by developing complex value kernels driven by classic data driven optimization criteria. Here we demonstrate its strength on a shape detection problem where we can compare it to state of the art classical convolution techniques. We also can compare to the MAP estimator of the Bayesian model for the shape detection problem.

To be clear, we do not provide (yet) a recipe on how to build kernels for the full CNN framework for machine learning, and so the title of this paper reflects that. Here, we plant a seed on the topic of building complex value kernels inspired in quantum theory, by demonstrating that for a given one layer problem of shape detection (where the classic data optimization criteria is well defined), we can build such complex value kernel and demonstrate the relevance of the interference phenomena. To our knowledge such a demonstration is a new contribution to the field. We also speculate on how this process can be generalized.

1.1 The Specific Problem

We are given an image $\mathcal{I}$ with some known objects to be detected. The data is a set of N *feature points*, $X = \{x_1, x_2, ..., x_N\}$ in $\mathbb{R}^D$. Here we focus on 2-dimensional data, so $D = 2$. An image $\mathcal{I}$ may be described by the set of feature points, as shown for example in figure 1. Or, the feature points can be extracted from $\mathcal{I}$, using for example, SIFT features, HOG features, or maximum of wavelet responses (which are convolutions with complex value kernels). It maybe be the first two or so layers of a CNN trained in image recognition tasks. The problem of object detection has been well addressed for example by the SSD machine Liu et al. (2016), using convolution networks. Here as we will demonstrate, given the points, we can construct a ONE layer CNN that solves the problem of shape detection and so we focus on this formulation. It allows us to study in depth its performance (including an analytical study not just empirical).



(a) A deformed circle (b) Uniformly distributed noise added to (a)

Figure 1: (a) A circle centered at $(8, 8)$, radius 3, and with 100 points, is deformed as follows: for each point x_i a random value drawn from a uniform distribution with range $(-\eta_i,\ \eta_i)$, $\eta_i = 0.05$, is added along the radius. (b) 1000 points of clutter are added by sampling from a uniform distribution inside a box of size 9×9, with corners at points: (4,4), (4,13), (13,4), (13,13).

1.2 Paper Organization

We organize the paper as follows. Section 2 presents a general description of shapes, which is easily adapted to any optimization method. Section 3 presents the Bayesian method and the Hough transform method (and a convolution implementation) to the shape detection problem. Section 4 lays out our main proposal of using quantum theory to address the shape detection problem. The theory also leads naturally to a classical statistical method behaving like a voting scheme, and we establish a connection to Hough transforms. Section 5 presents a theoretical and empirical analysis of the quantum method for shape detection and a comparison with the classical statistical method. We demonstrate that for large deformations or clutter noise scenarios the quantum method outperforms the classical statistical method. Section 6 concludes the paper.

2 SHAPE DETECTION

A shape $\mathcal{S}$ may be defined by the set of points x satisfying $\mathcal{S}_\Theta(x) = 0$, where Θ is a set of parameters describing $\mathcal{S}$. Let μ be a shape's center (in our setting $\mu = (\mu_x, \mu_y)$). The choice of μ is in general arbitrary, though frequently there is a natural choice for μ for a given shape, such as its "center of mass": the average position of its coordinates. We consider all the translations of $\mathcal{S}_\Theta(x)$ to represent the same shape, so that a shape is translation invariant. It is then convenient to describe the shapes as $\mathcal{S}_\Theta(x - \mu)$, with the parameters Θ not including the parameters of μ. Thus we describe a shape by the set of points X such that

$$\mathcal{S}_\Theta(x - \mu) = 0 \quad \text{for all } x \in X.$$

The more complex a shape is, the larger is the set of parameters required to describe it. For example, to describe a circle, we use three parameters $\{\mu_x, \mu_y, r\}$ representing the center and the radius of the circle, i.e., $\mathcal{S}_{\Theta=\{r\}}(x - \mu) = 1 - \frac{(x-\mu)^2}{r^2}$ (see figure 1 a.) An ellipse can be described by $\mathcal{S}_\Theta(x - \mu) = 1 - (x - \mu)^\mathsf{T} \Sigma^{-1} (x - \mu)$, where $\Theta = \{\Sigma\}$ is the covariance matrix, specified by three independent parameters.

We also require that a shape representation be such that if all the values of the parameters in Θ are 0, then the the set of points that belong to the shape "collapses" to just $X = \{\mu\}$. This is the case for the parameterizations of the two examples above: the circle and the ellipse.

Energy Model: Given a shape model we can create an *energy model* per data point x as

$$\mathrm{E}_{\mathcal{S}_\Theta}(x - \mu) = |\mathcal{S}_\Theta(x - \mu)|^p \tag{1}$$

where the parameter $p \geq 0$ defines the L_p norm (after the sum over the points is taken and the $1/p$ root is applied). The smaller $E_{\mathcal{S}_\Theta}(x - \mu)$, the more it is likely that the data point x belongs to the shape $\mathcal{S}_\Theta$ with center μ. In this paper, we set $p = 1$ because of its simplicity and robust properties.

2.1 DEFORMATIONS

To address realistic scenarios we must study the detection of shapes under deformations. When deformations are present, the energy (1) is no longer zero for deformed points associated to the shape $\mathcal{S}_\Theta(x - \mu)$. Let each ideal shape data point x_i^S be deformed by adding η_i to its coordinates, so $x_i = x_i^S + \eta_i$. Then the deformed points x_i satisfy

$$0 = |\mathcal{S}_\Theta(x_i^S - \mu)| = |\mathcal{S}_\Theta(x_i - \eta_i - \mu)|$$

Deformations of a shape are only observed in the directions perpendicular to the shape tangents, i.e., along the direction of $\nabla_x \mathcal{S}_\Theta(x - \mu)\Big|_{x_i}$, where ∇_x is the gradient operator.

For example, for a (deformed) circle shape, $\Theta = \{r\}$ and $\mathcal{S}_r(x-\mu) = 1 - \frac{(x-\mu)^2}{r^2}$, and so $\nabla_x \mathcal{S}_r(x - \mu)\Big|_{x_i} = 2\,\frac{(x_i-\mu)}{r^2} \propto \hat{r}_i$, where $\hat{r}_i$ is a unit vector pointing outwards in the radius direction at point x_i. Thus, $\eta_i = \eta_i\,\hat{r}_i$.

3 CLASSICAL METHODS: A BAYESIAN AND A HOUGH TRANSFORM

Given a set of data points $X = \{x_1, x_2, ..., x_N\}$ in $\mathbb{R}^D$ originated from a shape $\mathcal{S}_\Theta(x_i - \mu) = 0$. We assume that each data point is independently deformed by η_i (a random variable, since the direction is along the shape gradient), conditional on being a shape point.

3.1 A BAYESIAN APPROACH

Based on the energy model (1), for $p = 1$ (for simplicity and robust properties), we can write the likelihood model as $\mathrm{P}(X|\Theta, \mu) = \frac{1}{C} \prod_{i=1}^N \mathrm{e}^{-\lambda \mathrm{E}_{\mathcal{S}_\Theta}(x_i - \mu)} = \frac{1}{C} \prod_{i=1}^N \mathrm{e}^{-\lambda |\mathcal{S}_\Theta(x_i - \mu)|}$, where C is a normalization constant and λ a constant that scale the errors/energies. The product over all points

is a consequence of the conditional independence of the deformations given the shape parameter (Θ, μ). Assuming a prior distributions on the parameters to be uniform, we conclude that the a posteriori distribution is simply the likelihood model up to a normalization, i.e.,

$$\mathrm{P}(\Theta, \mu|X) = \frac{1}{Z}\prod_{i=1}^{N} \mathrm{e}^{-\lambda \mathrm{E}_{\mathcal{S}_\Theta}(x_i-\mu)} = \frac{1}{Z}\prod_{i=1}^{N} \mathrm{e}^{-\lambda|\mathcal{S}_\Theta(x_i-\mu)|}$$

where Z is a normalization constant (does not depend on the parameters). The parameters that maximize the likelihood $\mathcal{L}(\Theta, \mu) = \log \mathrm{P}(\Theta, \mu|X) = -Z - \lambda \sum_{i=1}^{N} |\mathcal{S}_\Theta(x_i - \mu)|$ also minimize the total energy $\mathcal{E}(\Theta, \mu) = \sum_{i=1}^{N} |\mathcal{S}_\Theta(x_i - \mu)|$.

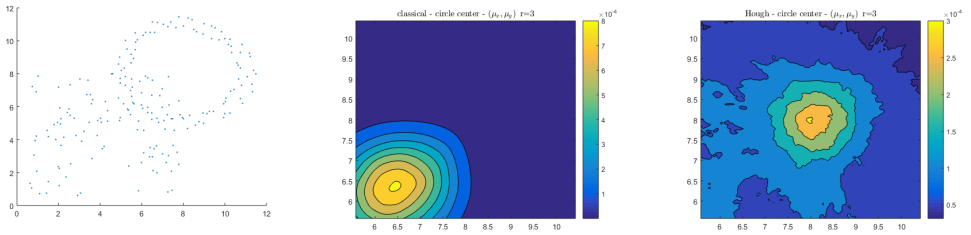
3.2 Hough Transform

A Hough transform cast binary votes from each data point. The votes are for the shape parameter values that are consistent with the data point. More precisely, each vote is given by

$$v(\Theta, \mu|x_i) = u\left(\frac{1}{\alpha} - |\mathcal{S}_\Theta(x_i - \mu)|\right) \tag{2}$$

where $u(x)$ is the Heaviside step function, $u(x) = 1$ if $x \geq 0$ and zero otherwise, i.e., $u = 1$ if $|\mathcal{S}_\Theta(x_i - \mu)| \leq \frac{1}{\alpha}$ and $u = 0$ otherwise. The parameter α clearly defines the error tolerance for a data point x_i to belong to the shape $\mathcal{S}_\Theta(x - \mu)$, the larger is α the smaller is the tolerance. One can carry out this Hough transform for center detection as a convolution process. More precisely, create a kernel, $K^H(x)$ centered at $(0, 0)$ and define it as $K^H(x) = u\left(\frac{1}{\alpha} - |\mathcal{S}_\Theta(x)|\right)$ for x in a rectangular (or square) shape that includes all x for which $u\left(\frac{1}{\alpha} - |\mathcal{S}_\Theta(x)|\right) = 1$. The Hough transform for center detection is then the convolution of the kernel with the input image. The result of the convolution at each location is the Hough vote for that location to be the center.

3.3 A comparison of the two methods



(a) Circle deformed with clutter noise.

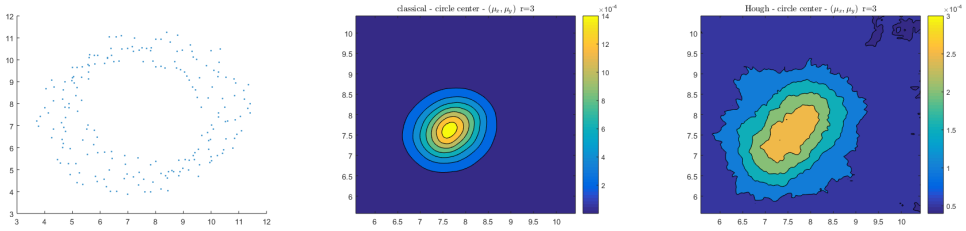
(b) Bayesian estimation

(c) Hough method

Figure 2: (a) A circle centered at $(8, 8)$, radius 3, and with 100 points, is deformed with $\eta = 0.5$. 100 clutter noise points are added uniformly to a box with diagonal corners at $(0, 0)$ and $(8, 8)$. (b) The Bayesian method (showing the probability value). The radius is fed to the method. The method mixes all data yielding the highest probability in the wrong place. increasing the parameter p can only improve a little as all data participate in the final estimation. (c) The Hough method with $\alpha = 2.769$ estimated to include all circle points that have been deformed. The method is resistant to clutter noise.

When we have one circle with deformations (e.g., see figure 1a), the Bayesian approach is just the "perfect" model. Even if noise distributed uniformly across the image is added (e.g., see figure 1b), the Bayesian method will work very well. However, as one adds clutter noise to the data (noise that is not uniform and "may correspond to clutter" in an image) as shown in figure 2, the Bayesian method mix all the data, has no mechanism to discard any data, and the Hough method outperforms the Bayesian one. Even applying robust measures, decreasing p in the energy model, will have limited effect compared to the Hough method that can discard completely the data. Consider another

scenario of two overlapping and deformed circles, shown in figure 3. Again, the Bayesian approach does not capture the complexity of the data, two circles and not just one, and end up yielding the best one circle fit in the "middle", while the Hough method cope with this data by widening the center detection probabilities (bluring the center probabilities) and thus, including both true centers. Still, the Hough method is not able to suggest that there are two circles/two peaks. In summary, the Bayesian model is always the best one, as long as the data follows the exact model generation. However, it is weak at dealing with *real world uncertainty* on the data (clutter data, multiple figures), other scenarios that occur often. The Hough method, modeled after the same true positive event (shape detection) is more robust to these data variations and for the center detection problem can be carried out as a convolution.



(a) Two overlapping circles (b) Bayesian estimation (c) Hough method

Figure 3: (a) Two overlapping circles deformed with $\eta = 0.5$, radius $r = 3$ and with 100 points each. The circle centers are at $(8, 8)$ and $(7.2, 7.2)$. (b) The Bayesian method (showing the probability value). The radius is fed to the method. The method mixes all data yielding the highest probability approximately in the "middle" of both centers and no suggestion of two peaks/circles/centers exists. (c) The Hough method with $\alpha = 2.769$ estimated to include all circle points that have been deformed. The method yields a probability that is more diluted and includes the correct centers, but does not suggest two peaks.

4 Quantum Shape Detection and Interference

Quantum theory was developed for system of particles that evolve over time. For us to utilize here the benefits of such a theory for the shape detection problem we invoke a hidden time parameter. We refer to this time parameter as hidden since the input is only one static picture of a shape. A hidden shape dynamics is not a new idea in computer vision, for example, scale space was proposed to describe shape evolution and allows for better shapes comparisons Witkin (1983); Lindeberg (1994). Hidden shape dynamics was also employed to describe a time evolution equation to produce shape-skeletons Siddiqi & Kimia (1996). Since our optimization criteria per point is given by "the energy" of (1), we refer to classic concept of action, the one that is optimized to produce the optimal path, as

$$\mathcal{A}_{0\to\Theta}^{\mu\to x}(\mathcal{P}_0^T) = T\,|\mathcal{S}_\Theta(x-\mu)|$$

where we are adopting for simplicity $p = 1$. The idea is that a shapes evolve from the center $\mu = x(t = 0)$ to the shape point $x = x(t = T)$ in a time interval T. During this evolutions all other parameters also evolve from $\Theta(t = 0) = 0$ to $\Theta(t = T) = \Theta$. The evolution is reversible, so we may say equivalently, the shape point x contracts to the center μ in the interval of time T.

Following the path integral point of view of quantum theory Feynman & Hibbs (2012), we consider the wave propagation to evolve by the integral over all path

$$\psi_0(\mu) = \int \mathrm{d}\mathcal{P}_0^T\, \mathcal{K}(\mathcal{P}_0^T)\, \psi_\Theta(x) \tag{3}$$

where $\psi_{\Theta(t)}(x(t))$ is the probability amplitude that characterize the state of the shape, $\mathcal{P}_0^T$ is a path of shape contraction, from an initial state $(x(0), \Theta(0)) = (x, \Theta)$ to a final state $(x(T), \Theta(T)) = (\mu, 0)$. The integral is over all possible paths that initialize in $(x(0), \Theta(0)) = (x, \Theta)$ and end in $(x(T), \Theta(T)) = (\mu, 0)$. The Kernel $\mathcal{K}$ is of the form

$$\mathcal{K}(\mathcal{P}_0^T) = \frac{1}{C}\, \mathrm{e}^{\mathrm{i}\frac{1}{\hbar}\,\mathcal{A}_{\Theta\to 0}^{x\to\mu}(\mathcal{P}_0^T)} = \frac{1}{C}\, \mathrm{e}^{\mathrm{i}\frac{T}{\hbar}\,|\mathcal{S}_\Theta(x-\mu)|} \tag{4}$$

where a new parameter, $\hbar$, is introduced. It has the notation used in quantum mechanics for the reduced Planck's constant, but here it will have its own interpretation and value (see section 5.1.3).

We now address the given image described by $X = \{x_1, x_2, ..., x_N\} \subset \mathbb{R}^2$ (e.g., see figure 1). We consider an empirical estimation of $\psi_\Theta(x)$ to be given by a set of impulses at the empirical data set X, i.e., $\psi_\Theta(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \delta(x - x_i)$, where $\delta(x)$ is the Dirac delta function. The normalization ensure the probability 1 when integrated everywhere. Note that $\psi_\Theta(x)$ is a pure state, a superposition of impulses. Then, substituting this state into equation (3), with the kernel provided by (4), yields the evolution of the probability amplitude

$$\psi_\Theta(\mu) \approx \sum_{i=1}^{N} \int \mathrm{d}x\, \frac{1}{C}\, \mathrm{e}^{\mathrm{i}\frac{T}{\hbar}\,|\mathcal{S}_\Theta(x-\mu)|}\, \frac{1}{\sqrt{N}}\, \delta(x - x_i) = \frac{1}{C\sqrt{N}} \sum_{i=1}^{N} \mathrm{e}^{\mathrm{i}\frac{T}{\hbar}\,|\mathcal{S}_\Theta(x_i-\mu)|}\,. \tag{5}$$

where $C = \int \mathrm{e}^{\mathrm{i}\frac{T}{\hbar}\,|\mathcal{S}_\Theta(x-\mu)|}\, \mathrm{d}\mu$. Thus shape points with deformations, x_i, are interpreted as evidence of different quantum paths, not just the optimal classical path (which has no deformation). Equation 5 is a convolution of the kernel $K(x) = \mathrm{e}^{\mathrm{i}\frac{T}{\hbar}\,|\mathcal{S}_\Theta(x)|}$ throughout the center candidates, except it is discretized at the locations where data is available. According to quantum theory, the probability associated with this probability amplitude (a pure state) is given by $\mathrm{P}(\Theta) = |\psi_\Theta(\mu)|^2$, i.e.,

$$\mathrm{P}_\Theta(\mu) = \psi_\Theta(\mu)\, \psi_\Theta^*(\mu) = \frac{1}{C^2\, N} \left(\sum_{i=1}^{N} \mathrm{e}^{\mathrm{i}\frac{T}{\hbar}\,|\mathcal{S}_\Theta(x_i-\mu)|} \right) \left(\sum_{j=1}^{N} \mathrm{e}^{-\,\mathrm{i}\frac{T}{\hbar}\,|\mathcal{S}_\Theta(x_j-\mu)|} \right),$$

which can also be expanded as

$$\mathrm{P}_\Theta(\mu) = \frac{1}{C^2\, N} \sum_{i=1}^{N} \left[1 + 2 \sum_{j>i}^{N} \cos\left[\frac{T}{\hbar}\, \left(|\mathcal{S}_\Theta(x_i - \mu)| - |\mathcal{S}_\Theta(x_j - \mu)| \right) \right] \right]. \tag{6}$$

It is convenient to define the phase $\phi_{ij} = \frac{T}{\hbar}\, \left(|\mathcal{S}_\Theta(x_i - \mu)| - |\mathcal{S}_\Theta(x_j - \mu)| \right)$.

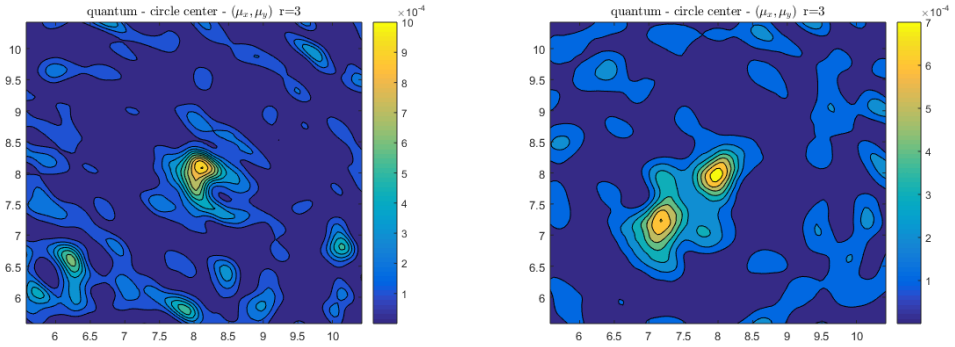
4.1 Interference

Note the interference phenomenon arising from the cosine terms in the probability (6). More precisely, a pair of data points that belongs to the shape will have a small magnitude difference, $|\phi_{ij}| \ll 1$, and will produce a large cosine term, $\cos\phi_{ij} \approx 1$. Two different data points that belong to the clutter will likely produce different phases, scaled inversely according to $\hbar$, so that small values of $\hbar$ will create larger phase difference. Pairs of clutter data points, not belonging to the shape, with large and varying phase differences, will produce quite different cosine terms, positive and/or negative ones. If an image contains a large amount of clutter, the clutter points will end up canceling each other. If an image contains little clutter, the clutter points will not contribute much. This effect can be described by the following property for large numbers: if $N \gg 1$ then $\sum_{i=1}^{N} \sum_{j=1}^{N} \cos(\epsilon_i - \epsilon_j) \approx \sqrt{N\,\pi/4} \ll N$, when each ϵ_k is a random variable.

Figure 4 shows the performance of the quantum method on the same data as shown in figure 2a and figure 3a. The accuracy of the detection of the centers and the identification of two centers shows how the quantum inspired method outperforms the classical counterparts. In figure 4a, due to interference, clutter noise cancels out (negative terms on the probability equation 6 balance positive ones), and the center is peaked. We do see effects of the noise inducing some fluctuation. In figure 4b the two circle center peaks outperform both classical methods results as depicted in figure 3. A more thorough analysis is carried out in the next section to better understand and compare the performance of these different methods.

4.2 LINEAR-COMPLEXITY COMPUTATION IN THE SIZE OF THE DATA SET

Note that even though the probability reflects a pair-wise computation as seen in (6), we evaluate it by taking the magnitude square of the probability amplitude (given by equation (5)), which is computed as a sum of N complex numbers. Thus, the complexity of the computations is linear in the data set size. After all, it is a convolution process.



(a) Quantum Probability on figure 2a

(b) Quantum Probability on figure 3a

Figure 4: Quantum Probability depicted for input shown in figure 2a and figure 3a, respectively. The parameters used where $T = 1$, $\hbar = 0.12$. The quantum method outperform the classical methods, as the center detection shown in (a) is more peaked than in the Hough method and in (b) the two peaks emerge. These are results of the interference phenomena, as cancellation of probabilities (negative terms on the probability equation 6) contribute to better resolve the center detection problem.

4.3 A CLASSICAL STATISTICAL VERSION OF THE QUANTUM CRITERION

we derive a classical probability from the quantum probability amplitude via the Wick rotation Wick (1954). It is a mathematical technique frequently employed in physics, which transforms quantum physical systems into statistical physical systems and vice-versa. It consists in replacing the term $\mathrm{i}\,\frac{T}{\hbar}$ by a real parameter α in the probability amplitude. Considering the probability amplitude equation (5), the Wick rotation yields

$$\mathrm{P}_{\Theta}(\mu) = \frac{1}{Z}\sum_{i=1}^{N} \mathrm{e}^{-\,\alpha\,|\mathcal{S}_{\Theta}(x_i-\mu)|}\,. \tag{7}$$

We can interpret this as follows. Each data point x_i produces a vote $v(\Theta, \mu|x_i) = \mathrm{e}^{-\,\alpha\,|\mathcal{S}_{\Theta}(x_i-\mu)|}$, with values between 0 and 1. The parameter α controls the weight decay of the vote. Interestingly, this probability model resembles a Hough transform with each vote (7) being approximated by the binary vote described by (2)

5 ANALYSIS FOR THE CIRCLE SHAPE

We analyze the quantum method described by the probability (6), derived from the amplitude (5), and compare it with the classical statistical method described by (7) and its approximation (2). This analysis and experiments is carried for a simple example, the circle.

We consider the circle shape, $\mathcal{S}_{r^*}(x-\mu) = 1 - \frac{(x-\mu)^2}{(r^*)^2}$ of radius r^* and its evaluation not only at the true center μ^* but also at small displacements from it $\mu = \mu^* + \delta\mu$ where $\frac{\delta\mu}{r^*} < 1$ with $\delta\mu = |\delta\mu|$.

5.1 QUANTUM PATHS AS DEFORMATIONS

The points of an original circle are deformed to create the final "deformed" circle shape. Each point is moved by a random vector η_i pointing along the radius, i.e., $\eta_i = \eta_i \hat{r}_i^*$ with $\hat{r}_i^*$ being the unit

vector along the radius. Thus, we may write each point as $x_i = x_i^C + \eta_i$ so that x_i^C belong to the original circle or $\mathcal{S}_{r^*}(x_i^C - \mu^*) = 0$ The deformation is assumed to vary independently and uniformly point wise. Thus, $\eta_i \in (-\eta, \eta)$ and $\mathrm{P}(\eta_i) = \frac{1}{2\eta}$. Plugging in the deformations and center displacement into the shape representation, $\mathcal{S}_i = \mathcal{S}_{r^*}(x_i - \mu)$, we get

$$\mathcal{S}_i = 1 - \frac{(x-\mu)^2}{(r^*)^2} = 1 - \frac{(x_i^C + \eta_i - \mu^* - \delta\mu)^2}{(r^*)^2} = -[2(a_i + b_i) + (a_i^2 + b^2 - 2a_i b_i)] \quad (8)$$

where $\delta\mu_i = \hat{r}_i^* \cdot \delta\mu$, $a_i = \frac{\eta_i}{r^*}$ and $b_i = \frac{\delta\mu_i}{r^*}$, $b = \frac{\delta\mu}{r^*} = \max_i b_i$. We are investigating small deformations $|a_i| < 1$ and small displacements from the true center $|b_i| < 1$. We interpret $|\mathcal{S}_i|$ as a random variable and we assumed the sampling of the circle is such that the variation of b_i is uniform. So $\mathrm{P}_b(b_i) = \frac{1}{2b} = \frac{r}{2\delta\mu}$ and with the deformations uniformly distributed we also have $\mathrm{P}_a(a_i) = \frac{1}{2a} = \frac{r}{2\eta}$. Finally, since the deformations are independent of the shape position, a_i and b_i are independent and

$$\mathrm{P}_{\mathcal{S}}(|\mathcal{S}_i|) = \mathrm{P}_a(a_i)\mathrm{P}_b(b_i) = \frac{1}{4ab} = \frac{(r^*)^2}{4\eta\,\delta\mu}$$

For the special case of the evaluation of the shape at the true center, $\delta\mu = 0$ we obtain $\mathcal{S}_i = -2\frac{\eta_i}{r^*} - \frac{\eta_i^2}{(r^*)^2} = -[2a_i + a_i^2]$. The action for each path is given by $|\mathcal{S}_\Theta(x_i - \mu)|$ and we have multiple paths samples from the data. Note that when we apply the quantum method, we interpret data derived from shape deformation as evidence of quantum trajectories (paths) that are not optimal while the classical interpretation for such data is a statistical error/deformation. Both are different probabilistic interpretations that lead to different methods of evaluations of the optimal parameters as we analyze next.

5.1.1 Quantum Probability Amplitude

We interpret the probability amplitude in equation (5), the sum over $i = 1, \ldots, N$, as a sum over many independent samples. In general given a function $f(|\mathcal{S}|)$, then the sum over all points, $\sum_{i=1}^{N_C} f(|\mathcal{S}_i|)$, can be interpreted as N_C times the statistical average of the function f over the random variable $\mathcal{S}_i$. In this case, the random variable $\mathcal{S}_i(\eta_i, \delta\mu_i)$ represent two independent and uniform random variables, $(\eta_i, \delta\mu_i)$, or (a_i, b_i).

Inserting shape equation (8) into the quantum probability amplitude of equation (5) result in

$$\psi_{r*}(\mu^* + \delta\mu) \approx \frac{1}{C\sqrt{N_C}} \sum_{i=1}^{N_C} \mathrm{e}^{\mathrm{i}\,\frac{1}{\hbar}\,|\mathcal{S}_{r^*}(x_i - \mu)|} = \frac{\sqrt{N_C}}{C}\mathcal{I}_{ab}(\mathrm{e})$$

where $\mathcal{I}_{ab}(\mathrm{e}) = \frac{1}{4ab}\int_{-a}^{a}\mathrm{d}a_i \int_{-b}^{b}\mathrm{d}b_i\,\mathrm{e}^{\mathrm{i}\,\frac{1}{\hbar}\,\left|2(a_i+b_i)+(a_i^2+b^2-2a_ib_i)\right|}$ and at the true center we get $\psi_{r*}(\mu^*) \approx \frac{\sqrt{N_C}}{C}\mathcal{I}_a(\mathrm{e})$, where $\mathcal{I}_a(\mathrm{e}) = \frac{1}{2a}\int_{-a}^{a}\mathrm{d}a_i\,\mathrm{e}^{\mathrm{i}\,\frac{1}{\hbar}\,|2a_i+a_i^2|}$. The ratio of the probabilities (magnitude square of the probability amplitudes) for the circle is then given by

$$Q_C(a, b, \hbar) = \frac{\mathrm{P}_{r*}(\mu^*)}{\mathrm{P}_{r*}(\mu^* + \delta\mu)} \approx \frac{|\mathcal{I}_a(\mathrm{e})|^2}{|\mathcal{I}_{ab}(\mathrm{e})|^2} \quad (9)$$

These integrals can be evaluated numerically (or via integration of a Taylor series expansions, and then a numerical evaluation of the expansion).

5.1.2 Classical Hough Transform

Inserting shape equation (8) into the vote for the Hough transform giving by equation (2) result in

$$v(r^*, \mu^* + \delta\mu|x_i) = u\left(\frac{1}{\alpha} - |2(a_i + b_i) + (a_i^2 + b^2 - 2a_ib_i)|\right)$$

and interpreting the Hough total vote, $V^H_{r*}(\mu^*) = \sum_{i=1}^{N_C} v(r^*, \mu^*|x_i)$, as an average over a function of the random variable $|\mathcal{S}_i|$ multiplied by the number of votes, we get

$$V^H_{r*}(\mu^* + \delta\mu) \approx N_C \mathcal{I}_{ab}(u) \quad \Rightarrow \quad V^H_{r*}(\mu^*) \approx N_C \mathcal{I}_a(u)$$

where $\mathcal{I}_{ab}(u) = \frac{1}{4ab}\int_{-a}^{a} \mathrm{d}a_i \int_{-b}^{b} \mathrm{d}b_i \, u\left(\frac{1}{\alpha} - \left|2(a_i + b_i) + (a_i^2 + b^2 - 2a_i b_i)\right|\right)$ and at the true center $\mathcal{I}_a(u) = \frac{1}{2a}\int_{-a}^{a} \mathrm{d}a_i \, u\left(\frac{1}{\alpha} - \left|2a_i + a_i^2\right|\right)$. The ratio of the votes at the true center and the displaced center is then given by

$$H(a, b, \alpha) = \frac{V^H_{r*}(\mu^*)}{V^H_{r*}(\mu^* + \delta\mu)} = \frac{\mathcal{I}_a(u)}{\mathcal{I}_{ab}(u)} \tag{10}$$

We now address the choice of hyper-parameters of the models, namely $\hbar$ for the quantum model and α for the classical counterpart so that the detection of the true center is as accurate as possible.

5.1.3 Evaluation of $\hbar$ and α for Center Detection μ

In this section, without loss of generality, we set $T = 1$. In this way we can concentrate on understanding $\hbar$ role (and not $\frac{T}{\hbar}$). The amplitude probability given by equation (5) has a parameter $\hbar$ where the inverse of $\hbar$ scales up the magnitude of the shape values. The smaller is $\hbar$, the more the phase $\varphi_i = \frac{1}{\hbar}|\mathcal{S}_\Theta(x_i - \mu)|$ reaches any point in the unit circle. A large $\hbar$ can make the phase φ_i very small and a small $\hbar$ can send each shape point to any point in the unit circle .

The parameter $\hbar$ large can help in aligning shape points to similar phases. That suggests $\hbar$ as large as possible. At the same time, $\hbar$ should help in misaligning pair of points where at least one of them does not belong to the shape. That suggests small values of $\hbar$. Similarly, if one is evaluating a shape with the "wrong" set of parameters, we woulud want the shape points to cancel each other. In our example of the circle, we would like that shape points evaluated at a center displacement from the true center to yield some cancellation. That suggests $\hbar$ small. One can explore the parameter $\hbar$ that maximize the ratio $Q_C(a, b, \hbar)$ given by equation (9). We can also attempt to analytically balance both requests (high amplitude at the true center and low amplitude at the displaced center) by choosing values of $\hbar$ such that $\varphi_i = \frac{1}{\hbar}|\mathcal{S}_{r^*}(x_i - \mu^*)| \leq \pi \; \forall i = 1, \ldots, N_C$. More precisely, by choosing

$$\hbar = \frac{1}{\pi} \max_i |\mathcal{S}_{r^*}(x_i - \mu^*)| \approx \frac{2a + a^2}{\pi} \tag{11}$$

Figure 5 suggests this choice of $\hbar$ gives high ratios for $Q(a, b, \hbar)$.

Now we discuss the estimation of α. we note that for $\frac{1}{\alpha} > 2a + a^2 = 2\frac{\eta}{r^*} + \left(\frac{\eta}{r^*}\right)^2$ all shape points will vote for the true center. Thus, choosing α so that largest value of its inverse is $\frac{1}{\alpha} = 2a + a^2$ will guarantee all votes and give a lower vote for shape points evaluated from the displaced center. One could search for higher values of α, smaller inverse values, so that reducing votes at the true center and expecting to reduce them further at the displaced center, i.e., to maximize $H(a, b, \alpha)$ in equation (10). Figure 5 suggests such changes do not improve the Hough transform performance.

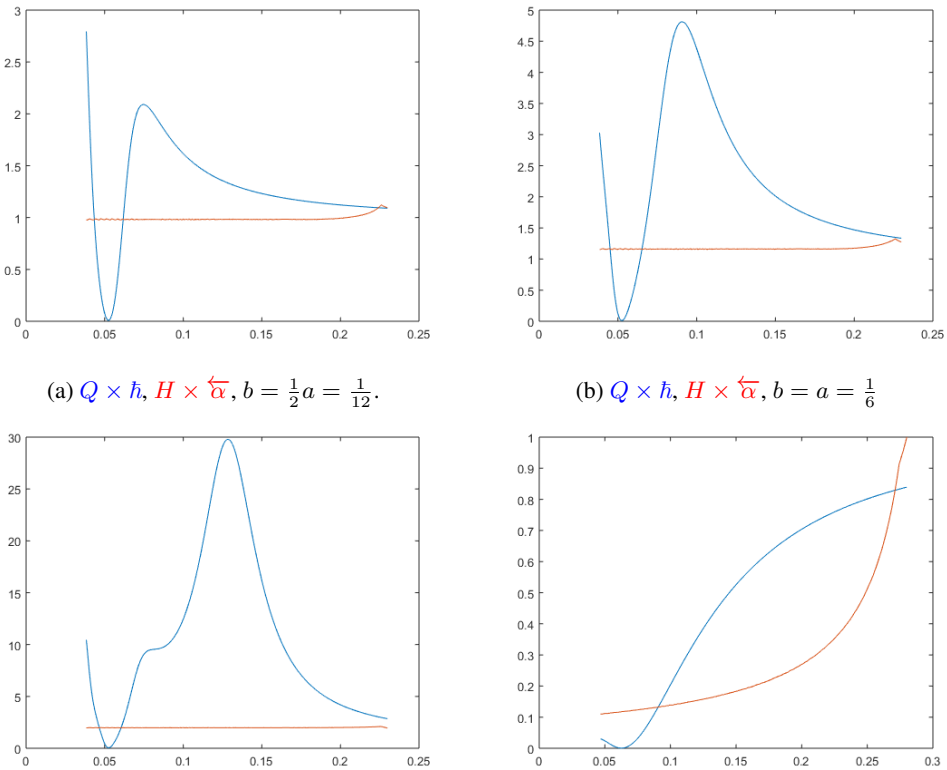
Figure 5 demonstrates that the quantum method outperforms the classical Hough transform on accuracy detection.

We can also perform a similar analysis adding noise, to obtain similar results. This will require another two pages (a page of analysis and a page of graphs), and if the conference permits, we will be happy to add.

6 Conclusions

Deep Convolution Neural Networks (CNNs), rooted on the pioneer work of Rumelhart et al. (1986); LeCun (1985); Krizhevsky et al. (2012), and summarized in LeCun et al. (2015), have been shown to be very useful in a variety of fields.

Inspired in quantum theory, we investigated the use of complex value kernel functions, followed by the local non-linear absolute (modulus) operator square. We studied a concrete problem of



(a) $Q \times \hbar$, $H \times \overleftarrow{\alpha}$, $b = \frac{1}{2}a = \frac{1}{12}$.

(b) $Q \times \hbar$, $H \times \overleftarrow{\alpha}$, $b = a = \frac{1}{6}$

(c) $Q \times \hbar$, $H \times \overleftarrow{\alpha}$, $b = 2a = \frac{1}{3}$.

(d) $|\psi|^2(\mu^*) \times \hbar$ and $V(\mu^*) \times \overleftarrow{\alpha}$

Figure 5: Quantum method vs classical Hough transform for accuracy of the detection of the center. For all figures we fixed the radius $r^* = 3$ and deformations $\eta = 0.5$, thus, $a = \frac{1}{6}$. For each of the figures 5a, 5b,5c we vary we vary $b = \frac{1}{2}a, a, 2a$ (or center displacements $\delta\mu = 0.25, 0.5, 1$), respectively. These figures depict ratios $Q(a, b, \hbar) \times \hbar$ (blue) for $\hbar \in (0.047, 0.2802)$ and $H(a, b, \alpha) \times \overleftarrow{\alpha}$ (red) for $\overleftarrow{\alpha} \in (22.727,\ 2.769)$ (The reverse arrow implies the x-axis start at the maximum value and decreases thereafter). All plots have 200 points, with uniform steps in their respective range. Note that our proposed parameter value is $\hbar = 0.1401$, the solution to equation (11), and indeed gives a high ratio. Also, $\alpha = 2.769$ is the smallest value to yield all Hough votes in the center. Clearly the quantum ratio outperforms the best classical Hough method, which does not vary much across α values. As the center displacement increases, the quantum method probability, for $\hbar = 0.1401$, decreases much faster than the Hough method probability. Final figure 5d display values of $|\psi|^2(\mu^*) \times \hbar$ (at the true center) in blue, for $\hbar \in (0.047, 0.2802)$, with 200 uniform steps. In red, $V(\mu^*) \times \overleftarrow{\alpha}$ for $\overleftarrow{\alpha} \in (22.727,\ 2.769)$, with 200 uniform steps.

shape detection and showed that when multiple overlapping shapes are deformed and/or clutter noise is added, a convolution layer with quantum inspired complex kernels outperforms the statistical/classical kernel counterpart and a "Bayesian shape estimator". It is worth to mention that the Bayesian shape estimator is the best method as long as the data satisfy the model assumptions. Once we add multiple shapes, or add clutter noise (not uniform noise), the Bayesian method breaks down rather easily, but not the quantum method nor the statistical version of it (the Hough method being an approximation to it). An analysis comparing the Quantum method to the Hough method was carried out to demonstrate the superior accuracy performance of the quantum method, due to the quantum phenomena of interference, not present in the classical CNN.

We have not focused on the problem of learning the shapes here. Given the proposed quantum kernel method, the standard techniques of gradient descent method should also work to learn the kernels, since complex value kernels are also continuous and differentiable. Each layer of the networks carries twice as many parameters, since complex numbers are a compact notation for two numbers,

but the trust of the work is to suggest that they may perform better and reduce the size of the entire network. These are just speculations and more investigation of the details that entice such a construction are needed. Note that many articles in the past have mentioned "quantum" and "neural networks" together. Several of them use Schrödinger equation, a quantum physics modeling of the world. Here in no point we visited a concept in physics (forces, energies), as Schrödinger equation would imply, the only model is the one of shapes (computer vision model). Quantum theory is here used as an alternative statistical method, a purely mathematical construction that can be applied to different models and fields, as long as it brings benefits. Also, in our search, we did not find an article that explores the phenomena of interference and demonstrate its advantage in neural networks. The task of brining quantum ideas to this field must require demonstrations of its utility, and we think we did that here.

REFERENCES

R. Feynman. *The Feynman Lectures on Physics*, volume 3. Addison Wesley, 1971.

R. Feynman and A. Hibbs. *Quantum Mechanics and Path Integrals: Emended by D. F. Steyer*. Dover Publications, 2012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1097–1105. 2012.

Y. LeCun. A learning scheme for assymetric threshold network, 1985.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning, 2015.

T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics (Supplement on Advances in Applied Statistics: Statistics and Images: 2)*, 2 (21):224270, 1994.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg. *SSD: Single Shot MultiBox Detector*, pp. 21–37. Springer International Publishing, 2016.

D. Rumelhart, G. Hinton, and R. Williams. Learning representations by back-propagating error, 1986.

K. Siddiqi and B. B. Kimia. A shock grammar for recognition. *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR '96*, 1996.

G. C. Wick. Properties of Bethe-Salpeter wave functions. *Phys. Rev.*, 96:1124–1134, Nov 1954.

A. P. Witkin. Scale-space filtering. *Proc. 8th Int. Joint Conf. Art. Intell.*, pp. 10191022, 1983.