Gaussian Processes for Digits Recognition

Data Science and Scientific Computing

Dipartimento di Matematica e Geoscienze

Authors:
Daniel ORZAN
Gabriele RUGGERI

SVHN Dataset

The Street View House Number (SVHN) is a digit classification benchmark dataset that contains 99289 32×32 RGB images of printed digits (from 0 to 9) cropped from pictures of house number plates. The cropped images are centered in the digit of interest, but nearby digits and other distractors are kept in the image. SVHN has three sets: training (73257), testing (26032) and an extra set helping with the training process, that however was not used in the experiment.

SVHN Complexity

The dataset is a generalized and more complex version of the MNIST dataset, where the complexity is given by the fact that the pictures come from a real world context and so they suffer from the typical problems of images:

- Distrorsion
- Rotations
- Poor quality
- Neaby Digits

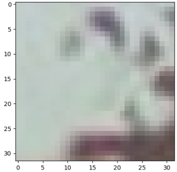


Figure: noisy image

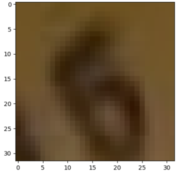


Figure: rotated digit

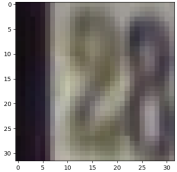


Figure: nearby digit

Goals

What is this experiment about?

The goals of this project are:

1. Study and test different preprocessing steps to get the best features out of the images
2. Build baseline classifiers to be used as benchmark for further improvements
3. Analyze different kernel for computer vision tasks and test their performances
4. Build a Gaussian Process Classifier with different kernels and compare the results with the naive classifier
5. Analyze the impact of the preprocessing pipelines on the results and draw conclusions

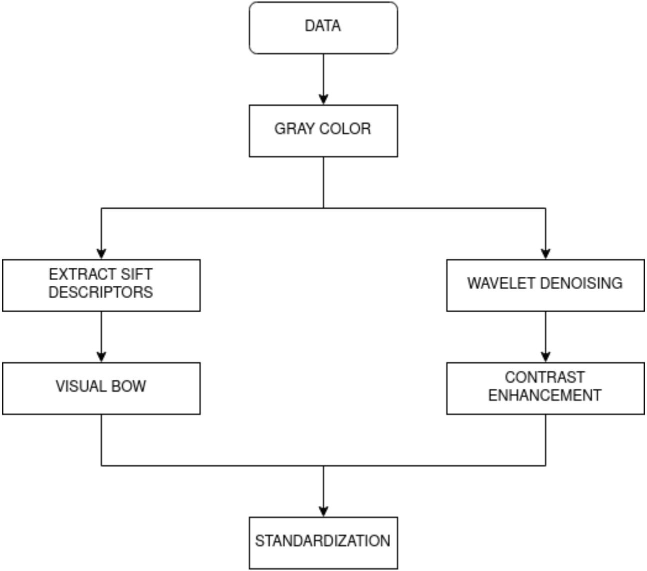
Preprocessing

- Data contain images from 10 classes, however to reduce the computations required we considered only images from class "2" and "5" turning the problem into a binary classification with a training set made of 17467 observations and a test set with 6533 ones.
- Since the classes are balanced during the model evaluation we will use accuracy as the only metric, mainly for its reliability and easy of interpretability.
- Images have 3 channels (RGB). To reduce the dimensionalily we casted them to gray scale using the formula:

$$GRAY = \frac{299}{100}R + \frac{587}{100}G + \frac{114}{100}B$$

As pointed out in [1] and [2] this is a de-facto standard which does not compromise model performances.

Preprocessing



SIFT

The *Scale Invariant Feature Transform (SIFT)* is a feature detection algorithm, developed in 1999 by D.Lowe [3], [4].

SIFT routine

1. Constructing a scale space to ensure features are scale indipendent
2. Keypoint localization
3. Orientation assignment to ensure features are rotation indipendent
4. Keypoint descriptors are computed for every keypoint

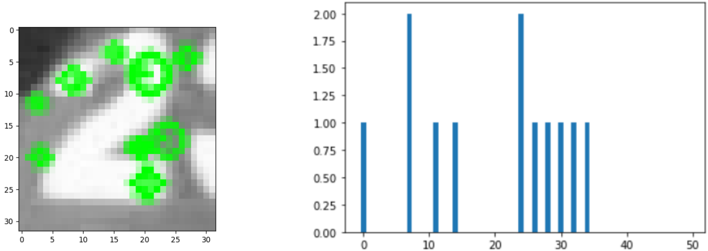


Figure: SIFT keypoints/descriptors

SIFT and Visual BoW

- The scale space is created by considering the images on different scales (octaves) generated by half-sizing the original image, then at every scale gaussian blur is applied iteratively to remove noise. At this point difference of gaussian (DoG) is used for feature enhancement; it considers subsequent differences of images at the same octave
- Keypoints detection is performed by looking at local maxima/minima by comparing each pixel with its neighboring (across octaves). Low contrast and edge keypoints are removed
- Descriptors are computed by considering magnitude and orientation of neighboring pixels for keypoints

Visual Bag of Words

In the end we have that every image has an N_Keypoints * 128 long feature vector. To overcome the fact that different images lie in different spaces we consider descriptors from all the images, cluster them in K groups and then every image gets represented with a K long feature vector obtained by grouping its own descriptors in one of the K clusters.

Wavelet Theory

What is a wavelet?

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\Big(\frac{x-b}{a}\Big) \quad a, b \in \mathrm{R}$$

a: scale parameter
b: translation parameter
If $a_0 = 2$, $b_0 = 1$ then there exist families of wavelets $\{\psi_{m,n}(t)\}$ that forms an orthonormal basis of $L^2(\mathrm{R})$ [5], [6].

Discrete Wavelet Transform

Given $f(t)$ discrete signal, ψ wavelets:

$$DWT(m,n) = W_\psi(m,n) =< f, \psi >= 2^{-\frac{m}{2}} \sum_{k=-\infty}^{\infty} f(k)\psi\Big(2^{-m}k - n\Big)$$

Wavelet Theory

Signal reconstruction

Given $f(t)$ discrete signal, ψ wavelets and φ scaling functions:

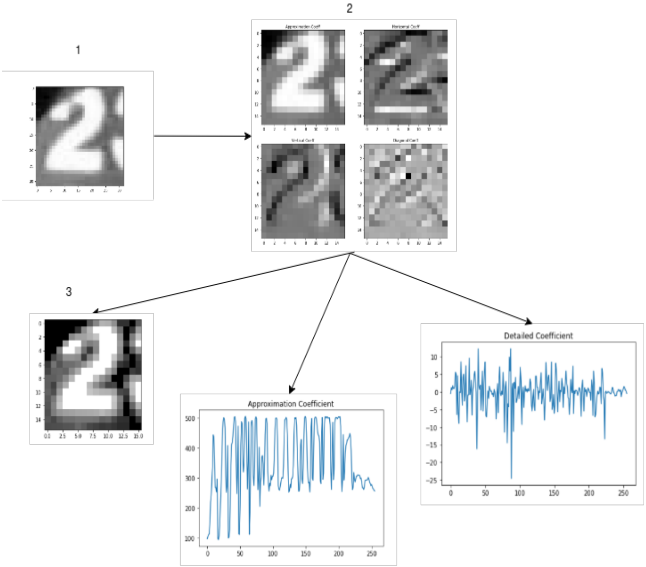
$$f(t) = \frac{1}{\sqrt{M}} \sum_n W_\varphi(m_0, n)\varphi_{m_0,n}(x) + \frac{1}{\sqrt{M}} \sum_{m=m_0}^{\infty} \sum_n W_\psi(m, n)\psi_{m,n}(x)$$

where M is a power of 2

Haar Wavelet

$$\psi(t) = \begin{cases} 1 & 0 \leq t < \frac{1}{2}, \\ -1 & \frac{1}{2} \leq t < 1, \\ 0 & \text{otherwise.} \end{cases} \qquad \varphi(t) = \begin{cases} 1 & 0 \leq t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

Wavelet pipeline



Base Models

KNN
Non parametric model with $k = 5$.

Naive Bayes
Discriminant Analysis algorithm that approximates the likelihood assuming conditional indipendence. The prediction is computed as

$$\hat{y} = arg \max_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

using Maximum A Posteriori (MAP) estimation to find $P(y)$ and $P(x_i|y)$.

The likelihood of the features is assumed to be Gaussian

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

where the parameters σ_y and μ_y are estimated using maximum likelihood.

Bayesian Logistic Regression

The goal in Bayesian logistic regression is to learn the conditional probability

$$P(y|\boldsymbol{x}, \boldsymbol{w}, b) = S(b + \boldsymbol{w}^t \boldsymbol{x}) = \frac{1}{1 + e^{-(b + \boldsymbol{w}^t \boldsymbol{x})}}$$

which was computed via Stochastic variational inference, with ELBO loss function and ADAM optimizer.

Since the data is standardised, the prior on **w** is $\mathcal{N}(\mathbf{0}, I_p)$ and it will likely converge to its true distribution thanks to the great amount of data.

Generative Model

$$\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, I_p)$$

$$b \sim \mathcal{N}(0, 1)$$

$$\hat{y} = S(b + \boldsymbol{w}^t \boldsymbol{x})$$

Guide Model

$$\mu_{w_i} \sim \mathcal{U}(0, 1), \sigma^2_{w_i} \sim \mathcal{U}(0, 1)$$

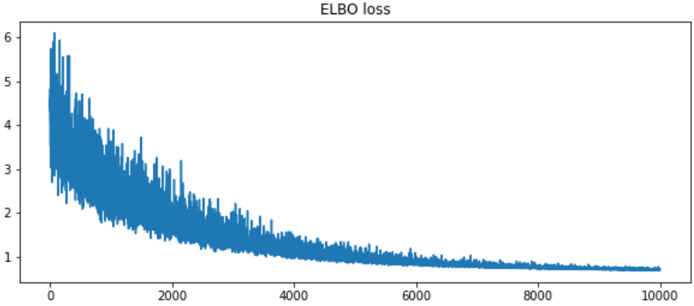
$$\mu_b \sim \mathcal{U}(0, 1), \sigma^2_b \sim \mathcal{U}(0, 1)$$

$$\boldsymbol{w} \sim \mathcal{N}(\mu_{\boldsymbol{w}}, diag(\sigma^2_{\boldsymbol{w}}))$$

$$b \sim \mathcal{N}(\mu^2_b, \sigma^2_b)$$

ELBO loss

$$ELBO(\nu) = \mathbb{E}_q[log(p(x, z)) - log(q(z, \nu))]$$



Gaussian Processes

Gaussian Processes are implemented using Laplace's approximation [2].

Laplace's method consists in approximating the posterior $p(\boldsymbol{f}|X, \boldsymbol{y})$ by a Gaussian distribution

$$q(\boldsymbol{f}|X, \boldsymbol{y}) = \mathcal{N}(\hat{\boldsymbol{f}}, (K^{-1} + W)^{-1})$$

where $W = -\nabla\nabla p(\boldsymbol{y}|\boldsymbol{f})$ [9].
When predicting on test data x_* we must compute:

$$p(f_*|X, y, x_*) = \int p(f_*|X, x_*, f) q(f|X, y) df$$

$$\pi_* = p(y_* = 1|X, y, x_*) = E_{p(f_*|X, y, x_*)}[\sigma(f_*)].$$

The hyperparameters of the kernel are optimized during the fit by maximizing the log marginal likelihood

$$\mathcal{L} = log(p(y|X)) = log \int p(y|f) p(f|X) df.$$

Kernel

Stationary/Isotropic

- **RBF**: $k(x,x') = \exp\left(- \frac{||x-x'||^2}{2l^2} \right)$
- **Matérn**: $k(x,x') = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{l} ||x-x'|| \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{l} ||x-x'|| \right)$
- **Rational Quadratic**: $k(x,x') = \left(1 + \frac{||x-x'||^2}{2\alpha l^2} \right)^{-\alpha}$

With a linear combination of kernels an improvement in classification performance is possible [7].

Non stationary

- **Dot-Product** $k(x,x') = \sigma_0^2 + x \cdot x'$

Accuracy

Base Models	**Wavelet**	**SIFT**
KNN	90.2%	75.0%
Naive Bayes	64.0%	78.8%
Bayesian Logistic Regression	63.2%	78.6%

Gaussian Processes	**Wavelet**	**SIFT**
RBF	93.1%	80.4%
Matérn	92.2%	80.4%
Rational Quadratic	93.1%	80.4%
Dot-Product	63.3%	78.8%
0.5**RBF**+0.5**Matérn**	94.6%	80.7%

Possible improvement: Pyramid Match Kernel

The Pyramid Match Kernel [8] employs multi-resolution histogram pyramid in the feature space to form a partial matching between two sets of feature vectors. This matching may be used as a robust measure of similarity to perform content-based image retrieval, as well as a basis for learning object categories.

$$K(X, Y) = \sum_{i=0}^{L} \frac{1}{2^i} N_i$$

$$N_i = \mathbb{I}(H_i(X), H_i(Y)) - \mathbb{I}(H_{i-1}(X), H_{i-1}(Y))$$

Moreover

$$K(X, Y) \in O(mL)$$

Where m is the number of features and L is the number of levels of the pyramid.

Conclusions

- We can see that performances on SIFT data are fairly constant, due to the fact that in the new feature space data are probably linearly separable.
- We get the best results on data coming from wavelet preprocessing by using Gaussian Processes with non linear kernel, in particular RBF + Matern (94.6%). On this dataset the only kernel that performed poorly was the DotProduct (linear) with $\sim 60\%$.

[1] The Effects of Image Pre- and Post-Processing, Wavelet Decomposition, and Local Binary Patterns on U-Nets for Skin Lesion Segmentation, Sara Ross-Howe, H.R. Tizhoosh, Kimia Lab, University of Waterloo, Waterloo, Ontario, Canada

[2] GAUSSIAN PROCESS MODELS FOR COMPUTER VISION, Hakeem Frank, Faculty of California State Polytechnic University, Pomona

[3] Object recognition from local scale-invariant features, Proceedings of the International Conference on Computer Vision, D.Lowe

[4] Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, D.Lowe

[5] Evaluation of wavelet transform preprocessing with deep learning aimed at palm vein recognition application, Meirista Wulandari, Basari, and Dadang Gunawan, AIP Conference Proceedings 2193, 050005 (2019)

[6] Wavelet Analysis for Image Processing, Tzu-Heng Henry Lee, Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, ROC

[7] Support Vector Machine with Mixture of Kernels for Image Classification, Tian D., Zhao X., Shi Z., IFIP Advances in Information and Communication Technology, vol 385. Springer, Berlin, Heidelberg (2012)

[8] The Pyramid Match Kernel: Efficient Learning with Sets of Features, Kristen Grauman, Trevor Darrell, Journal of Machine Learning Research 8 (2007) 725-760

[9] C. E. Rasmussen C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X. c 2006 Massachusetts Institute of Technology.