

CNN-FA: Theory and Data Presentation

Michael Person and Chris Everett

Florida Institute of Technology

Fall 2016

Outline

CNN Overview

FA Overview

Data Overview

Bibliography

CNN Overview

- ▶ CNN's are algorithms that map an input to an output
- ▶ The most common example of a CNN is when a 2D image is forward fed into the CNN and a 1D classification vector is outputted
- ▶ CNN's use many layers to generate this mapping function
- ▶ Layers are used to detect features and if the feature is found, pass it onto the next layer

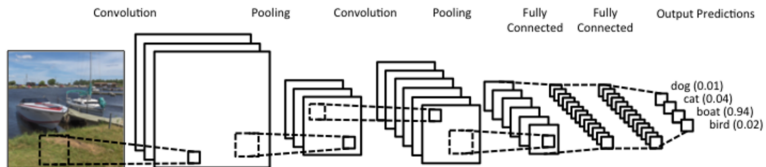
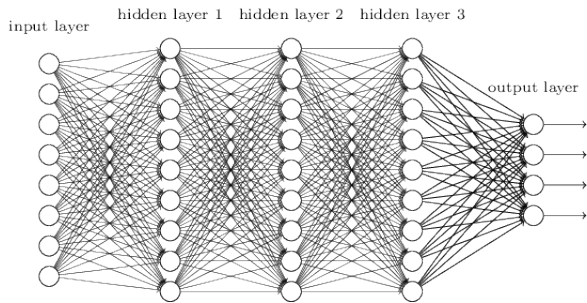
CNN History

- ▶ Due to the large computational requirements of training, accurate CNN's were not tractable until 2012
- ▶ Geoffrey Hinton life's work was the reason for the breakthrough paper in 2012
- ▶ Since 2015, there has been an emergence of more accurate Very Deep CNNs

CNN Process

- ▶ A CNN is composed of three distinct stages, a Construction stage, a Training stage, and an Operational stage
- ▶ Construction is when you generate the layering of the CNN, how many nodes there are, how the layers are connected through the nodes, the activation function to be used, and the initial parameters
- ▶ Training is when the CNN undergoes supervised learning with a set of training images using backpropagation to tune the parameters
- ▶ Operation is when the weights have been adequately set and the CNN is used to classify images

Example

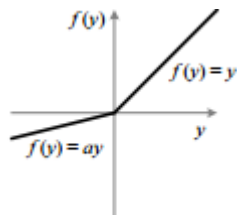
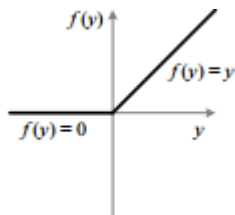


Layers

- ▶ Layers are collection of nodes at the same depth in the network. Each node in a layer performs the same operation
- ▶ Fully connected layers node's in the current layer output to every node in the next layer
- ▶ Convolutional layers perform Feature Extraction by convolving that layer's filter with the inputted feature map
- ▶ Pooling layers reduce the feature maps dimensions. The most common pooling operation is Max Pooling. Pooling both makes features translationally invariant and reduces the computational load for future layers

CNN Activation

- ▶ Features are detected from the previous layer by activation functions
- ▶ Setting the parameters that are used to determine if a neuron is activated or not is how a CNN "learns" to classify images
- ▶ Rectified Linear Units (ReLU) or Parametric Rectified Linear Units (PReLU) are the activation functions that are used because they do not suffer from vanishing gradient



Activation Function Equation

The first step is to calculate the activation function on the previous layer during the forward pass through the net.

$$z_j^l = \sum_k w_{jk}^l a_k^{l-1} + b_j^l \quad (1)$$

$$a_j^l = \sigma(z_j^l) \quad (2)$$

THIS IS A BUNCH OF PARAMETERS!!!!!!!!!!!!!! Hundreds of millions which equates to billions of flops

Backpropagation

- ▶ CNN's have many parameters that need to be optimized
- ▶ A forward pass would require as many passes through the CNN as there are parameters to optimize
- ▶ Backpropagation allows for a single backward pass of the CNN to optimize each parameter at once

Begin Backpropagation by finding the error of the current images classification.

$$\delta^l = \nabla_a C \odot \sigma'(z^L) \quad (3)$$

$$\nabla_a C \equiv \frac{\partial C}{\partial a_j^L} \quad (4)$$

Backpropage Error

Pass the error from the the current layer to the previous layer in order to find it's error via a ridiculous application of the chain rule.

$$\delta^l = (\mathbf{w}^{l+1})^T \delta^{l+1} \odot \sigma'(z^l) \quad (5)$$

After having found the errors at every node in every layer, solve for the partial derivatives of error function wrt to the parameters.

$$\frac{\partial C}{\partial b_j^l} = \delta_j^l \quad (6)$$

$$\frac{\partial C}{\partial w_{jk}^l} = a_k^{l-1} \delta_j^l \quad (7)$$

Optimization Algorithm

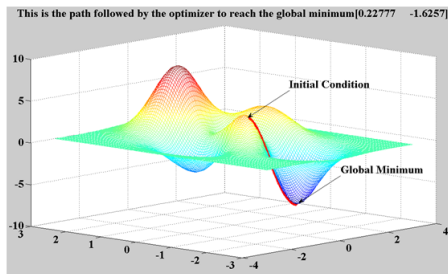
- ▶ Minimization of the error function is done by adjusting the parameters used in the activation functions
- ▶ Stochastic Gradient Descent is an efficient algorithm that uses minibatches to minimize the objective function by adjusting the parameters

$$w'_k = w_k - \frac{\eta}{m} \sum_{j=0}^{M-1} \frac{\partial C_j}{\partial w_k} \quad (8)$$

$$b'_l = b_l - \frac{\eta}{m} \sum_{j=0}^{M-1} \frac{\partial C}{\partial b_l} \quad (9)$$

SGD Discussion

- ▶ ReLU and PReLU activation functions are used because their derivatives are large and allow for fast convergence and move away from saddle points
- ▶ η is the learning rate and can be thought of the step size in your error space
- ▶ Momentum changes the learning rate to span more of the error space which makes the solution more likely to converge to a global minima instead of a local minima



Generalization

- ▶ CNN's can have problems generalizing from the training data due to overfitting of the parameters
- ▶ Overfitting is combated by dropout and data augmentation
- ▶ Dropout randomly drops a neuron out of the network in order to make each neuron rely on itself instead of on others for classification
- ▶ Data augmentation is done by image rotation, mirroring, and pixel alteration by PCA

Famous CNNs

- ▶ LeNet - 1998 (Backpropagation with Gradient Descent)
- ▶ AlexNet - 2012 (ReLU, Dropout, Data Augmentation)
- ▶ RNN - 2015 (Supervised Pretraining)
- ▶ GoogLeNet - 2014 (Inception Module)
- ▶ MSRA - 2015 (Xavier Initialization)
- ▶ (VGG - 2015 ())
- ▶ ResNet - 2015 (Residual Learning, Batch Normalization)

FA Overview

- ▶ Identifies unobserved shared factors among data sets [1]
- ▶ Created for use in psychology research by Charles Spearman as a way to reduce a complex data set of observed variables to a simpler set of unobserved factors that provide the latent information about the relationships in the data set [2]
- ▶ *For example:* Two hundred variables covering human responses to a personality test is reduced to sixteen core factors
- ▶ Two types: [1]
 - ▶ *Exploratory factor analysis (EFA)* is utilized to find the unobserved shared factors among data sets [3]
 - ▶ *Confirmatory factor analysis (CFA)* is utilized to confirm hypothesized shared factors among data sets [4]

FA Graphs

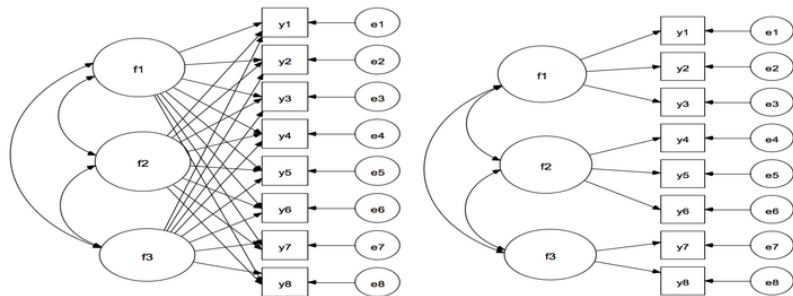


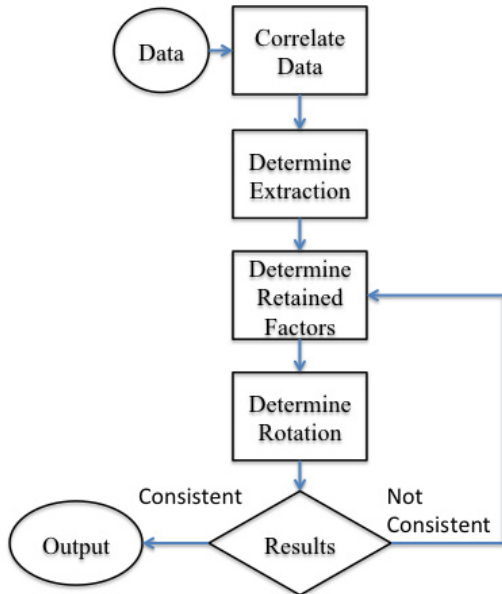
Figure 1: Factor Analysis: EFA illustrated on left and CFA illustrated on right as illustrated in [4]

FA Factors

Table 1: Factor Loading Table [5]

Variable	Factor 1	Factor 2
Income	0.65	0.11
Education	0.59	0.25
Occupation	0.48	0.19
House Value	0.38	0.60
Number of Public Parks in Neighborhood	0.13	0.57
Number of Violent Crimes per Year in Neighborhood	0.23	0.55

Exploratory Factor Analysis Process



Correlation

- ▶ The first step in the EFA process is to correlate the relationships between the variables [3].
- ▶ The correlation occurs by generating a correlation matrix or covariance matrix. If a covariance matrix is utilized then the covariance matrix is converted to a correlation matrix for the next step.
- ▶ Equation 10 illustrates a formula for computing each correlation value in a correlation matrix where the correlation matrix includes a comparison of each variable to every other variable

Correlation

		Bacteria	Pesticides	Imported_food
Correlation	Bacteria	1.000	.505	.334
	Pesticides	.505	1.000	.388
	Imported_food	.334	.388	1.000
	Tap_water	.294	.334	.301
	Food_irradiation	.317	.502	.340
	Antibiotics_food	.280	.510	.329
	Mad_cow	.357	.418	.276
	Wild_game	.378	.344	.245
	Foot_mouth	.397	.411	.274
	Food_additives	.269	.490	.395
	Bottled_water	.216	.298	.280
	GMO	.282	.497	.350
	Improper_label	.350	.475	.327
	Mercury_fish	.365	.476	.287
	Growth_hormones	.330	.540	.354
	Artificial_sweet	.257	.385	.301
	Food_packaging	.266	.442	.335

Figure 6. Truncated SPSS output for Correlation matrix. The Determinant score available below this matrix is not shown.

Correlation

$$r = \frac{N \sum(xy) - (\sum x)(\sum y)}{\sqrt{[N \sum(x^2) - \sum(x)^2][N \sum(y^2) - \sum(y)^2]}} \quad (10)$$

Where

- ▶ N is the number of pairs of scores,
- ▶ $\sum xy$ is the sum of the products of paired scores,
- ▶ $\sum(x)$ is the sum of x scores,
- ▶ $\sum(y)$ is the sum of y scores,
- ▶ $\sum(x^2)$ is the sum of squared x scores,
- ▶ $\sum(y^2)$ is the sum of squared y scores, and
- ▶ r is the correlation score.

Extraction

- ▶ The second step is to extract factors from the correlation data [3].
- ▶ *Purpose*: Determine how the various variables are related to one another by extracting the factors from the correlation matrix.
- ▶ Extraction method is utilized to determine the unobservable variables [1].
- ▶ Common extraction methods: Principal components analysis, unweighted least squares (ULS), maximum likelihood (ML), principal axis factoring (PAF), iterated principal axis factoring (iterated PAF), alpha factoring, image factoring, and Harris factoring [3].

Retain

- ▶ The third step is to determine how many factors to retain [3].
- ▶ *For example:* if the data set includes one hundred variables then five factors can be selected for extraction from the data set.
- ▶ The number of retained factors can be determined using:
 - ▶ Kaiser's criterion such that all factors with an eigenvalue above 1 are retained
 - ▶ Jolliffe's criterion such that all factors with factors above 0.70 should be retained
 - ▶ Scree test which uses a combination of eigenvalues and factors [1].

Retain Details

- ▶ As an example of Jolliffe's criterion, all of the factors in Table 1 would be discarded since all of the factors are below 0.70. In this case, either the factor limit would need to be lowered or another test such as the scree test or Kaiser's criterion would need to be utilized for the factors.
- ▶ The eigenvalue for use in Kaiser's criterion and the scree test is calculated using an eigen decomposition $AX = \lambda X$ for the square matrix of the correlation matrix such that A is the matrix, in this case the correlation matrix, λ is the eigenvalue, and X is a eigenvector [6].

Rotation

- ▶ The fourth step is to rotate the factors [3].
- ▶ *Purpose:* Distinguish the factors from each other [1].
- ▶ The rotation can be orthogonal or oblique.
 - ▶ The orthogonal rotation is designed to separate the factors by 90 degrees [3].
 - ▶ The oblique rotation is designed to correlate related factors [1].
- ▶ Figure 4 illustrates a factor matrix before and after a varimax rotation
 - ▶ Varimax rotation is a type of orthogonal rotation that "minimizes the number of variables that have high loads on each factor and works to make small loadings even smaller" [1]
 - ▶ Varimax rotation achieved the goal of clarifying the factors by separating factors 1, 2, and 3 from each other.

Rotation

Factor Matrix^a

	Factor		
	1	2	3
Growth_hormones	.741		-.352
Pesticides	.709		
GMO	.708		
Food_irradiation	.681		
Food_additives	.672		
Foot_mouth	.664	.560	
Antibiotics_food	.654		
Mad_cow	.653	.441	
Mercury_fish	.638		
Improper_label	.634		
Food_packaging	.632		
Artificial_sweet	.600		
Agroterrorism	.591		
Wild_game	.570	.472	
Bacteria	.523		
Imported_food	.510		
Bottled_water	.480		
Tap_water	.478		

Extraction Method: Principal Axis Factoring.

a. 3 factors extracted. 9 iterations required.

Rotated Factor Matrix^a

	Factor		
	1	2	3
Growth_hormones	.802		
GMO	.614		.380
Antibiotics_food	.598		.344
Mercury_fish	.519	.399	
Pesticides	.514	.344	.359
Food_additives	.505		.497
Improper_label	.490	.385	
Foot_mouth		.833	
Mad_cow		.730	
Wild_game		.707	
Agroterrorism		.471	
Bacteria		.419	
Food_packaging	.341		.578
Food_irradiation	.349		.547
Bottled_water			.534
Artificial_sweet	.406		.525
Tap_water			.486
Imported_food			.377

Extraction Method: Principal Axis Factoring.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 5 iterations.

Figure 13. SPSS output for Factor Matrix before and after Varimax rotation to illustrate how rotation aids interpretation.

Results

- ▶ The fifth step is to determine if the results from the EFA process can be replicated [3].
- ▶ The results are reviewed to ensure that the results are feasible and are within the context of the problem [1].

Data

- ▶ MNIST database [7, 8]
 - ▶ Database of handwritten digits
 - ▶ Training set: 60,000 examples with images and labels
 - ▶ Test set: 10,000 examples
- ▶ Caltech 256 database [9]
 - ▶ Database of objects
 - ▶ 257 Object Categories
 - ▶ 30,608 Images
 - ▶ Images per category range from 80 to 827

Combination

- ▶ CNN and FA could be combined to decrease the processing time in CNN [10, 11].
- ▶ Combination is illustrated in figure 5.
- ▶ Steps:
 - ▶ First, the raw data is pre-processed by EFA to reduce the dimensions of the data set by identifying the unobserved latent factors and associating the data sets with these unobserved latent factors.
 - ▶ Second, the dimensionally reduced data set is processed by CNN to classify the data set into a single dimensional vector.

CNN-FA Flow Chart

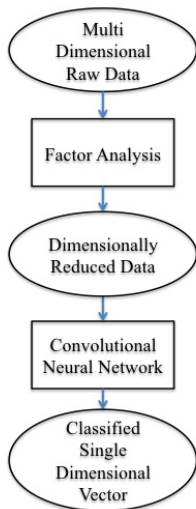











Figure 5: Proposed Combination of FA and CNN with FA being used to reduce the dimensionality of the input data before classification by CNN [11]

Bibliography I

-  A. G. Yong and S. Pearce, "A beginner's guide to factor analysis: Focusing on exploratory factor analysis," *Tutorials in Quantitative Methods for Psychology*, vol. 9, no. 2, pp. 79–94, 2013.
-  B. Thompson, *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association, 2004.
-  J. W. Osborne and E. S. Banjanovic, *Exploratory Factor Analysis with SAS*. SAS Institute, 2016.
-  Columbia University Mailman School of Public Health. Exploratory factor analysis. [Online]. Available: <https://www.mailman.columbia.edu/research/population-health-methods/exploratory-factor-analysis>

Bibliography II

-  M. Rahn. Factor analysis: A short introduction, part 1. [Online]. Available: <http://www.theanalysisfactor.com/factor-analysis-1-introduction/>
-  E. W. Weisstein. Eigenvalue. [Online]. Available: <http://mathworld.wolfram.com/Eigenvalue.html>
-  Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.
-  D. Ciregan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 3642–3649.
-  G. Griffin, A. Holub, and P. Perona, “The caltech 256,” California Institute of Technology, Tech. Rep., 2006.

Bibliography III



A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>



B. Yang, X. Fu, and N. D. Sidiropoulos, “Learning from hidden traits: Joint factor analysis and latent clustering,” *IEEE Transactions on Signal Processing*, 2016.

Bibliography