DEEP LEARNING
IMAGE & VIDEO RECOGNITION

BRUCE HO, PHD
CHIEF DATA SCIENTIST

Boston | Harrisburg | New York | San Jose | www.bigr.io

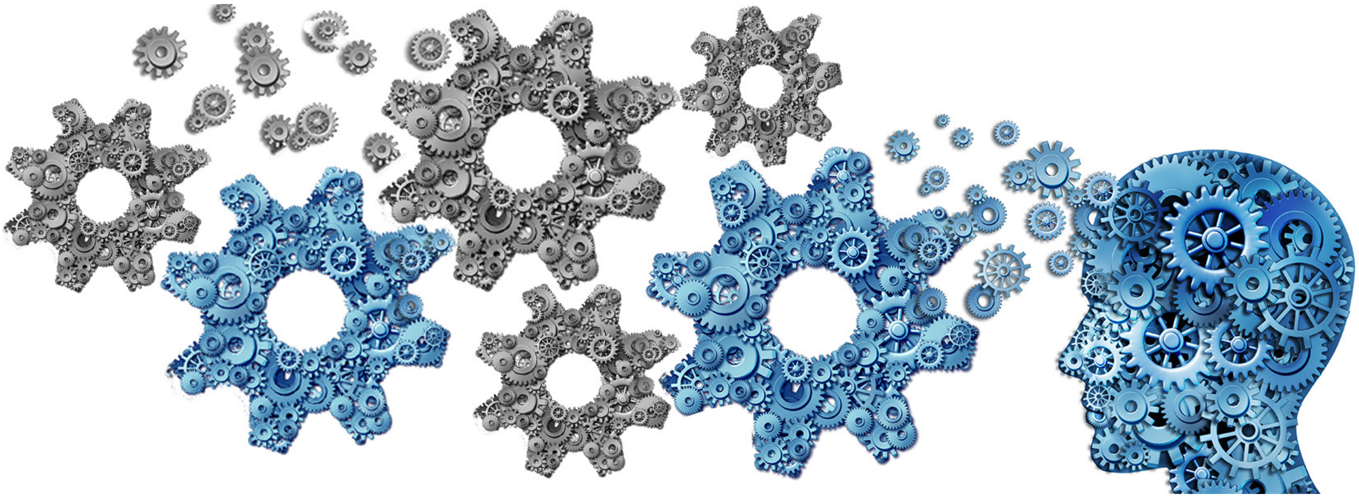
Deep Learning: Image and Video Recognition

By Bruce Ho, Chief Data Scientist

Abstract

This paper illustrates advancements in implementing Deep Neural Networks for automatic feature extraction in image and video for applications including facial recognition, programmatic video highlights, and image segmentation and object classification. Given the limitations of earlier extraction methods, these networks significantly increase accuracy, output, and available feature selection options for further analysis. BigR.io has developed solutions for the following industry use cases:

- Image Insights
- Video Highlights
- Anomaly Detection



Over the past few years, Deep Neural Network (DNN) capabilities have surpassed human parity in recognizing and interpreting images. These DNNs use Convolutional Neural Networks (CNNs) to automatically extract features from an input image with the use of convolution filters. Backpropagation then facilitates the learning by these filters of their kernel functions, starting with random values and ending up with elemental features that best represent the class of images being trained (for instance, nose, eye, and jaw shapes for face images).

Image recognition is also where the highly coveted idea of transfer learning got its early foothold. Pre-trained models based on certain categories of images can be repurposed for various classification applications using only a small dataset. Since data preparation and labeling is one of the most challenging steps when carrying out supervised learning, the impact this concept has on accelerating this process cannot be overstated. Published models and datasets by some of the biggest players in the field (Google, Microsoft, etc.) now serve as a strong starting point to build robust application-specific models for businesses with only modest means for development.

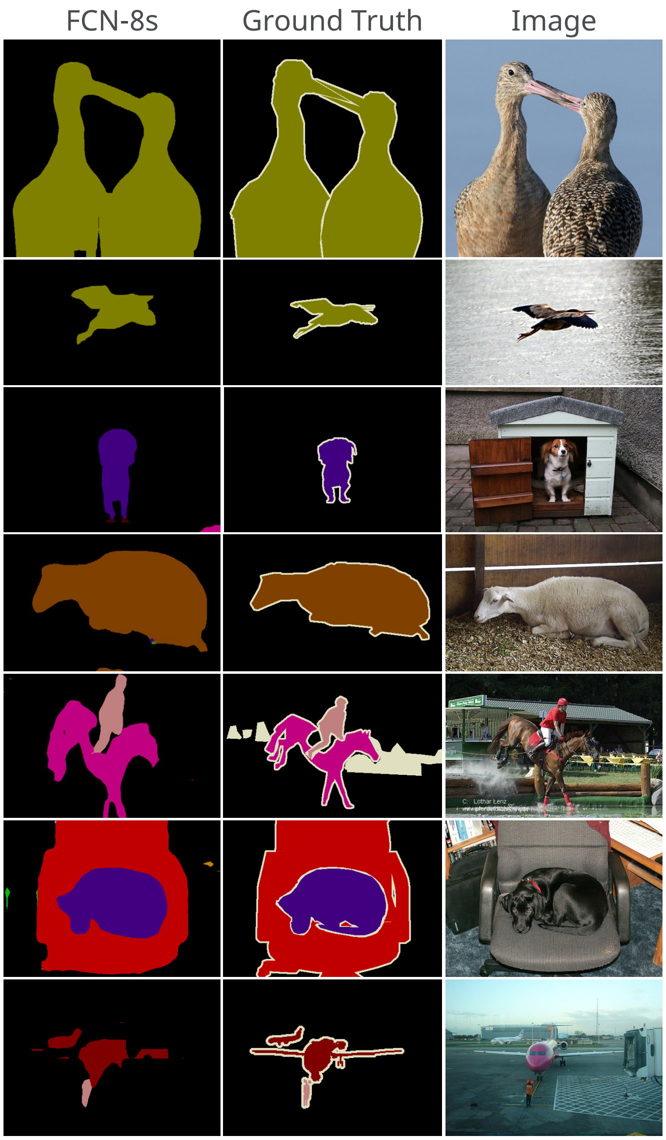
Industry Use Cases

Similar to the adoption of best practices in big data and data science across several industry verticals, image video recognition solutions affect business outcomes across diverse government agencies and businesses. In this paper, we specifically examine use cases in the security and professional sports segments, but these solutions illustrate applications across all areas of video content creation, consumption, and monitoring.

Image Insights

Image recognition can go beyond classification tasks for an entire image. In dense prediction, we are asking the neural network to detect the semantic context of any given pixel in a document or image.

CNNs work by first finding image features that resemble certain filter functions, then floating such features to a top-level representation as a translation-invariant descriptor (e.g., detection of a nose, regardless of its position within the image). By combining both coarse- and fine-grained features at different scales, we obtain both the semantic context and location information of any one pixel. This opens the door for pixel-level semantic segmentation (aka dense prediction).

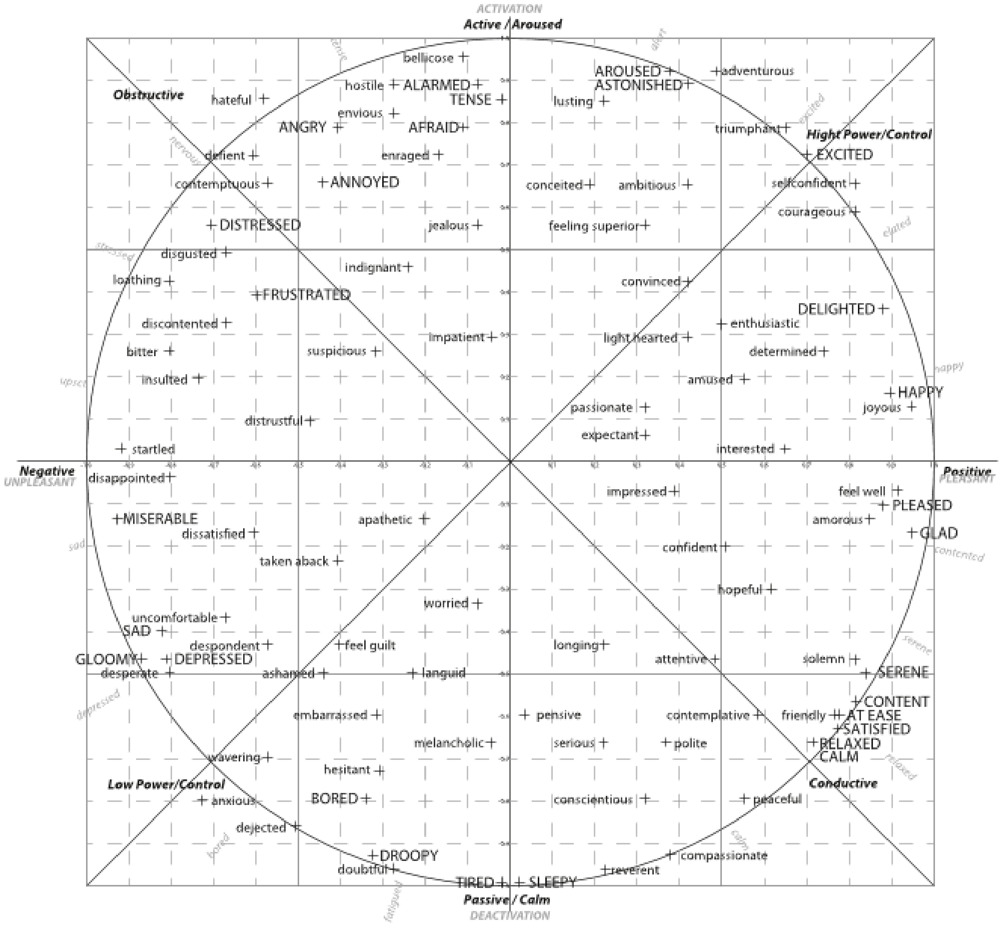


Images: *An example of using an FCN for image segmentation*

Recent work on Fully Convolutional Networks (FCNs) leverages this capability to extract semantic context of a digitized document. One could, for example, detect whether a particular pixel is a title, section header, figure caption, an image, or part of a long paragraph using FCNs. A mobile user could then easily re-layout or restyle an electronic document using the extracted semantic context. FCNs have also been successfully applied to segment parts of an image, as well as full documents, with remarkable accuracy. How does this system pick potential customers from an image of a crowd, a soccer team, or a room full of event attendees? Given a close-up face shot, is this person happy to be here, in the target age group, or giving a positive response to the last sales message? Being able to answer these audience measurement questions for marketing is one of the hot areas in need of a deep learning solution.

Many classic approaches to facial feature extraction and classification, Support Vector Machines, for example, have been devoted to this long-standing problem. Deep learning research in facial identification is relatively new but already outperforming older techniques by a wide margin. This development, and many other impressive improvements achieved by deep learning, are generally attributed to the automatic feature extraction function of neural networks and the incremental accuracy boost that deep learning techniques achieve when given a huge training dataset.

In many applications, a high-quality, close-up facial shot is not always available. Picking faces out of an ordinary action photo may be the first step before applying any facial feature analysis. For this, the region-based CNNs (R-CNNs) excel in both speed and accuracy. The R-CNN approach proposes a number of bounding boxes in the original photo using what is called Selective Search. In this method, initial object boundaries are set using a graphical pixel similarity approach. Neighboring boxes with high pixel similarity metrics are then merged to further reduce the object count. Finally, each boxed object can be classified based on a pre-trained image recognition model.



Points on the valence arousal plot can be translated to commonly understood emotions.

In other efforts, researchers have extended facial analysis to emotion detection. Classically, this simply involved image labeling where the subject exhibits a range of facial expressions and a group of volunteers would mark each as happy, sad, angry, etc. — typically up to eight emotions. More recent work also incorporates dynamic facial movements, for example, capturing the complete sequence of facial movements for a smile or frown. A more generalizable model can be developed using linear scoring along the valence-arousal graph. A prediction of valence and arousal scores on future subjects can then be interpreted using a wider range of emotion states instead of the initial selection of about eight.

Reference: *G Paltoglout, M Thelwall, Seeing Stars of Valence and Arousal in Blog Posts. Issue No. 01 Jan-Mar 2013 Vol. 4, IEEE Transactions on Affective Computing.*

Video Highlights

There are numerous highlights in every major sporting event. Manual real-time extraction of these highlights by fully attentive labelers is error-prone, requires significant manpower, is very expensive, and doesn't scale well. Furthermore, while the most recent games may benefit from manual labeling, there are years of archived footage that remain unprocessed. Most off-stats highlights are overlooked by human observers who are instructed to look for only specific events, for example, looking for a ball boy slipping while chasing a tennis ball or a Major League splitter in a Little League game.

Today, we can automate programmatic video highlights using video recognition techniques. In addition to applying CNNs to static image features, Recurrent Neural Networks (RNNs) are able to classify video segments using optical flow between image frames. This technique is easily trained not only to extract official stat events, but also to extract any interesting player motion not explicitly logged and indexed — for example, an alley-oop in basketball. Due to the automated nature of these extraction tasks, studios can come up with new ideas at any time to build upon an existing menu of highlights.

Image: *Durant eyeing Rihanna after his 3-pointer (she was cheering for LeBron).*

Going beyond sporting events, any kind of motion picture, video ad, or short-form video opens itself up for potential indexing and repurposing. For example, a DC Comics fan may want the ability to easily find all instances of girl superhero encounters within the DC universe. This task requires automatic video highlight extraction, which is the key to reviving and monetizing unlimited archive contents that would otherwise remain buried and forgotten.

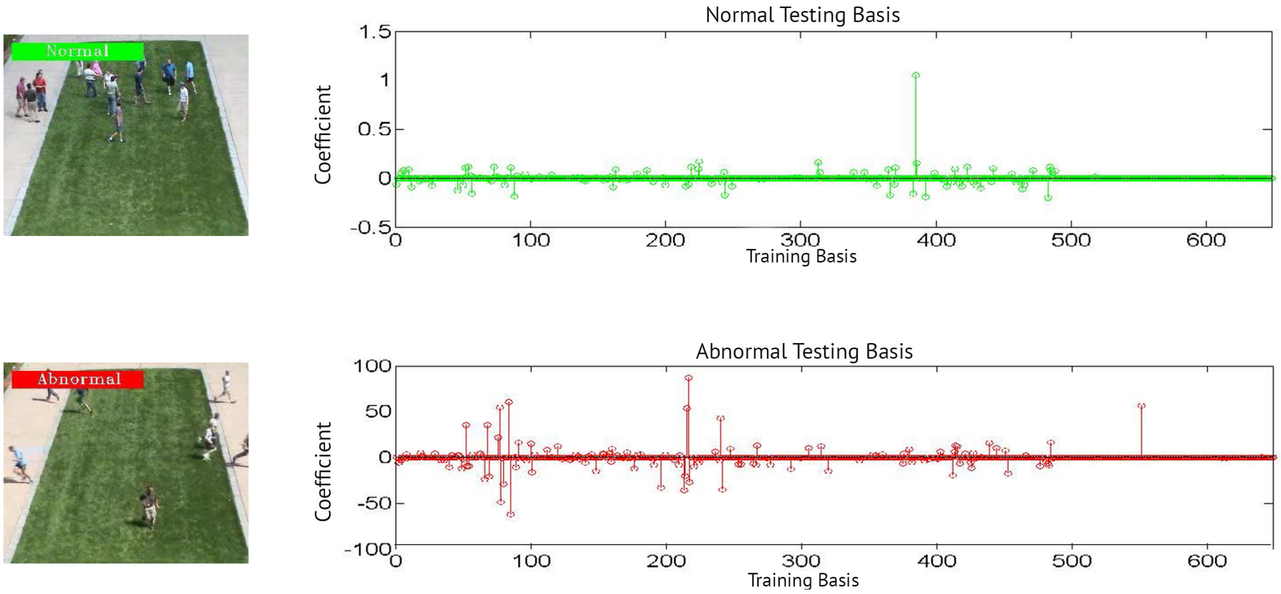
Anomaly Detection

Current practice for critical video surveillance for alarming actions is based on hired staff watching closed-circuit video. Keeping an eye on scenes that rarely change and that are highly unlikely to amount to an event of interest is a very tedious task for a human observer. Boredom leads to inattention and defeats the original purpose of catching a rapidly developing security alert as it happens. Applications from homeland security surveillance to casino monitoring all involve a constant look-out for extremely rare acts of violation. A dedicated machine trained to detect anomaly from live video can augment human monitoring and is guaranteed to never get distracted.

The optimal strategy for machine monitoring for visual anomalies generally rides on accurate and unsupervised hierarchical feature extraction. It has been shown that deep feature extraction methods are highly generalizable, effective, and have yielded state-of-the-art performances. While CNNs are a common component of initial layers in a DNN ensemble, a wide variety of approaches have been proposed for generating that final high-dimensional representation from which to best perform anomaly detection and action recognition. Many of these approaches converge on the concept of sparse coding where the representation is overcomplete, meaning the dimension is higher than needed to fully represent all input image/video features. However, the overriding advantage of sparse coding is that any given input can usually be described by a single non-zero coefficient in the representation vector. In contrast, whenever an action in the video triggers multiple non-zero coefficients, it also triggers the suspicion of an anomalous behavior.

Independent Component Analysis (ICA) is one such approach with many proposed variants. An ICA-based deep sparse feature extraction strategy combined with a non-parametric Bayesian approach can automatically determine the most optimal dimension for the latent feature vector, removing the heavy labor in parameter tuning that a full deep learning approach would entail. The reported accuracy improvement exceeds 10% over previous results.

Variants of Restricted Boltzmann Machines (RBMs) are another major direction of research for deep-sparse representation. While much progress has been made on the theoretical front, the experimental results thus far lag behind the best ICA models.



Reference: *Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2011, pp. 3449–3456*

The graph on the right is a sparse vector representation of the image on the left. The vector dimensions, called training bases, are laid out along the x-axis, with the bars representing the coefficients for the bases needed to represent the image. A normal sample (top) can be represented as a sparse linear combination of the training bases, while an anomalous sample (bottom) requires a large number of base elements.

Conclusion

Recent advancements in image and video recognition pave the way for many business applications that would have been unimaginably hard or expensive to implement before. BigR.io excels at the application of deep learning to images and electronic documents for use cases ranging from facial recognition, to programmatic video highlights, to image segmentation and object classification.

About BigR.io

BigR.io is a technology consulting firm empowering data to drive innovation and advanced analytics. We specialize in cutting-edge Big Data, Machine Learning, and Custom Software strategy, analysis, architecture, and implementation solutions. We are an elite group with MIT roots, shining when tasked with complex missions. Whether it's assembling mounds of data from a variety of sources, surfacing intelligence with Deep Learning, or building high-volume, highly-available systems, we consistently deliver.

With extensive domain knowledge, BigR.io's scientists and engineers design and build best-in-class solutions across a variety of verticals. This diverse industry exposure and our constant run-in with cutting-edge technology equips us with invaluable tools, tricks, and techniques. Our knowledge and horsepower bring innovative, cost-conscious, and extensible results to complex business challenges.

About Bruce Ho

Bruce is a top rated problem solver, communicator, and data scientist. Bruce has a unique strength in bridging software engineering with mathematical algorithms. He has over 15 years of IT experience, engaging in high-visibility, large-scale projects in marketing automation, cloud solutions, Big Data engineering and predictive analytics for top technology companies like Amazon, TeraData, and Life Technologies. He is a certified AWS solutions Architect, with a specialty in Big Data practices, particularly Spark architecture. Bruce published over 100 scientific papers during his academic career, which culminated with a Harvard faculty appointment. In his business pursuits, he continues to draw the best ideas from state-of-the-art research to fuel his Data Science practice.

His previous work includes pioneering research in hospital digitization where he served as a principal investigator for an NIH Grant to investigate efficient network transfer of high density medical images. He subsequently founded a tech startup pursuing eCommerce automation, for which he raised the funding and built up the team and operation. Since then, he has consulted widely for industries ranging from IoT and eCommerce to Ad tech, Finance, BioPharma Research, and Healthcare.

Bruce is well versed in the application of advanced Machine Learning techniques such as Dynamic Hidden Markov Model, MCMC, Vector Autoregression, and Deep Learning models including CNN, LSTM and RBM across multiple domains. His most recent fascinations are with video classification and spoken dialogue system, both of which leverage innovative use of the latest Neural Network techniques.

Bruce holds a BS in Physics from MIT, an MS in Electrical Engineering and PhD in Applied Physics, both from Caltech.