Artificial Intelligence: Business Strategies and Applications

Module 4 - Key Applications: Computer Vision and Natural Language Processing

Quick Reference Guide

Learning Outcomes

- Build an understanding of the types of problems computer vision studies
- Reflect on the rapid progressions of the computer vision and natural language processing fields and the future of the technologies
- Identify the current best performing systems, their level of performance, and axes along which you want to measure performance
- Explore advances in computer vision that are achieved by annotating images, otherwise known as providing supervision/supervised learning, as well as semi-supervised learning, which tries to reduce the need for annotation by also learning from images that aren't annotated

Computer Vision

- Computer vision studies **classification**: the task to recognize what's in an image, **object detection**: putting bounding boxes around classified objects, **instance segmentation**: pixel-level segmentation of each object, and **3-D reconstruction**: the reconstruction of the 3-D geometry of what's around us

Classification

- The ImageNet competition, a competition to build a system that can best recognize which object is in each image, led to major breakthroughs in computer vision
- The introduction of deep learning into the competition in 2012 led to the massive acceleration in progress of computer vision, to the point where error was driven down to 2.3% in 2017 and the competition was retired
- Two main reasons for the breakthrough:
    - Deep neural networks tend to excel where lots of data is available, and ImageNet was the first of such large-scale image recognition data sets
    - State-of-the-art GPUs allowed training the neural network to be shorted significantly
- **Data set augmentation**: using a small pool of labeled data to automatically generate a larger pool of labeled data

- Very good automatic augmentation leads to outperforming state of the art across many image recognition benchmarks
- **Train-test mismatch**: when a system was trained to recognize certain objects in an image, but was trained on a specific data set, like identifying cats vs. dogs, but all the training data was taken indoors. The system wouldn't perform that well when presented with outdoor test images.

Detection and Segmentation

- In **classification**, you assign a single class label to the entire image
- In **localization**, it's not enough to just say what's in the image, but you also need to put a bounding box around where an object is in the image
- In **Object Detection**, you can't just put a single bounding box around all the object, you need to put a separate bounding box around each individual object
- In **semantic segmentation**, bounding boxes aren't enough anymore, now you are expected to annotate each pixel in the image with what's in it
- In **Instance Segmentation**, if there are multiple instances of an object, you need to individually annotate them
- Neural net architectures for segmentation are a bit different than for classification. In classification we are trying to bring all information from the image together, gradually extracting local information, until finally we know what's in the image as a whole and have just a classification output at the end. But in segmentation, we need each pixel annotated with what's in that pixel. At the same time, we do want to use full image context similar to what's done in classification

3-D Reconstruction

- There are 3 main technology thrusts for 3-D reconstruction, often also referred to as depth perception: **LiDAR**, **Stereo**, and **Monocular**
- **LiDAR**: sends a laser pulse, and measures how long it takes for the laser pulse to come back after reflection of the surface to build up a depth map
  - Limits: since measurements are done by multiplying return-time with speed of light, this requires a very precise clock to get precise readings. LiDAR depends on sending light out and getting it back, and typical units have only about 50m range. LiDAR struggles to get readings on dark surfaces, as well as specular and semi-transparent surfaces
- **Stereo**: uses triangulation, which is having two cameras detecting the same object in both cameras' images and then triangulating where the object is in 3-D
  - Limits: we only have a finite number of pixels, thus it's impossible to perfectly pinpoint where an object is in the camera view. The further apart the cameras

    - are, the harder to match up points across images, as they start having very different views onto the scene
- **Monocular**: extracting depth from a single image. Computers have started to do this well and a lot of progress is being made in training large neural networks to predict the depth of each pixel

Image Generation

- **pixelCNN**: a neural network that is trained to generate the next pixel, when given all previous pixels in the image. Once done with training, we can generate images one pixel at a time, in each step sampling from a probability distribution over values for the next pixel
- **GANs (Generative Adversarial Networks)**: a generator network tries to generate images that achieve high realness score, while the image is evaluated by the discriminator network, which is trained on pairs of images, one real, one fake, to determine if the generated image is real or fake. The generator tries to fool the discriminator into thinking the generated images are real
    - Key drivers of progress in this field: (1) More compute, allowing to train larger networks for longer, and (2) Architectural improvements
- **VAEs (Variational AutoEncoders):** Two networks are trained: An encoder network, and a decoder network. The encoder network takes in an image often referred to as x, and turns it into a code z. Then the decoder is supposed to re-generate the original image x from the code z. To ensure this works, of course, the encoder has to generate a code z that contains enough info about the original image to allow to reconstruct it
    - VAEs are great at sharpening blurry images

Text Mining and Natural Language Processing

- **Information Retrieval**: searches for documents that are relevant to the question
- **Information Extraction**: embodies a number of different sub-tasks, among them Named Entity Recognition and Relationship Learning, used to extract and organize the details of a piece of text. If we can extract entities and the relationships between them, we can often discover what we are looking for without needing to read the entire document
- **NLU (Natural Language Understanding)**: a subset of NLP that both enriches the ability of systems to parse documents or break sentences into their constituent, grammatical pieces
- **NLG (Natural Language Generation):** the task of writing an answer to the question in prose

- **Automatic Speech Recognition**: takes spoken voice and converts spoken query to text
- **Text-To-Speech**: converts a written answer back to an audio stream to hear

Basics of Information Retrieval

- In the simplest model of data, the **Bag of Words**, the task of retrieval is the task of searching for documents that are likely to contain the answer to a specific question. If a document contains all the terms in a query, the documentation is probably relevant to the query meaning that the answer is more likely to appear inside. This basic model is also known as the **Vector Space Model (VSM).**
- The **TF*IDF** is an acronym used to normalize the vector representation of words in a document based on how frequently they appear and how semantically substantive they are. The TF refers to Term Frequency, the number of times a word appears in a document, and the IDF refers to Inverse Document Frequency, the number of documents in the collection that contain the given word. If document frequency is high, the word likely has little discrimination value.
- To solve the challenge of synonyms, **Latent Semantic Analysis** captures the idea that words that co-occur with common words, are perhaps likely to have the same meaning.
- **Word Embeddings**: a word in a document is characterized not by its frequency but by its word context. Word embeddings have the same effect of modeling semantically related terms.
- Given a large enough collection of documents, the word documents can begin to represent an entire language.
- **Word2vec** is a collection of word embeddings open-sourced by Google. It contains embeddings for billions of English words and their respective contexts. It captures the concept of gender, tenses, and even geographic relationships.
- However, this described approach is still limited. If you employ Word2vec in your applications, you are employing a dictionary of word embeddings generated from the universe of Google search documents. That knowledge may inaccurately reflect the semantics of highly specialized vocabularies such as medicine or finance.

Basics of Information Extraction

- Data extraction from text falls into two categories:
    - **Semi-structured data:** text documentation that loosely follows a template
        - Ex) Facebook page, Amazon listing, Yelp Reviews
    - Supervised machine learning techniques can be used to learn information extraction patterns in semi-structured data
    - **Unstructured data**: text that has no apparent, underlying template

    - Ex) a narrative text announcing a change in leadership
  - There is a branch in linguistics is all about the inherent structure in any language, and there have been many sub-fields of text analysis dedicated to relationships in unstructured text data.
  - **Information Extraction** is a general process of transforming semi-structured and unstructured text into structured data.

Basics of Language Modeling

- **Latent Dirichlet Allocation**: the model determines the latent topic structure of a document through the distribution of topics within documents and the distribution of words within topics. To write a document, LDA pulls words from the topic bags in proportion to the topical distribution in the document.
- The intuition behind word embeddings asks: “what is the probability of a word conditioned on the words surrounding it?”
- The **Continuous Bag of Words (CBOW)** model makes concrete the idea of word embeddings as a prediction. Every sentence of every document in a collection serves as a self-labeled, supervised training set. For a fixed context window of n-words before and after a target word, the model predicts the target.
- Analogous to CBOW is the **Skip-gram model** that inverts the prediction. Rather than predicting a target word based upon its context, the Skip-gram model attempts to predict the surrounding words based upon a target word. While the theory behind word embeddings was developed years earlier, the practical implementation is extremely computationally intensive.

Differences Between Language Models

- A traditional neural network and a CNN, or Convolutional Neural Network, both have a fixed size input layer and fixed size output layer.
- RNNs are different in at least two major respects. First, RNNs take in a variable number of inputs and/or outputs. More generally, they process data sequentially in the same way that humans can read variable length text from left to right. Second, unlike traditional neural networks and CNNs, the RNN has a memory. It explicitly keeps track of items that it has seen in the past as it reads forward.
- One variant of the RNN is called the **LSTM** or **Long and Short-Term Memory** model. These networks have proven particularly powerful for language tasks because in addition to carrying forward a memory of past sequence items or words, it also has an ability to shift focus. The LSTM can selectively forget items from the past as it adjusts.
- A newer generation of Neural Language Models are called transformers. **Transformers** can look both forward and backward simultaneously and can

selectively adjust their attention rather than rolling forward all memory in a single state and activation. Transformer architectures have established the state-of-the-art neural language modeling performance. They require fewer embeddings upon which to train and perform even better at transferring lessons learned from one domain to that of another.