Module 3: Fairness in AI/ML

Thursday, November 21, 2024 3:22 PM

- Word embedding (nlp)
  - A set of techniques for identifying similarities between words
  - Predicting the co-occurrence of words within a small chunk of text
  - Transforms human language meaningfully into numerical form - allows computers to understand the nuances simplicity encoded into our languages
- Word similarity & relatedness
  - Compute similarities (semantic)
  - Vectorization is the process of converting text to numbers
  - Vector space model (algebra) model for representing as a vector of identifiers in which semantically similar wards are mapped to proximate points in geometric space
  - Document occurrence - assign identifiers corresponding to count of words
  - Word context - quantify co-occurance of terms in a corpus by constructing a co- occurrence matrix which captures the number of times a tem appears in context of another term
- Cosine similarity & word analogy
  - Compute similarity between two word vectors - notion of similarity depends on what vector representation is selected
  - To solve analogy problem- c+b-a
- Word embeddings(word2vec) '
  - Stores each word as a point in space
  - Vector of a fixed number of dimensions
  - unsupervised, built by reading a huge corpus of data
  - Dimensions are projections along different axes
  - Vector space model:
    - Predict context of a given word by learning probabilities of co-occurance from a corpus - generate vectors that can predict the context of a word based on its surrounding words
  - Predicting words using context
    - Two versions: CBOW (continuous bag of words) and skip-gram.

        - CBOW - a neural network that is trained to predict which word fits in a gap
            - Predict high probability of what should fill the gap
        - Toy training data - predicting the target word given the surrounding words
        - Skip gram
            - Start with a single word embedding and tries to predict surrounding words
            - Uses words a few positions away from each center word
            - Pairs of center word/context word
            - Student passed the exam
            - Center word = yellow, others context
- Bias in word embedding
    - How to learn word 2vec embeddings
    - Start with N random 300 - dimensional vectors as initial embeddings
    - Using a machine learning classifier:
        - Take a carpus and take pairs of words that co-occur as pos-examples
        - Take pairs that want occuras neg examples
        - Train classifier to distinguish these by slowly adjusting all the embeddings to improve classifier performance
    - Cultural biases
    - Biased framings of women
    - Ethnic stereotypes
- Weat
    - Target ward set:
        - s= physics, chemistry = science
        - t= poetry, literature = arts
    - Attribute word set:
        - a= he, him, man = male
        - b= she, her, woman = female
    - Measures relative association between four concepts

1. Identify bias direction
    i. Calculate difference between
    ii. Ehe-eshe
    iii. Emale-efemale
    iv. Average

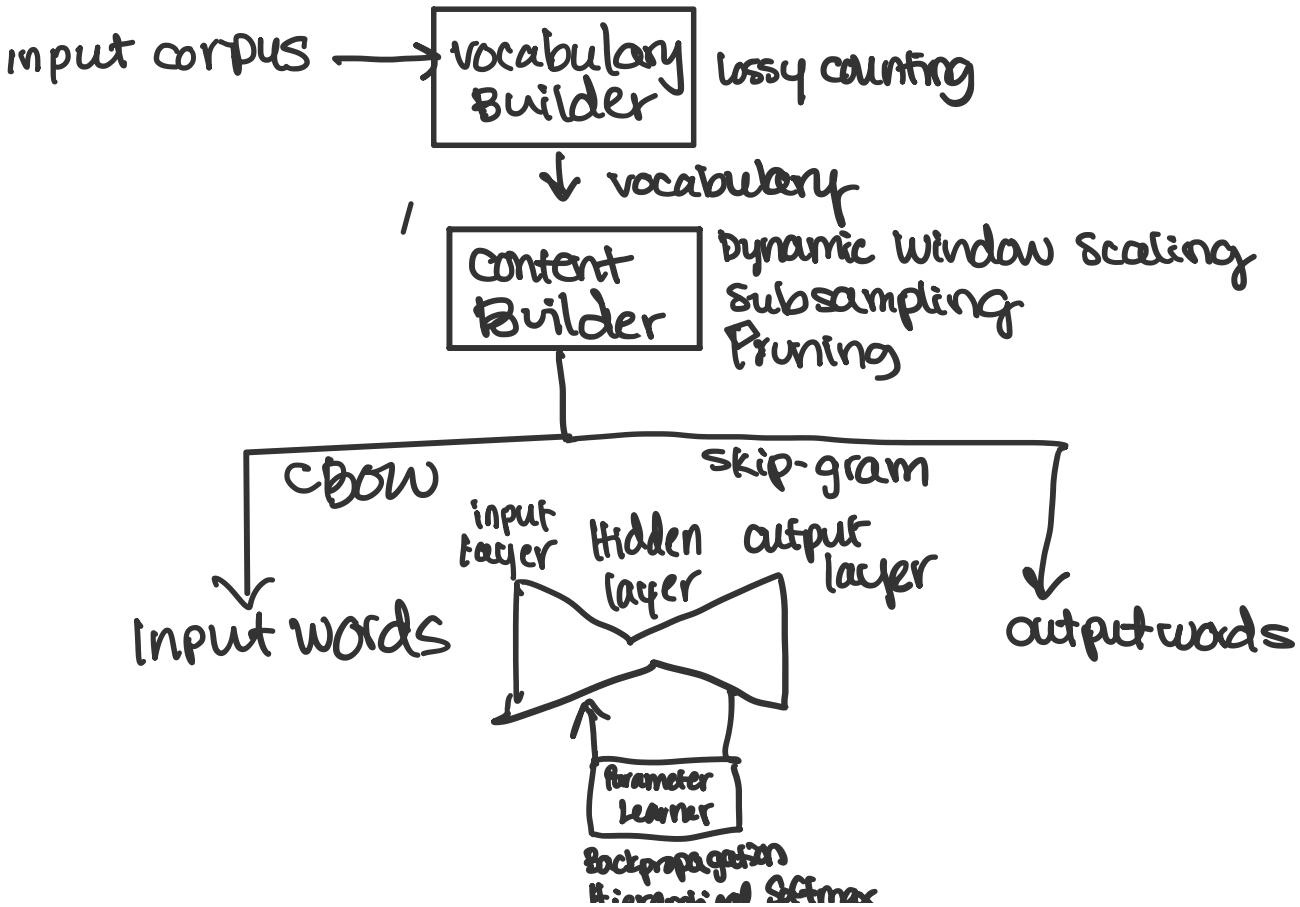
2. N

2. Neutralize - for every word that is not definitional, project to get rid of bias
3. Equalize pairs

- Facial recognition algorithms
  - Timeline of innovations
    - 1960s - facial recognition is possible
    - 1980s - 90s: advancement occurs in development of mapping and recognition software
    - 2000s: facial recognition became integrated with surveillance applications
    - 2010s: faster, portable, more powerful processors + deep learning
  - face detection: identify and locates human faces in an image regardless of positions, scale, in plane rotations orientation I prose out of plane rotation), illumination
  - two-class classification: face vs. Non-face
  - The first step: for any automatic face recognition system
  - The objective of any face detection algorithm is to locates all faces, irrespective of:
    - positioning (frontal, side new, upside down)
    - Rotation and pose
    - Occlusion
    - Resolution or image quality
    - In a single image or a sequence of images involving motion (video)
  - After face detection:
    - Image is segmented based on info and normalizes for translation, scale, rotations
    - multi-class classification: one person vs. All the others
    - Face identification: given an image that belongs to a person in a database
    - Face verification: given an image, verify whether it is from the same person it is claiming to be
    - Facial algorithms "measure" nodal points on the face
    - Face space is a theory in psychology that defines a multidimensional space in which recognizable faces are stored
    - Representation of faces within this space are according to invariant features of the face itself
  - Valentines (1991) "multidimensional face space " model

- Valentines (1991) "multidimensional face space" model
  - Human bias
    - "own race" bias
    - Twice as likely to identify own race
  - "own gender" bias
  - "own-age" bias
  - appearance-based methods (classifiers) can be trained for face recognition applications
  - Deep neural networks most common methods
- Emotions: facial recognition
  - Responding appropriately to the users emotional state will be perceived as more natural, engaging and trusting
- Ekman cross-cultural universal emotional expressions
  - Happiness
  - Sadness
  - Anger
  - Fear
  - Disgust surprise
  - Used to classify image features into one of the 6 feature categories
- Aus - Facial action units - represent distinct muscular activities
- Issues with facial recognition algorithms
  - Identification of mugshots in the wild is difficult
  - Resolution (notenoughpixels)
  - Facial pose- angulated
  - Illumination
  - Occupied facial areas
  - Results in:
    - Facial feature points not found or distorted
    - Higher errors in algorithms measurement
    - Not enough data or feature points to analyze
  - No standards exist for "acceptable" error rates meaning "success" is subjective
- Bias in facial recognition
  - Training sets are hard to get
  - If you are testing on equally non-diverse image set, you will not get real- world accuracy measurements
- What are predictive algorithms
    - Models to estimate likely outcomes based on historical

data


input corpus
vocabulary Builder
lossy counting
vocabulary
Content Builder
Dynamic Window Scaling
Subsampling
Pruning
CBOW
skip-gram
input layer
Hidden layer
output layer
input words
output words
parameter learner
Backpropagation

Neg. Sampling

Priors	demographics in law enforcement databases ≠ general population
Bias	most prominent study found several leading algorithms performed worse on underrepresented minorities, women, etc.
Consequences	If suspect is an underrepresented minority, system is more likely to erroneously fail to identify the right person causing innocent people to be investigated
Awareness	Booking Photo Comparison system is not biased against minorities because it does not see sex, race, orientation, age — incorrect in reality
No tests for Bias	there is no independent testing regime for biased error rates