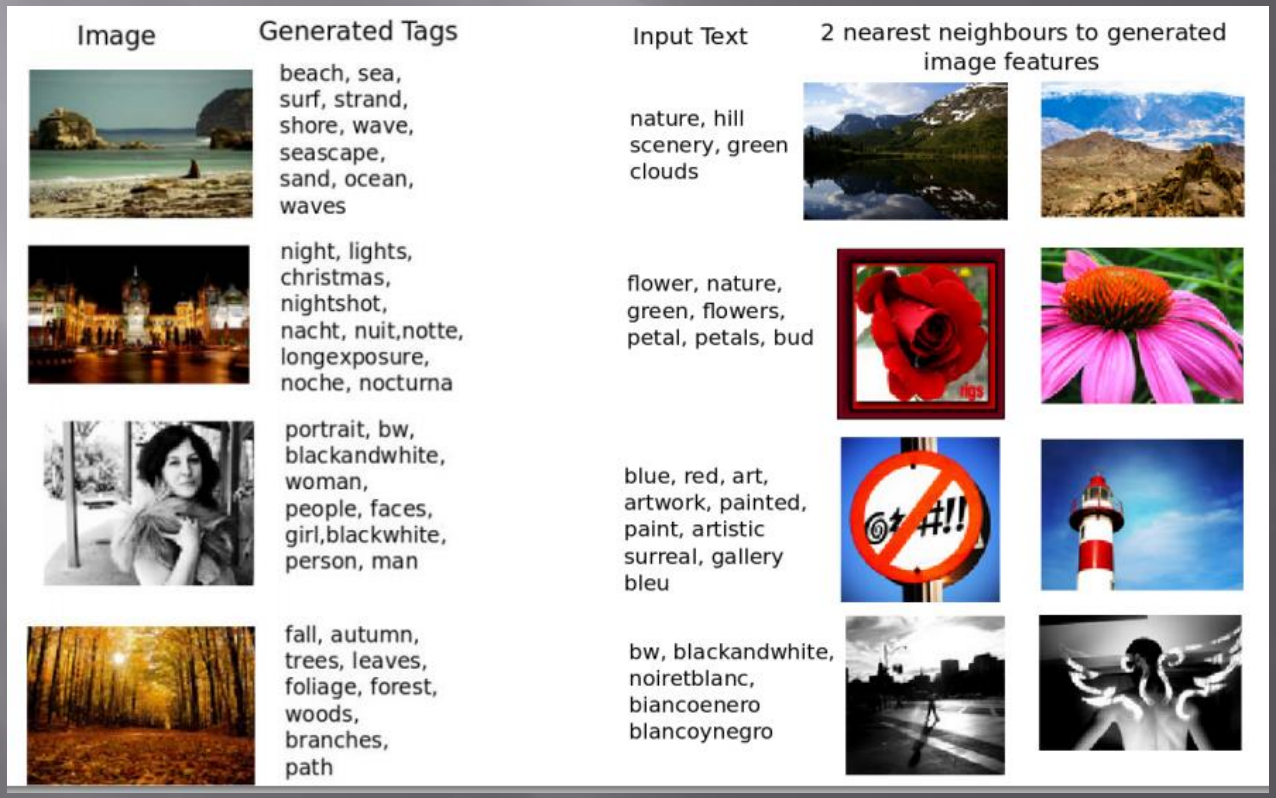
Deep Learning the Relationship between Visual and Textual Data

Patrick Poirson
University of South Alabama

Problem

- Learn the relationship between the very different but correlated visual (images, videos, audio) and textual (descriptions, tags) domains

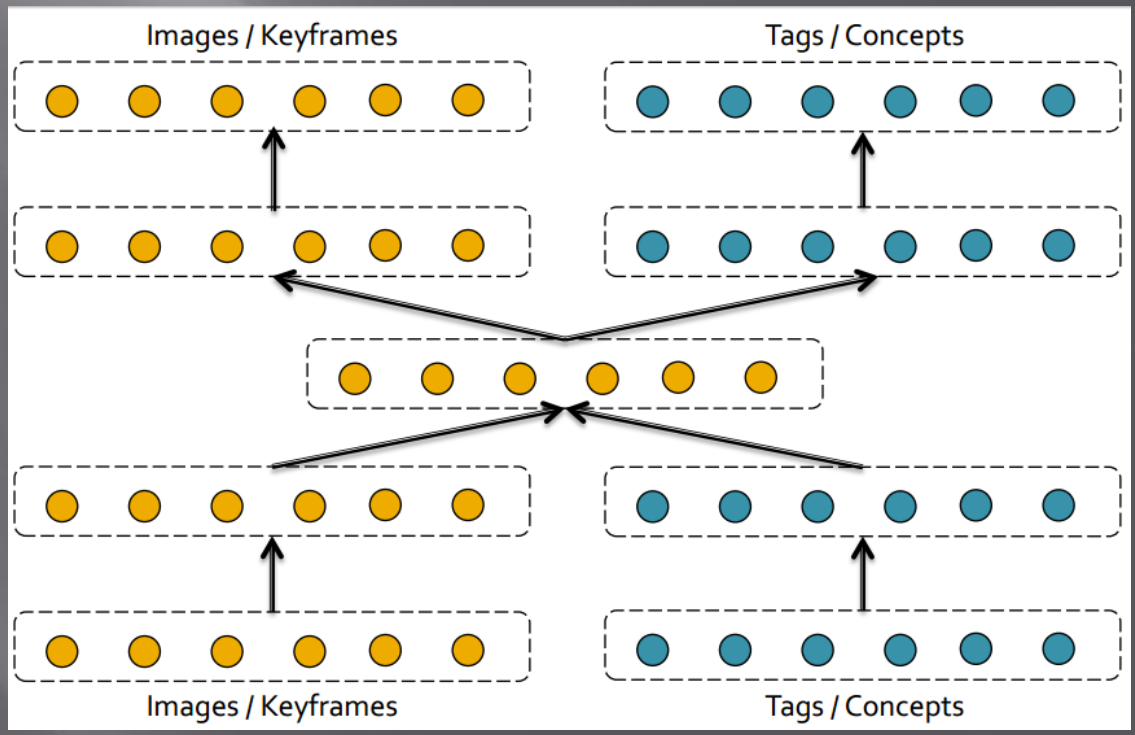


Approach

- Deep learning finds underlying non-linear representations for input data. We will use Stacked Denoising Autoencoders to learn this representation.
- Use MIR

FLICKR data set

25K and 1M



Papers

- Vincent et al, *Stacked Denoising Autoencoders: Learning Useful Representation in a Deep Network with a Local Denoising Criterion*
- Vincent et al, *Extracting and Composing Robust Features with Denoising Autoencoders*
- Chen et al, *Marginalized Denoising Autoencoders for Domain Adaptation*
- Ngiam et al, *Multimodal Deep Learning*
- Srivastava et al, *Multimodal Learning with Deep Boltzmann Machines*

Vincent, Stacked Denoising Autoencoders

- DAEs are trained to reconstruct a clean "repaired" input from a corrupted version. This is done by first corrupting the initial input.
- Forces the hidden layer to discover more robust features and prevent it from simply learning the identity
- Variation on stacking of ordinary autoencoders
- Significantly lower classification error
- Able to learn Gabor-like edge detectors
- SAE vs. DAE
  - SAE constrain the representation
  - DAE change the reconstruction criterion to denoising

Chen, Marginalized DAE

- Two limitations of SDAs: high computational cost and lack of scalability to high-dimensional features
- mSDAs marginalize noise and thus do not require optimization algorithms to learn parameters. Computed in closed-form
- Uses linear decoders as the basic building block
- Significantly speeds up SDAs by two orders of magnitude

Ngiam, Multimodal Deep Learning

- Uses autoencoders for deep multimodal learning
- Uses only video and audio
- Sparse criterion was used for training the autoencoder. Instead of denoising criterion

Srivastava, Multimodal Learning DBM

- Tackles the same problem but using DBM instead of DAE
- Compared their model to autoencoders and deep belief networks, and show that it achieves noticeable gains

Srivastava, Features for the MIR FLICKR Dataset

- Text features are word counts of 2000-most frequent tags
- Features extracted for the images
  - (1) Color Layout Descriptor (192 dimensions)
  - (2) Color Structure Descriptor (256 dimensions)
  - (3) Scalable Color Descriptor (256 dimensions)
  - (4) Homogenous Texture Descriptor (43 dimensions)
  - (5) Edge Histogram Descriptor (150 dimensions)
  - (6) Gist (960 dimensions)
  - (7) Pyramid Histogram of Words (using SIFT) (2000 dimensions)

Next steps

- Start training
  - Using matlab toolbox for SDAs
  - Use features that have already been extracted by Srivastava
- Continue researching and reading papers