Deep Learning the Relationship Between Visual and Textual Data

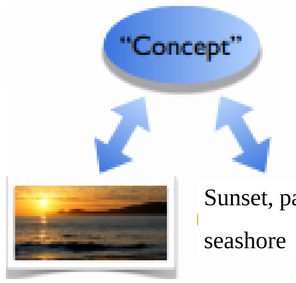
Patrick Poirson, Haroon Idrees
University of South Alabama

University of Central Florida

UCF Center for Research in Computer Vision

1. Problem

- **Introduction:**
  - Data is often a collection of modalities
  - Images and captions, audio and video, etc.
  - Need to fuse the data into a shared representation that captures the "concept"
- **Applications:**
  - Image Retrieval
  - Image Classification



2. MIR Flickr Dataset

- **Images and tags collected from Flickr**
- **1 Million Images and their Tags**
- **25K have been annotated for 38 classes**
- **Each image can belong to several classes**
- **Example classes people, female, tree, and water**

Tag	# Images	Tag	# Images
sky	845	people	330
water	641	city/urban	308/247
portrait	623	sea	301
night	621	sun	290
nature	596	girl	262
sunset	585	snow	256
clouds	558	food	225
flower/flowers	510/351	bird	218
beach	407	sign	214
landscape	385	car	212
street	383	lake	199
dog	372	building	188
architecture	354	river	175
graffiti/streetart	335/184	baby	167
tree/trees	331/245	animal	164

Average number of tags is 5.15

Pre-Processing

- **Features extracted from images:**
  - **I. Color Layout Descriptor**
  - **II. Color Structure Descriptor**
  - **III. Scalable Color Descriptor**
  - **IV. Homogeneous Texture Descriptor**
  - **V. Edge Histogram Descriptor**
  - **VI. Gist**
  - **VII. SIFT PHOW**
- **Combine Feature Vectors**
- **Text features are word counts of 2000-most frequent tags**

3. Theory

Autoencoders: The aim of an autoencoder is to learn a (encoding) for a set of data

Input: $X \in [0,1]^d$ **Output:** $Z \in [0,1]^d$

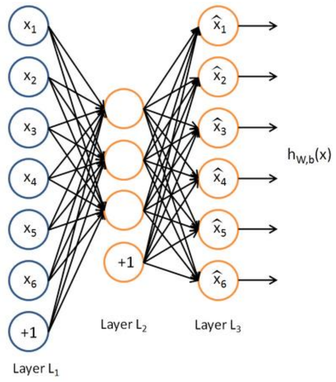
Maps X to Y through a mapping $y = s(Wx + b)$

Decoder maps Y into a reconstruction Z through

$z = s(W'y + b')$

Parameters W and W' are optimized such that the average reconstruction error is minimized

$$L_H(\mathbf{x}, \mathbf{z}) = -\sum_{k=1}^{d}[\mathbf{x}_k \log \mathbf{z}_k + (1-\mathbf{x}_k)\log(1-\mathbf{z}_k)]$$



Denoising Autoencoders: Trains the autoencoder to reconstruct the input from a corrupted version of the input

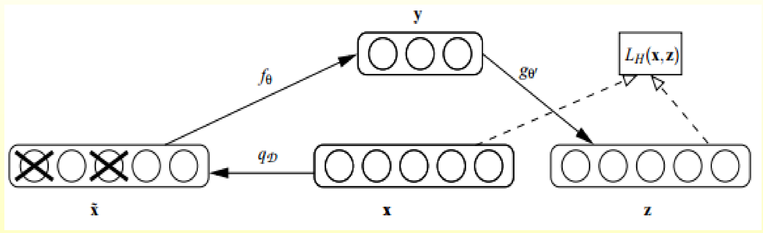
Input: $X \in [0,1]^d$ **Corrupted input:** X'

X' is mapped to Y through $y = f_\theta(x') = s(Wx' + b)$

We reconstruct $z = g_{\theta'}(y)$

Paramaters θ and θ' are trained to minimize the reconstruction error,

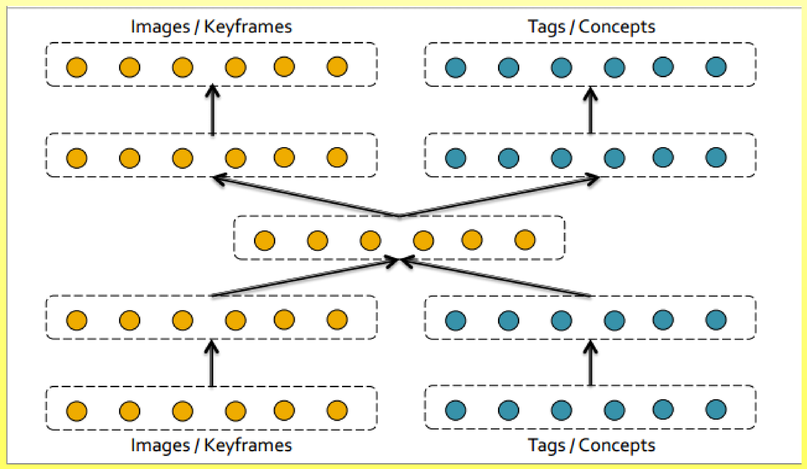
$$\boldsymbol{L(x,z) = \|x - z\|^2}$$



Multimodal Stacked Denoising Autoencoders

Pretrained using 975K unlabeled Flickr images and tags

Shared representation used for classification and retrieval



4. Results

Image Retrieval

Given: Tags Output: Similar images

Retrieved Images

Input Tags

NYC, New York, Taxi,

Times Square, Night







- **Multimodal Classification**
- **Given: Images and tags**
- **Training: MIR Flickr 25K**
  - **Randomly select 15K images and tags**
- **Testing: MIR Flickr 25K**
  - **Randomly select 10K images and tags**

Method	MAP	Prec@50
Random	.124	.124
LDA	.492	.754
SVM	.475	.758
DBM	.609	.873
SDA (Proposed)	.533	.858

