Effective Multi-Modal Retrieval based on Stacked Auto-Encoders

Author: Wei Wang, etc. NUS

Speaker: Luo Zhiyi

jessieluo1991@gmail.com

Oct 15th, 2014

Problem

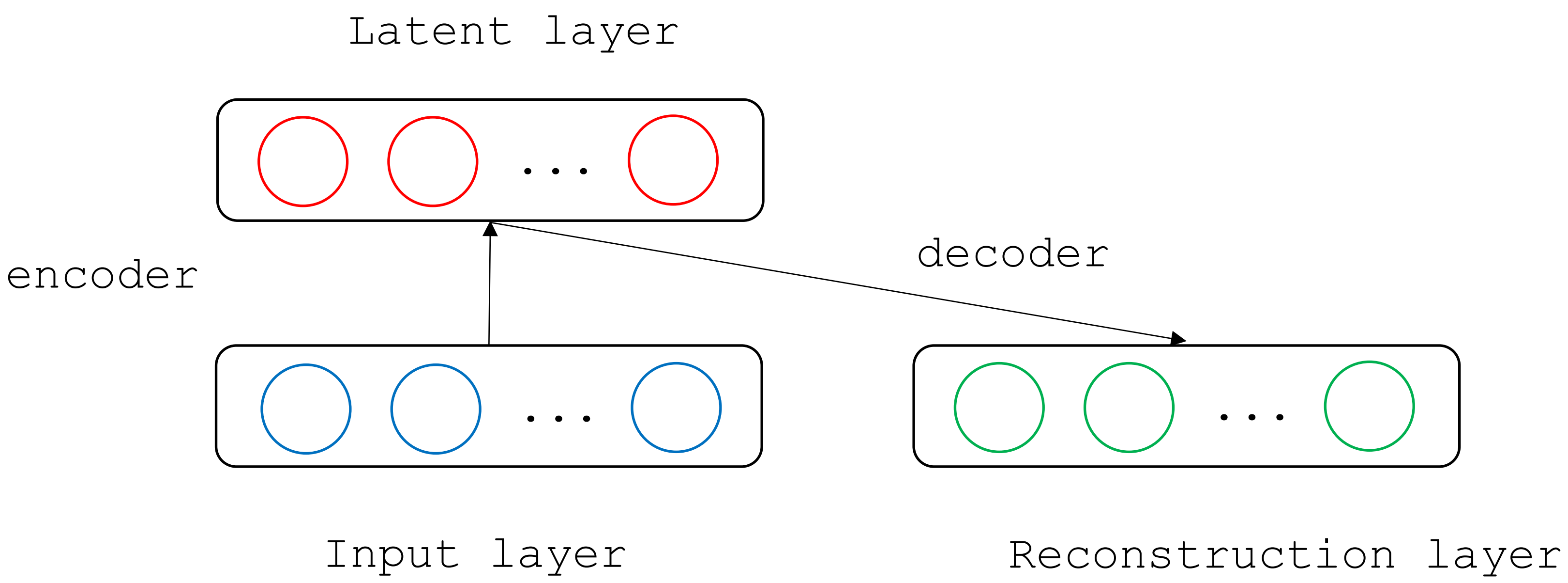
- Large-scale information retrieval from multiple modalities (**text**, **image**, **video**)

- Give me Hobbit trailers

Outline

AE	Auto-encoder
SAE	Stacked Auto-encoder
MSAE	Multi-modal Stacked Auto-Encoders
Training Algorithm	Single SAE Training Stage Multi-Modal Training Stage
Experiment	Demo

Auto-encoder



Auto-encoder

Learning parameters: W, b

Loss function: Reconstruction error and L2 regularization

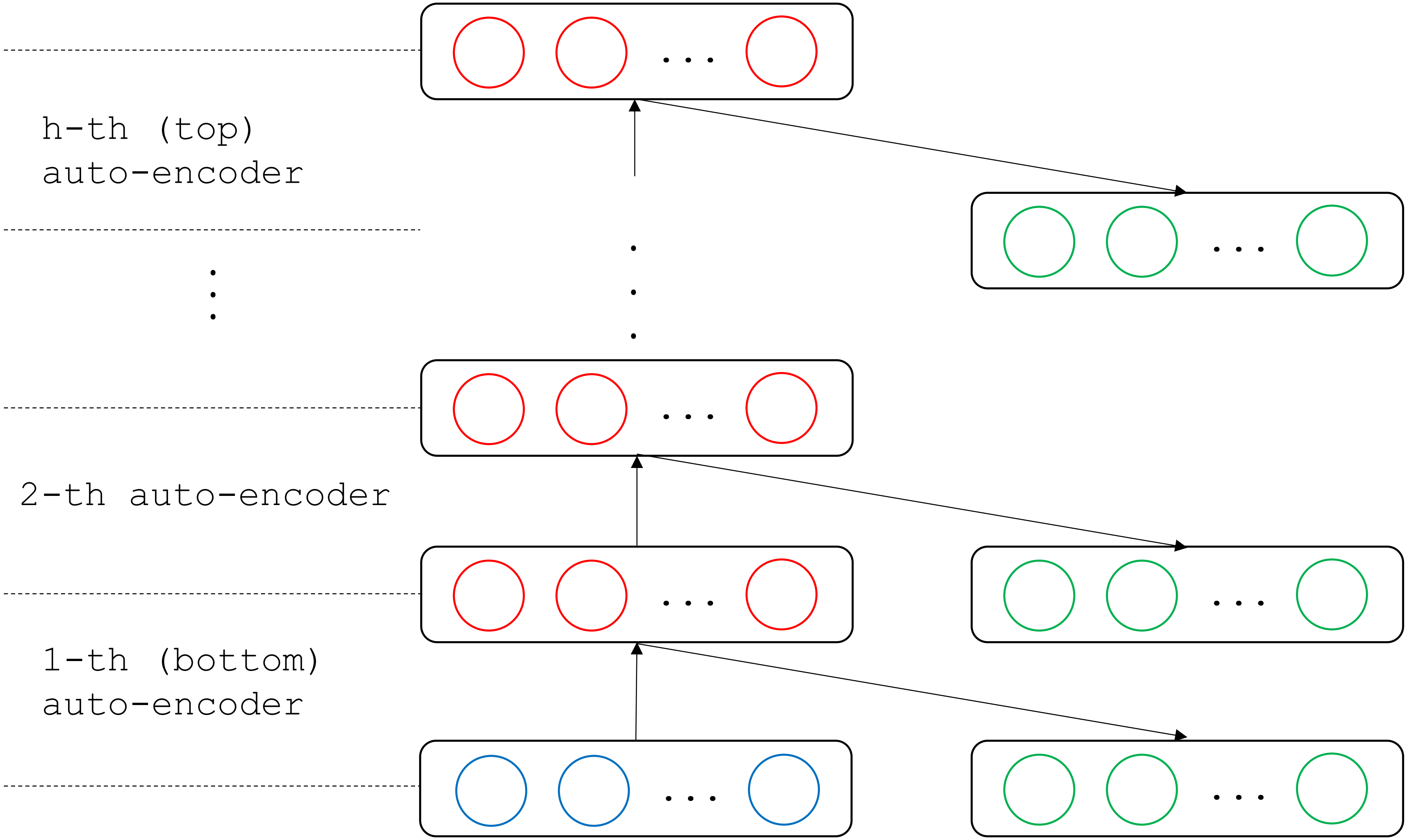
$$\mathcal{L}(x_0, x_2) = \mathcal{L}_r(x_0, x_2) + 0.5\xi(||W_1||_2^2 + ||W_2||_2^2)$$

Stacked Auto-Encoder

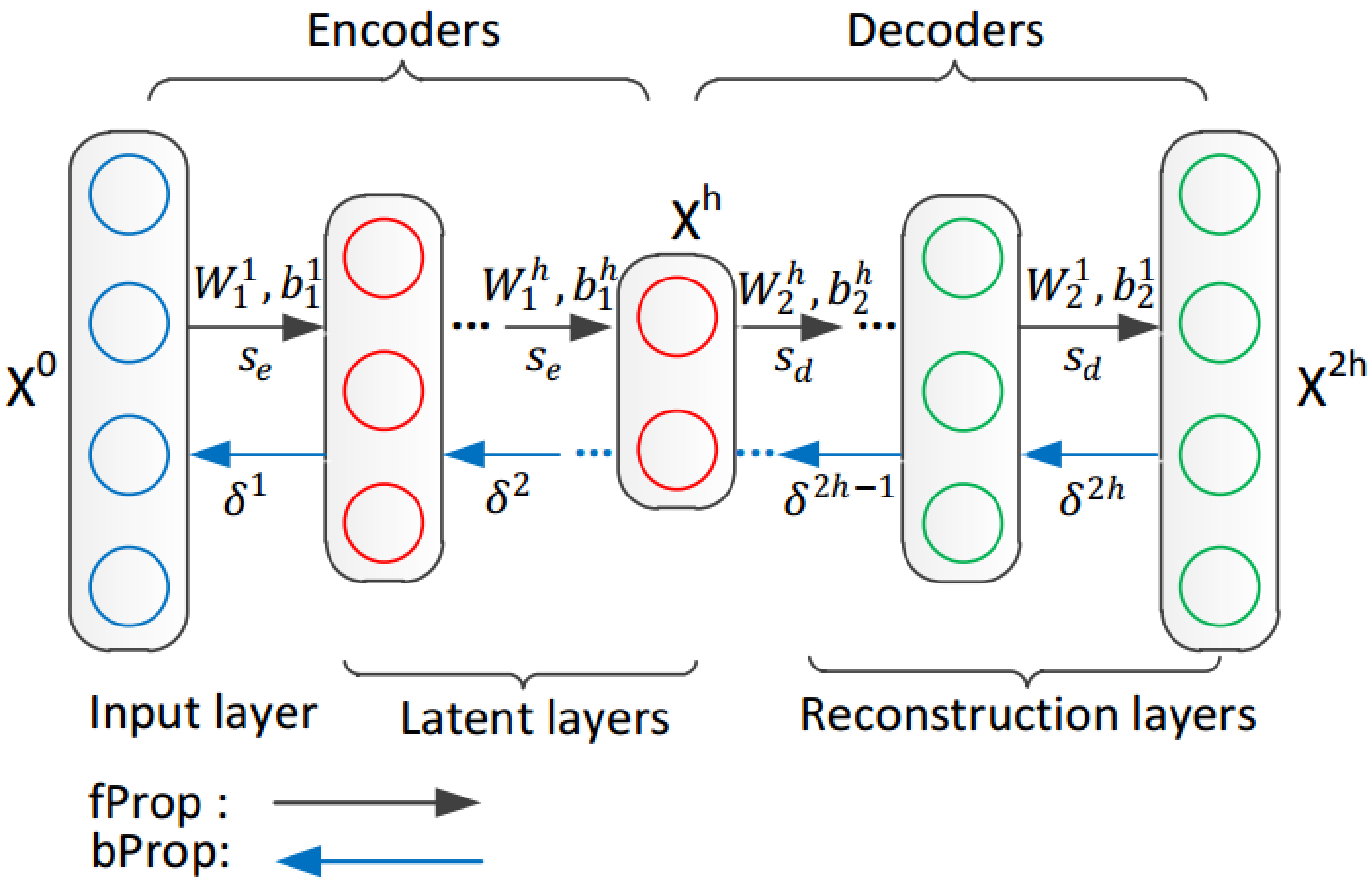
SAE — Stacked version of auto-encoder

Training — Pre-training

Fine-tuning



Stacked Auto-encoders



Unfolded Stacked Auto-Encoders

MSAE

Definitions

Latent Space Mapping

Given an image feature vector $x \in \mathbb{D}_I$ and a text feature vector $y \in \mathbb{D}_T$, find two mapping functions $f_I : \mathbb{D}_I \rightarrow \mathbb{Z}$ and $f_T : \mathbb{D}_T \rightarrow \mathbb{Z}$ such that if x and y are semantically relevant, the distance between $f_I(x)$ and $f_T(y)$, denoted by $dist(f_I(x), f_T(y))$, is small in the common latent space $\mathbb{Z}$.

Multi-Modal Search

Given a query object $Q \in \mathbb{D}_q$ ($q \in \{I, T\}$) and a target domain $\mathbb{D}_t \subset \mathbb{D}$ ($t \in \{I, T\}$), find a set $O \subset \mathbb{D}_t$ with k objects such that $\forall o \in O$ and $o' \in \mathbb{D}_t/O$, $dist(f_q(Q), f_t(o')) \geq dist(f_q(Q), f_t(o))$.

MSAE Training

Intuition

Intra-modal semantics can be preserved or even enhanced through inter-modal relationships with other modalities whose features are of high quality.

Training Algorithm

Single SAE Training

Multi-Modal Training

Experiment

Single SAE Training

- One SAE is trained for each modality
- Capture the intra-modal semantics

Algorithm

Algorithm 1 trainSAE(h, X^0, d)

Input: h, height of SAE
Input: X^0, training data, one example per row
Input: d, a sequence of dimensions for each layer
Output: $\theta = \{\theta^i\}_{i=1}^h$, parameters of SAE

1. **for** $i = 1$ to h **do**
2. random init $\theta^i \leftarrow d_{i-1}, d_i$
3. (θ^i, X^i)=**trainNN**$(1, X^{i-1}, \theta^i)$
4. $\theta \leftarrow$**trainNN**(h, X^0, θ)

trainNN(h, X, θ)

1. **repeat**
2. **for** batch B^0 in X **do**
3. Z, B=**fProp**$(2h, B^0, \theta)$
4. $\delta^{2h} = \frac{\partial \mathcal{L}(B^0)}{\partial Z^{2h}}$
5. **bProp**$(2h, \delta^{2h}, B, Z, \theta)$ //(see Appendix)
6. **until** converge
7. return **fProp**(h, X, θ)

Multi-Modal Training

- Intuition: Enhance the latent features even when original feature is bad
- Objective function: $L(X^0, Y^0) = \alpha L_r^I(X^0, X^{2h}) + \beta L_r^T(Y^0, Y^{2h}) + L_d(X^h, Y^h) + \xi(\theta)$
- Step: Iterate over all SAEs
  Adjusting the parameters in one SAE at a time(fixed others)
- Goal: Capture both intra-modal semantics and inter-modal semantics.

Algorithm

Algorithm 2 trainMSAE(h, X^0, Y^0, θ)

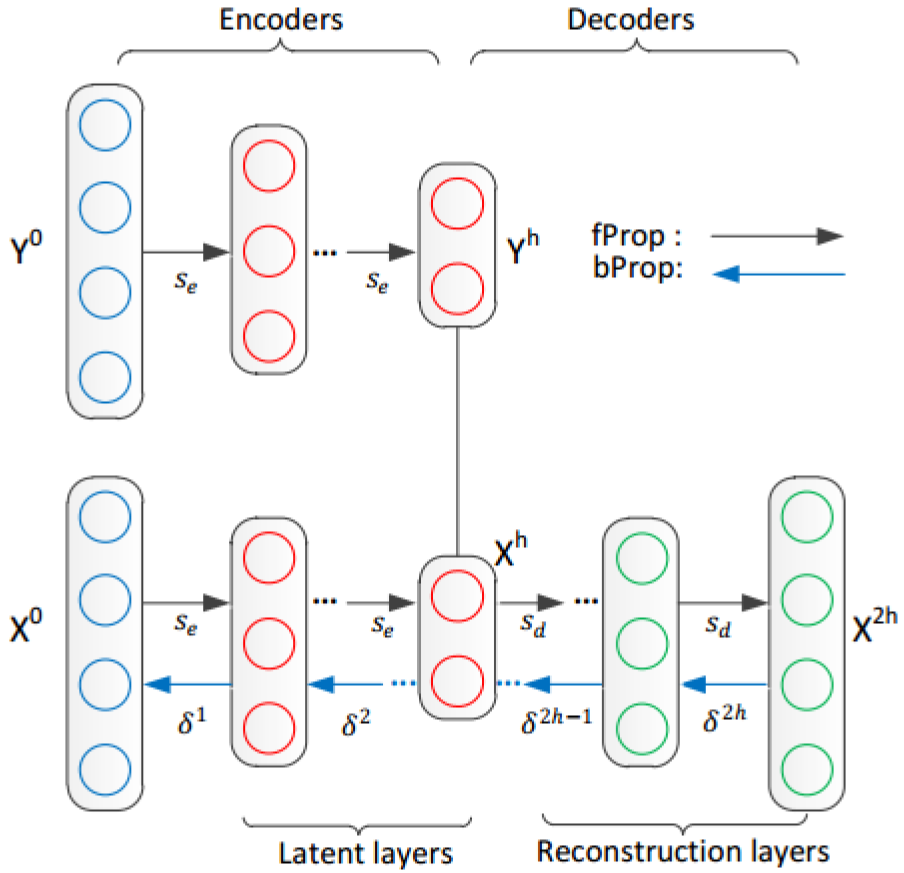
Input: h, height of MSAE
Input: X^0, Y^0, image and text input data
Input: θ=(θ_X, θ_Y), parameters of MSAE, initialized by **trainSAE**
Output: θ, updated parameters

repeat
  trainMNN$(h, X^0, Y^0, \theta_X, \theta_Y)$//train image SAE
  trainMNN$(h, Y^0, X^0, \theta_Y, \theta_X)$//train text SAE
until converge

trainMNN$(h, X, Y, \theta_X, \theta_Y)$

Input: X, input data for the modality whose SAE is to be updated
Input: Y, input data for the modality whose SAE is fixed
Input: θ_X, θ_Y, parameters for the two SAEs.

repeat
  for batch (B_X^0, B_Y^0) in (X, Y) **do**
    B_X, Z_X=**fProp**$(2h, B_X^0, \theta_X)$
    B_Y, Z_Y=**fProp**(h, B_Y^0, θ_Y)
    $\delta^{2h} = \frac{\partial \mathcal{L}(B_X^0, B_Y^0)}{\partial Z_X^{2h}}$
    δ^h=**bProp**$(h, \delta^{2h}, \{B_X^i\}_{i=h}^{2h}, \{Z_X^i\}_{i=h}^{2h}, \{\theta_X^i\}_{i=h}^{2h})$
    $\delta^h += \frac{\partial \mathcal{L}_d(B_X^h, B_Y^h)}{\partial Z_X^h}$
    bProp$(h, \delta^h, \{B_X^i\}_{i=0}^{h}, \{Z_X^i\}_{i=1}^{h}, \{\theta_X^i\}_{i=1}^{h})$
until converge



Experiment

Datasets

NUSWIDE
Wiki
Flickr1M

Dataset	**NUS-WIDE**	**Wiki**	**Flickr1M**
Total size	190,421	2,866	1,000,000
Training set	60,000	2,000	975,000
Validation set	10,000	366	6,000
Test set	120,421	500	6,000
Average Text Length	6	131	5

Evaluation Metric

Mean Average Precision (MAP)

Results

Task		$\mathbb{Q}_{I\to I}$				$\mathbb{Q}_{T\to T}$				$\mathbb{Q}_{I\to T}$				$\mathbb{Q}_{T\to I}$			
Algorithm		LCMH	CMSSH	CVH	MSAE	LCMH	CMSSH	CVH	MSAE	LCMH	CMSSH	CVH	MSAE	LCMH	CMSSH	CVH	MSAE
Dimension of	16	0.353	0.355	0.365	**0.417**	0.373	0.400	0.374	**0.498**	0.328	0.391	0.359	**0.447**	0.331	0.337	0.368	**0.432**
Latent Space	24	0.343	0.356	0.358	**0.412**	0.373	0.402	0.364	**0.480**	0.333	0.388	0.351	**0.444**	0.323	0.336	0.360	**0.427**
L	32	0.343	0.357	0.354	**0.413**	0.374	0.403	0.357	**0.470**	0.333	0.382	0.345	**0.402**	0.324	0.335	0.355	**0.435**

NUSWIDE

Task		$\mathbb{Q}_{I\to I}$				$\mathbb{Q}_{T\to T}$				$\mathbb{Q}_{I\to T}$				$\mathbb{Q}_{T\to I}$			
Algorithm		LCMH	CMSSH	CVH	MSAE	LCMH	CMSSH	CVH	MSAE	LCMH	CMSSH	CVH	MSAE	LCMH	CMSSH	CVH	MSAE
Dimension of	16	0.146	0.148	0.147	**0.162**	0.359	0.318	0.153	**0.462**	0.133	0.138	0.126	**0.182**	0.117	0.140	0.122	**0.179**
Latent Space	24	0.149	0.151	0.150	**0.161**	0.345	0.320	0.151	**0.437**	0.129	0.135	0.123	**0.176**	0.124	0.138	0.123	**0.168**
L	32	0.147	0.149	0.148	**0.162**	0.333	0.312	0.152	**0.453**	0.137	0.133	0.128	**0.187**	0.119	0.137	0.123	**0.179**

Wiki

Task		$\mathbb{Q}_{I\to I}$		$\mathbb{Q}_{T\to T}$		$\mathbb{Q}_{I\to T}$		$\mathbb{Q}_{T\to I}$	
Algorithm		CVH	MSAE	CVH	MSAE	CVH	MSAE	CVH	MSAE
Dimension of	16	**0.622**	0.621	0.610	**0.624**	0.610	**0.632**	**0.616**	0.608
Latent Space	24	0.616	**0.619**	0.604	**0.629**	0.605	**0.628**	**0.612**	**0.612**
L	32	0.603	**0.622**	0.587	**0.630**	0.588	**0.632**	0.598	**0.614**

Flickr1M

Conclusion

Propose MSAE mechanism
Multi-modal Stacked Auto-Encoders

- An effective mapping mechanism for multi-modal retrieval

Design Learning Objective Function

- Capture intra-modal semantics
  Capture inter-modal semantics

Demo

- Visualization of Training Process
- http://www.comp.nus.edu.sg/~wangwei/code/msae/index.html

Thank you!