Collective Visual Inference

Gang Hua

Microsoft Research Asia
ganghua@microsoft.com

Standard recognition regime

IMAGENET

Microsoft
Common Objects in Context

Dog, Cat

A dog is in love with a cat.

Sheep, Horse

A goat rides a horse.

Deep Learning

- Each visual instance is handled independently

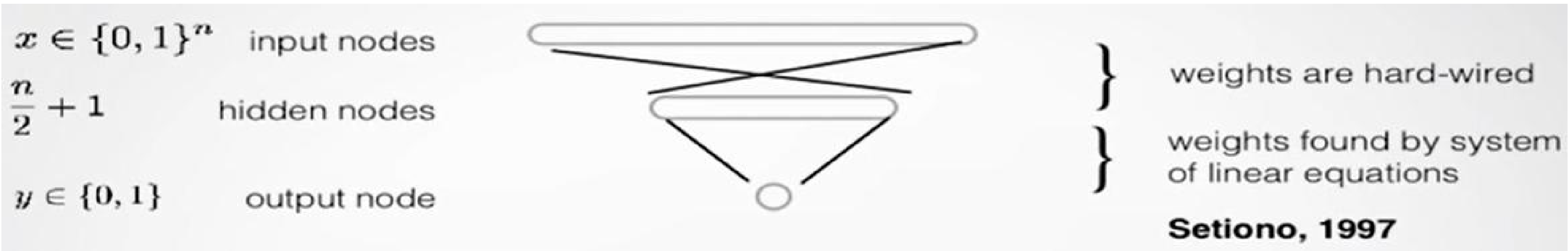
Caution about "end-to-end" learning…

- There are problems that DNNs solve very well
  - Object category recognition
  - Object detection
  - Face recognition
  - Speech to text

- But, DNNs fail on some seemingly simple problems
  - N-bit parity problems
  - Multiply numbers
  - Simple visual tasks

The importance of prior knowledge

- Example: the n-bit parity problem



- Even though there exists weights that solve the n-bit parity problems, “learning” them using the available training techniques does not work for n>30.
- This failure to train a DNN holds true also for overly-subscribed architecture.

The importance of prior knowledge

- Example: learning arithmetic operations

[Hoshen & Peleg, 2015]

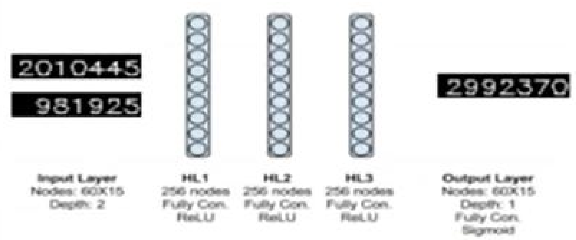


Figure 2. A diagram showing the construction of a neural network with 3 hidden layers able to preform addition using visual data. Two pictures are used as input and one picture as output. The network is fully connected and uses ReLU units in the hidden layers and sigmoid in the output layer. The hidden layers have 256 units each.

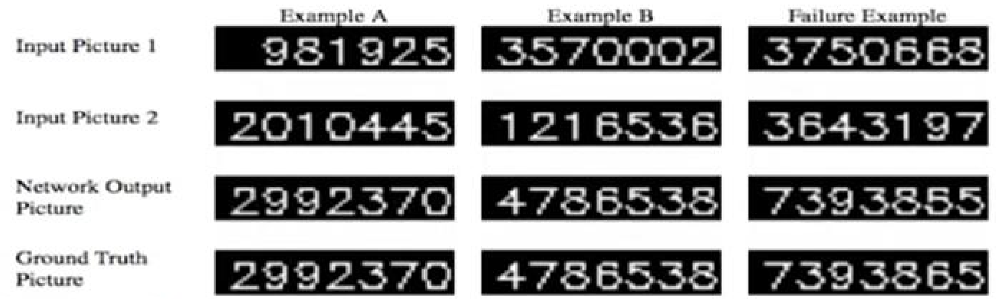


Figure 1. Input and output examples from our neural network trained for addition. The first two examples show a typical correc
The last example shows a rare failure case.

- DNN failed on the task of multiplication – whatever architecture they used they were unsuccessful in training the DNN

The importance of prior knowledge

- Example: Pentomino Dataset

[Gulcehre & Bengio, 2015]

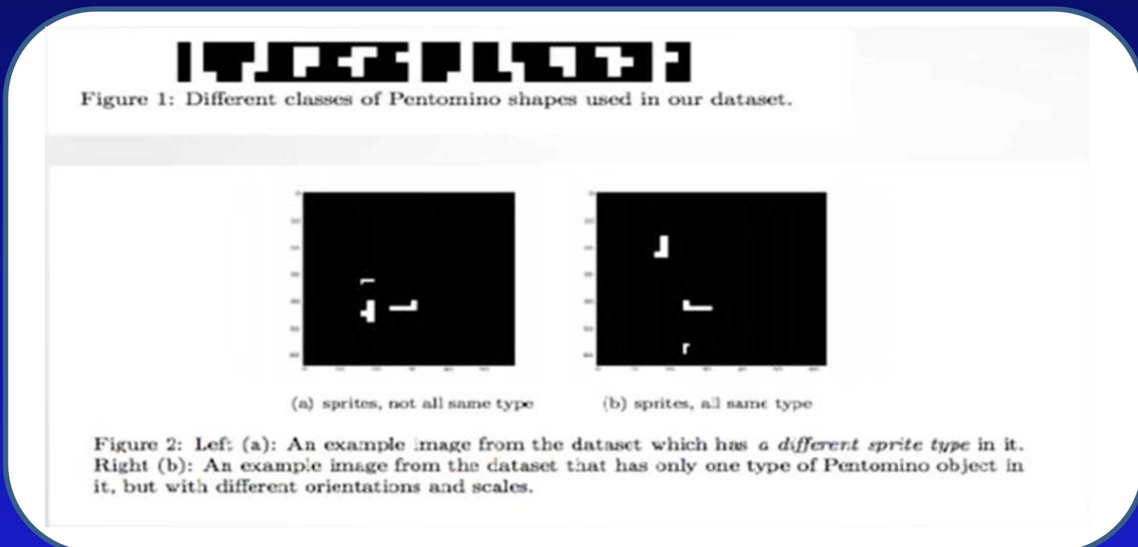


Figure 1: Different classes of Pentomino shapes used in our dataset.

(a) sprites, not all same type

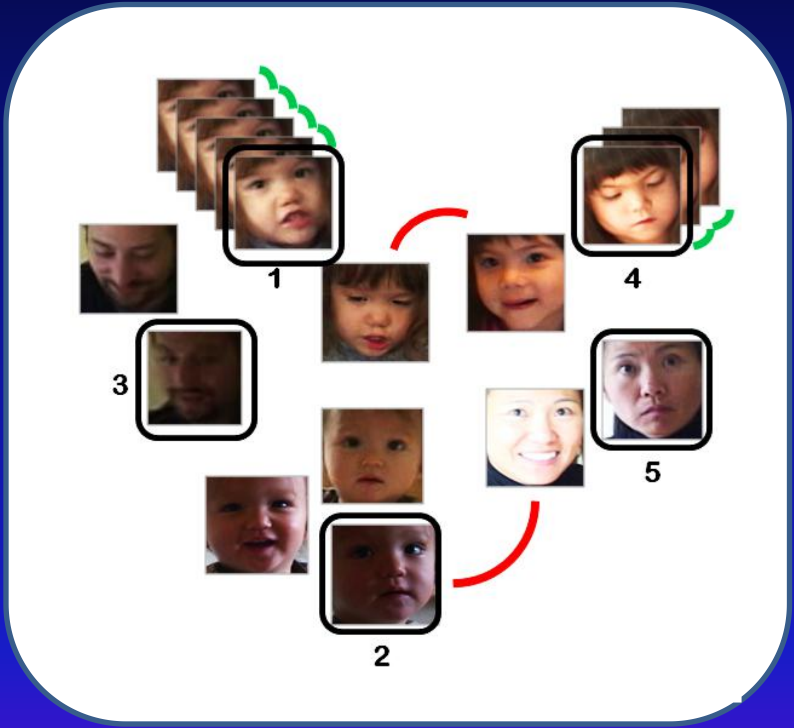
(b) sprites, all same type

Figure 2: Left (a): An example image from the dataset which has *a different sprite type* in it. Right (b): An example image from the dataset that has only one type of Pentomino object in it, but with different orientations and scales.

- Different part types, which can appear following some 2D geometric transformations
- Task: find out whether all parts are of the same class or not

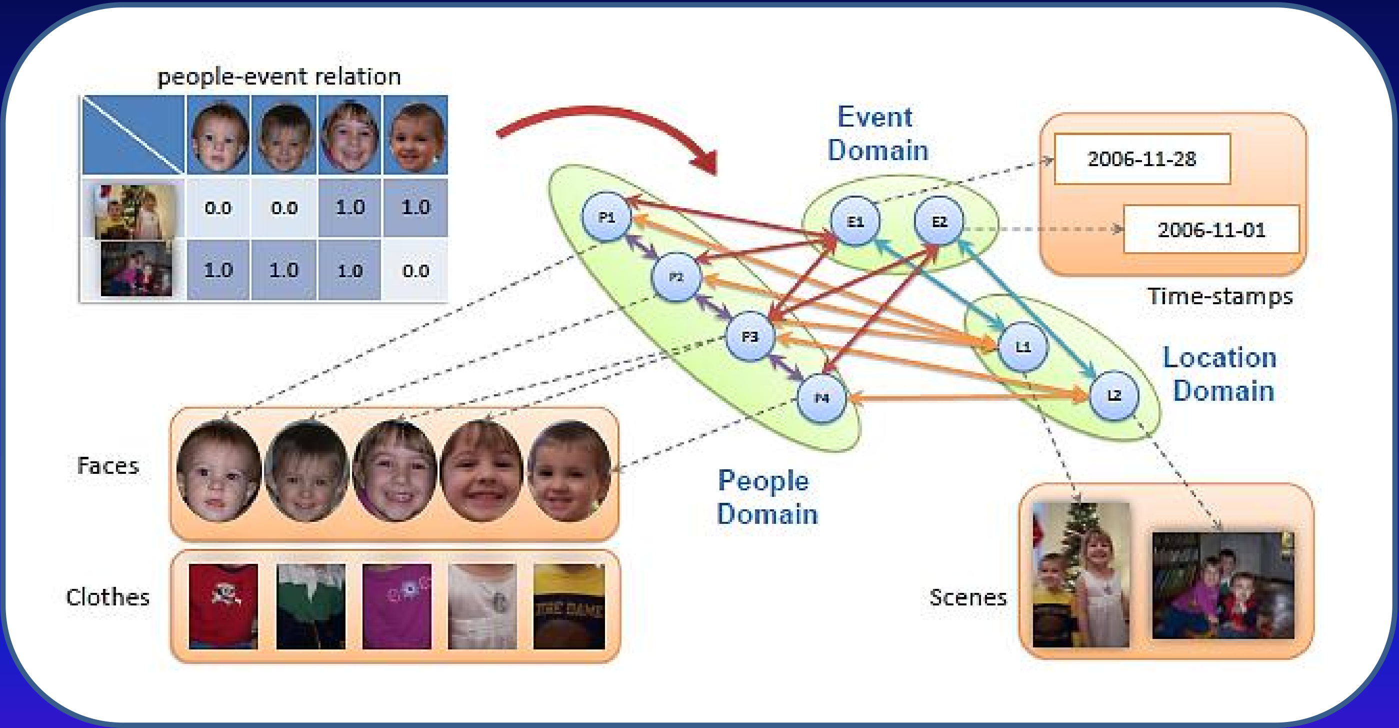
- DNN failed on the task when end-to-end was concerned
- DNN succeeded when the task was broken down into first finding the category of each part and then making a decision whether all categories are the same

Relations among images….



- Faces from the same natural image is unlikely to be the same person
- Faces from the same face track from a video have the same identity

Relations among images….

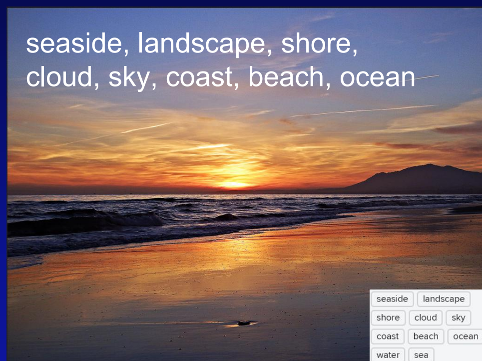


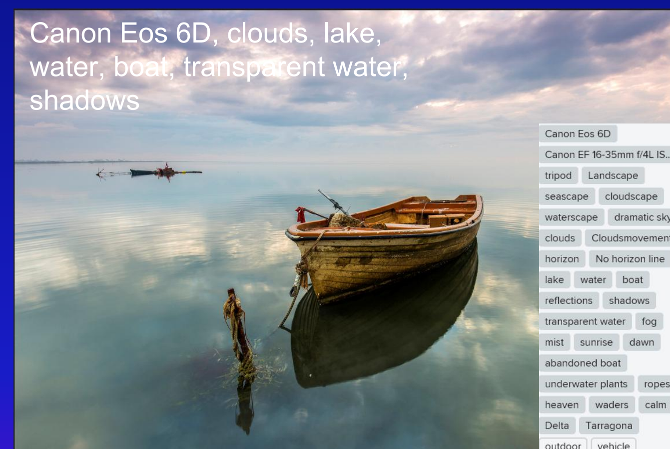
- Visual instances from different semantic domains may have strong correlations

Relations among images….

- A set of images may present the same object category

Relations among images….





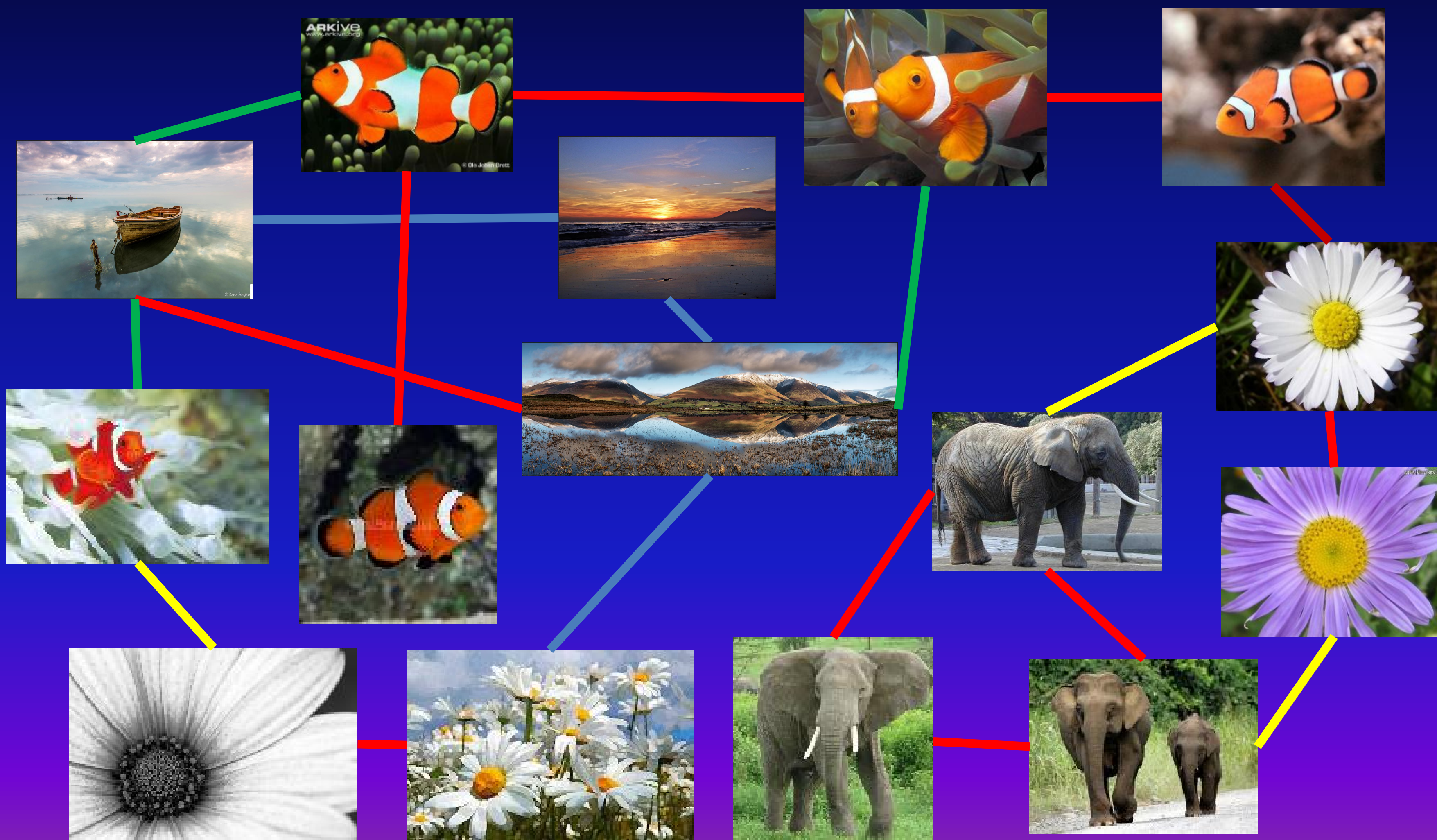
- Online photos in social media are often tagged with text keywords

Relations are common…

- Hyper-links
- Geo-tags and locations
- Spatial configurations of cameras
- Social networks
- Temporal correspondences
- … … …

Can these relations be leveraged to benefit visual understanding?

Paradigm shift: collective modeling

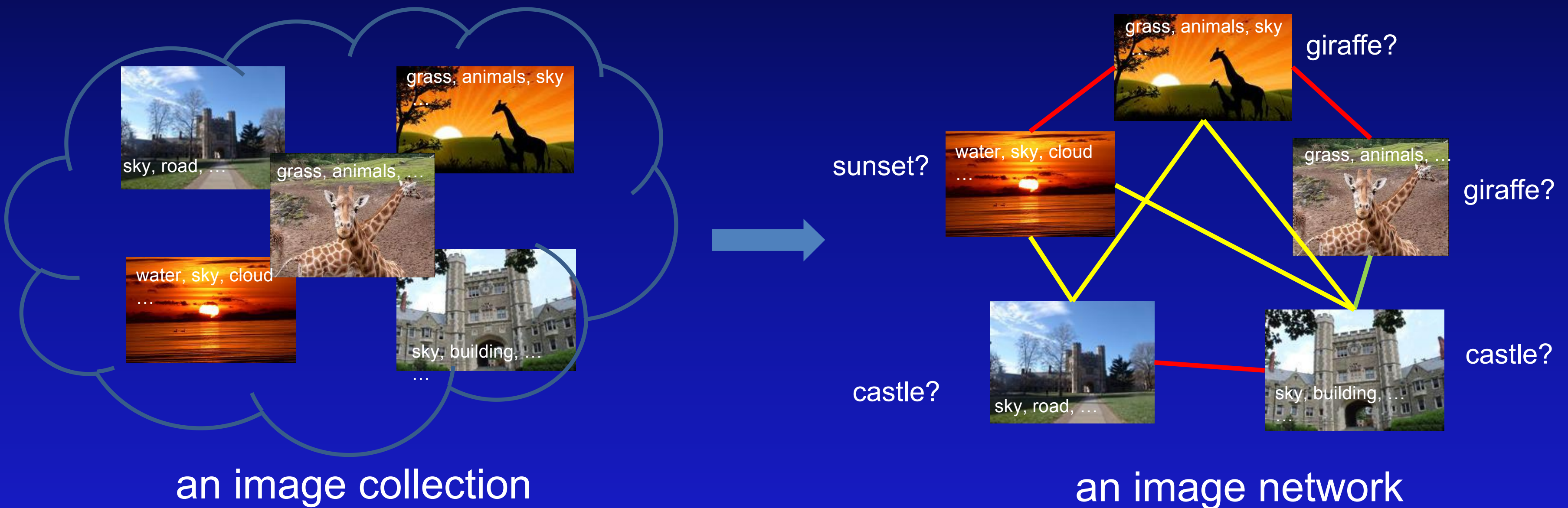


Outline

- Image understanding/recognition in social media
  - ***Visual Topic Network***: a relational statistical model of an image collection
  - ***Inference& learning***: collective visual inference across a set of images
  - ***Experiments***: validation on the NUS-WIDE and MIRFLICKR dataset
  - ***Conclusion***: discussions and future work

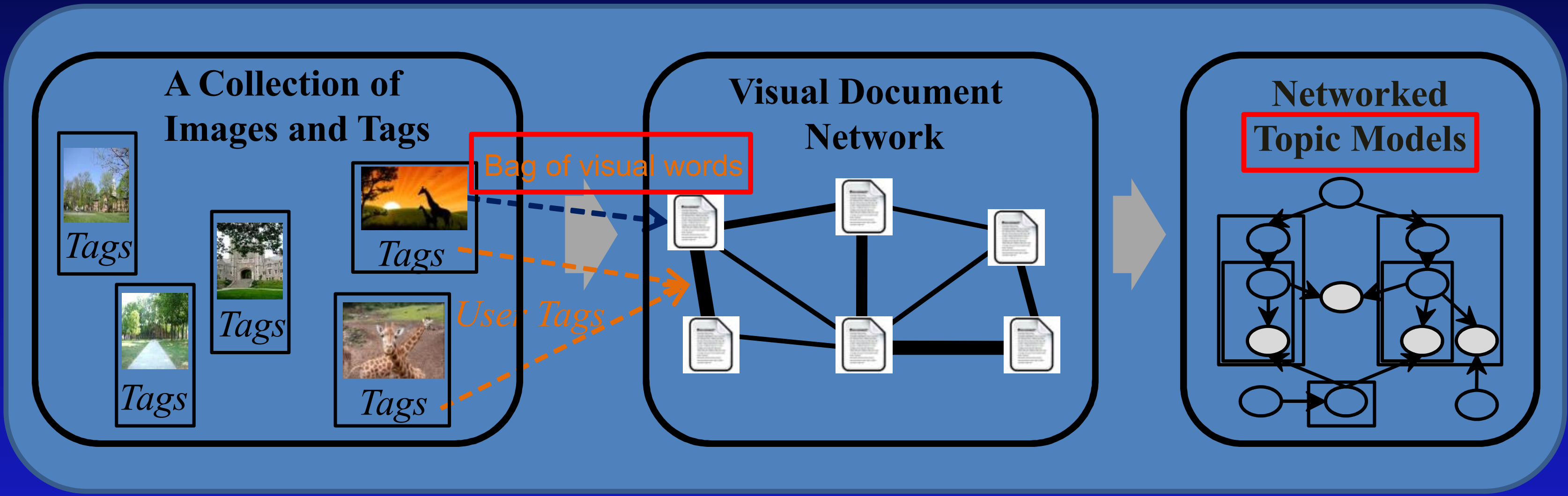
- Other works
  - Active learning with prior context for interactive face tagging
  - Joint people, event, and location recognition in personal photo album
  - Image and video object co-segmentation

Problem to solve



- Collective inference over an image ***network***

Our model: visual topic network



- Visual Topic Network: a set of networked topic models

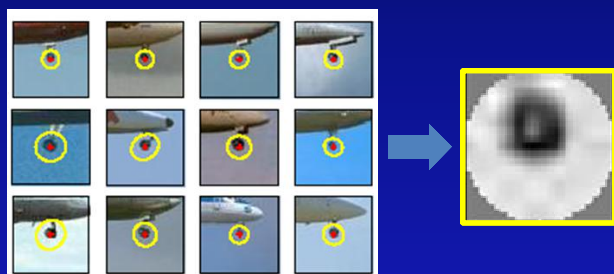
Bag-of-words

- Frequency of words from a dictionary in a document [Salton & McGill (1983)]

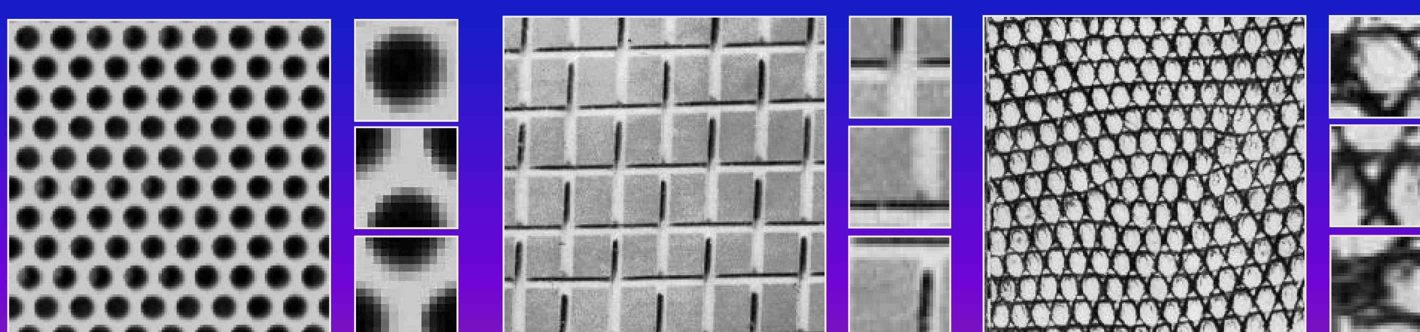
US Presidential Speeches Tag Cloud
http://chir.ag/phernalia/preztags/

Visual words

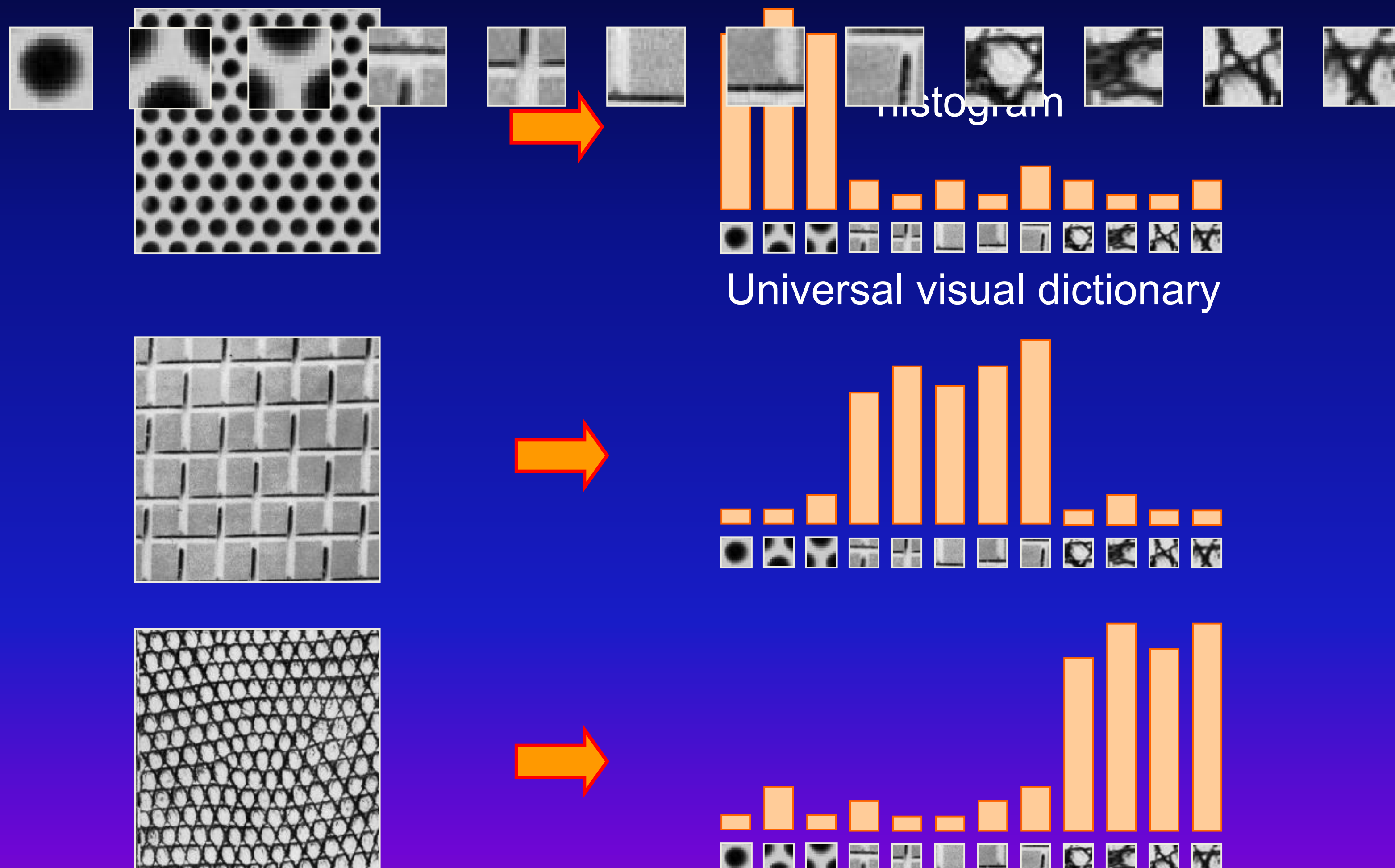
- Prototype local image patches
  - From local feature detector



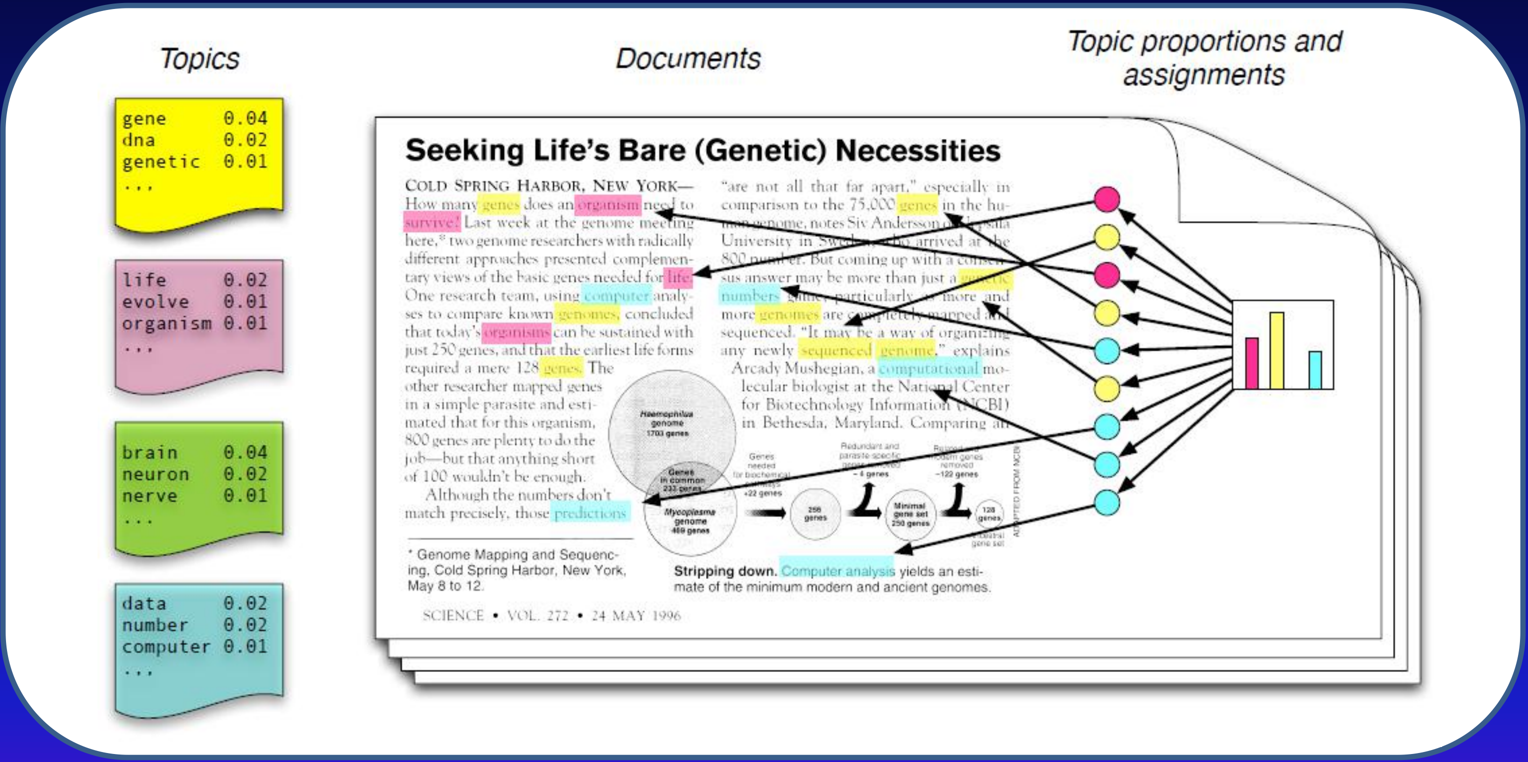
  - Repetition of basic elements of textures (texton)



Universal visual dictionary



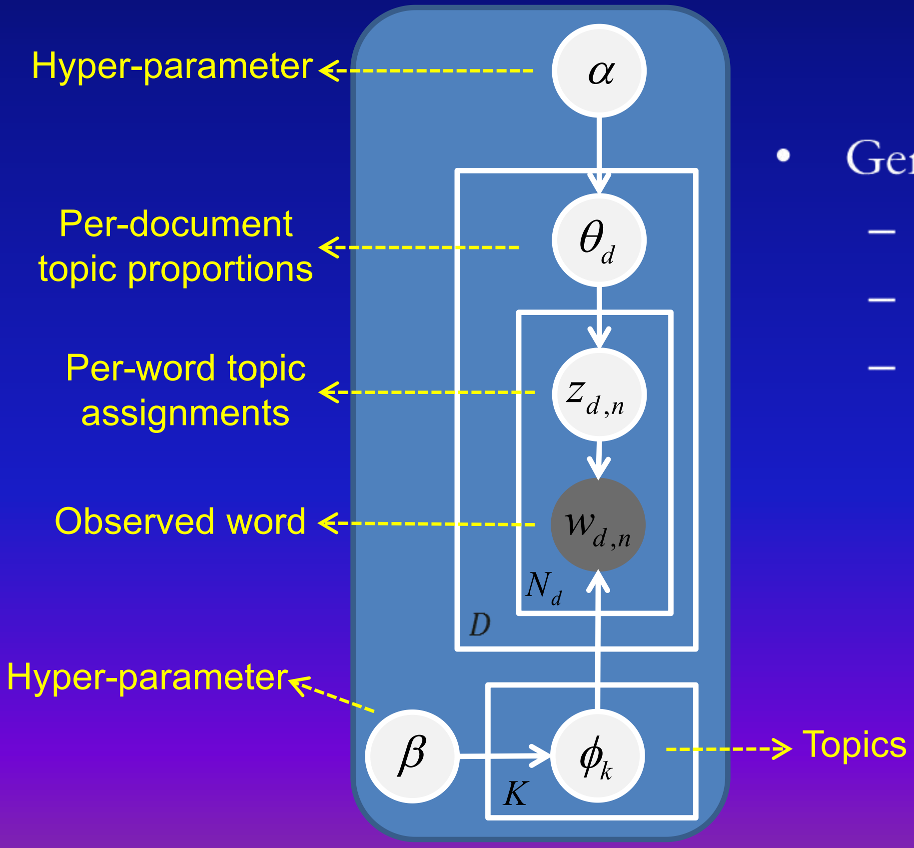
Topic models



- Each topic corresponds to a distribution over words (terms)
- Each document corresponds to a distribution over topics

Latent Dirichilet Allocation (LDA)

$$P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\phi} | \alpha, \beta) = \prod_{k=1}^{K} P(\phi_k|\beta) \prod_{d=1}^{D} P(\theta_d|\alpha) \prod_{n=1}^{N_d} P(z_{d,n}|\theta_d) P(w_{d,n}|\phi_{z_{d,n}})$$



- Generative model
  - Choose $\theta_d \sim \mathrm{Dir}(\alpha)$, where $d \in \{1, \ldots, D\}$
  - Choose $\phi_k \sim \mathrm{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$
  - For each word position d, n, where $n \in \{1, \ldots, N_d\}$
    - Choose a topic $z_{d,n} \sim \mathrm{Multinomial}(\theta_d)$
    - Choose a word $w_{d,n} \sim \mathrm{Multinomial}(\phi_{z_{d,n}})$

[Blei et al, 2003]

LDA: Inference

- Objective: Estimate $\boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\phi}$ given all observed words $\boldsymbol{W}$
  - Collapsed Gibbs sampling (fixed α, β)

$$\boxed{P(\boldsymbol{W}, \boldsymbol{Z} \mid \alpha, \beta) = \int P(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\phi} \mid \alpha, \beta) d\boldsymbol{\Theta} d\boldsymbol{\phi}}$$

  - Approximate the topic posterior $P(Z|W; \alpha, \beta)$ by Gibbs sampling from $P(\boldsymbol{W}, \boldsymbol{Z} \mid \alpha, \beta)$

$$p(z_{dn} = k \mid \boldsymbol{Z}^{-dn}, \boldsymbol{W}, \alpha, \beta) \propto \left(\alpha + m_{d,k}^{-dn}\right) \frac{\boxed{n_{k,w_{dn}}^{-dn}} + \beta}{\sum_w n_{k,w_{dn}}^{-dn} + V\beta}$$

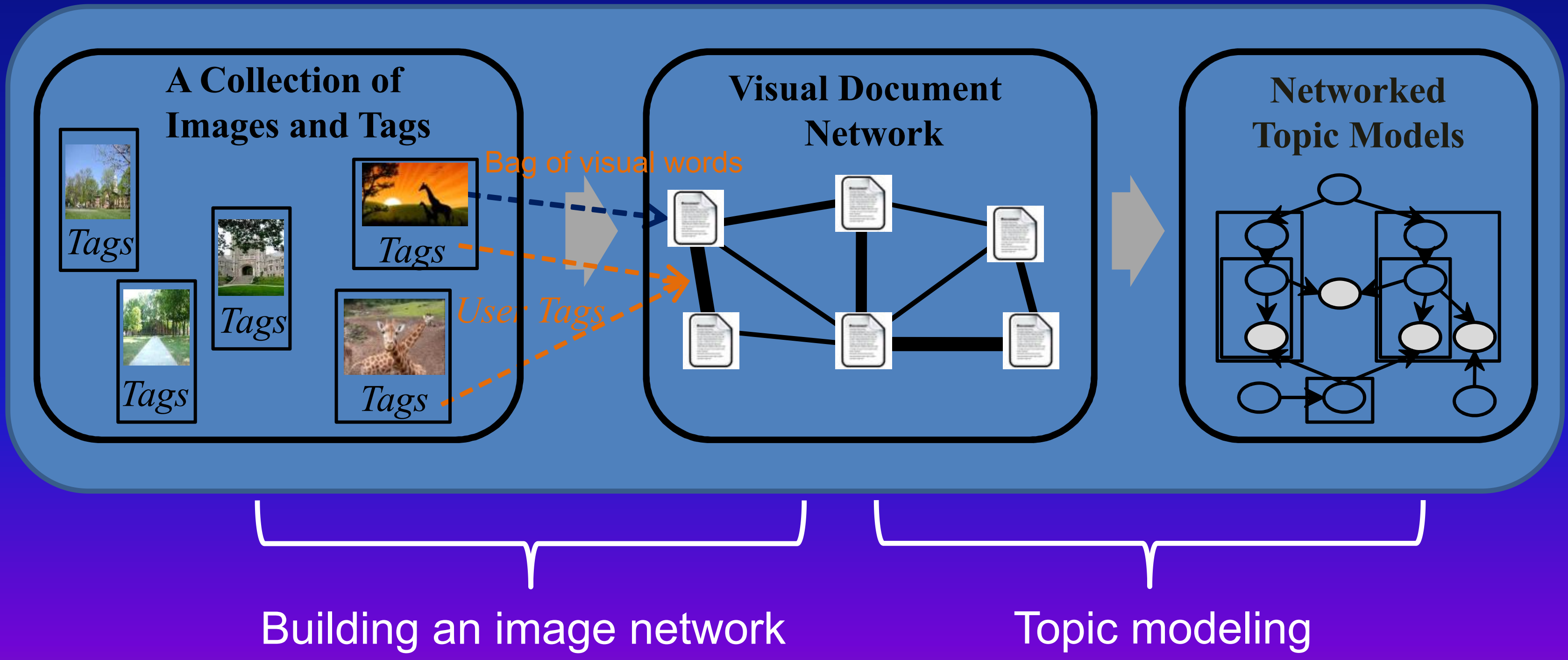
Excluding the current word w_{dn}, the number of all other words equal to w_{dn} from all documents that have been assigned to topic k

  - Statistics of topics and documents

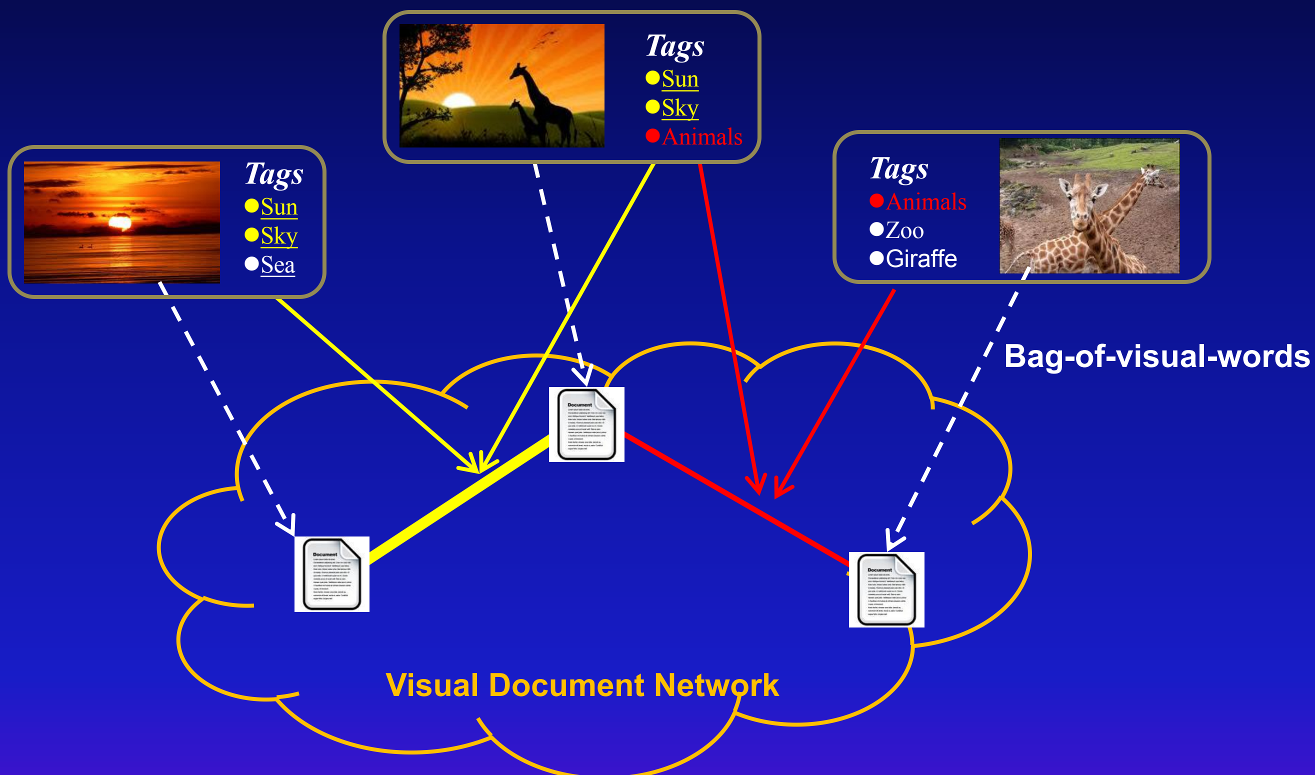
$$\phi_k(w) = \frac{n_{w,k} + \beta}{\sum_w n_{w,k} + V\beta} \qquad \theta_d(k) = \frac{n_{d,k} + \alpha}{\sum_k n_{d,k} + K\alpha}$$

Our model: visual topic network

- Visual Topic Network: a set of networked topic models



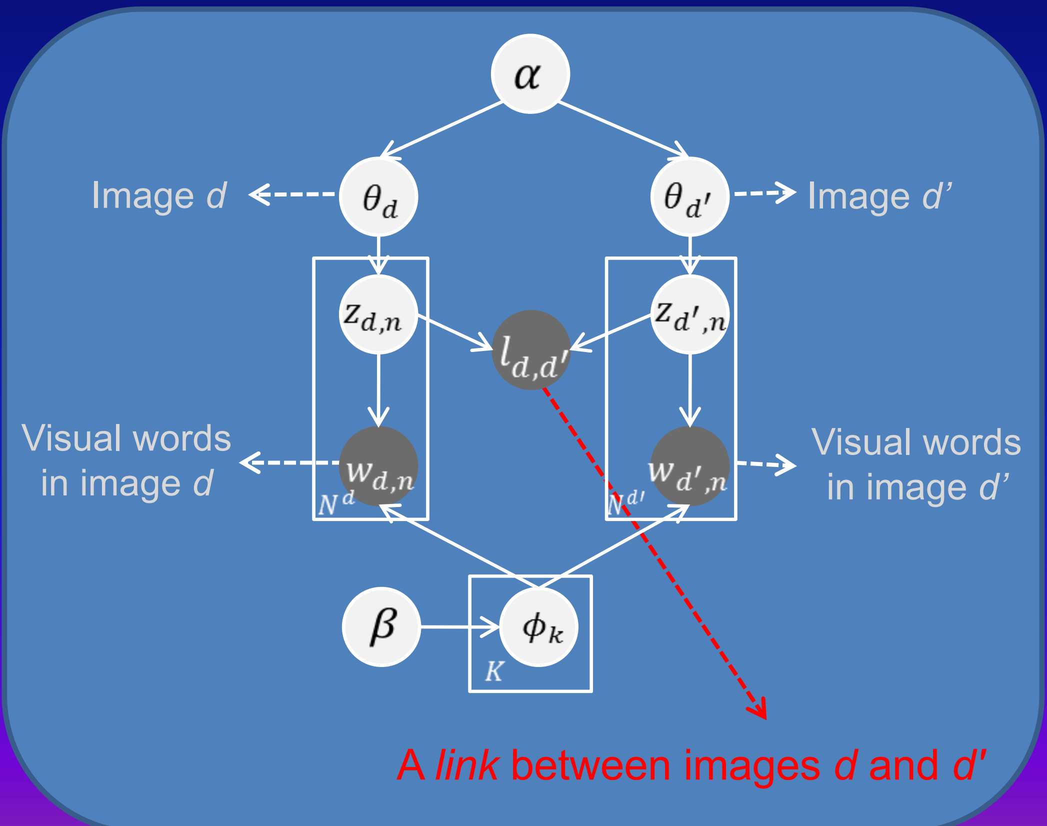
Building an image network



- Building image links according to the correlation of two tag sets

Visual topic network (VTN)

$$p(\boldsymbol{W}, \boldsymbol{L}, \boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi} | \alpha, \beta) = \prod_{d \in D} p(\theta_d | \alpha) \prod_{n \in N_d} p(z_{dn} | \theta_d)\, p(w_{dn} | \phi_{z_{dn}}) \prod_{k \in K} p(\phi_k | \beta) \prod_{d,d'} \psi(l_{d,d'} | z_d, z_{d'})$$



- 1. For each topic *k* :
  - (a) Draw topic distribution over codebook ϕ_k
- 2. For each image *d* :
  - (a) Draw topic proportions θ_d
  - (b) For each visual word:
    - (i) Select a topic $z_{d,n}$
    - (II) Draw a visual word $w_{d,n}$
- 3. For each link *l* :
  - (a) Draw a link $l_{d,d'}$ from a *link probability function* $\psi(l_{d,d'} \mid \mathbf{z}_d, \mathbf{z}_{d'})$

Multimodal image understanding

- Image content and contextual information are fused in our VTN
  - Image link is modeled with user text tags
  - Image content is modeled with visual words and latent topics

The link probability function

- $\psi(l_{d,d'} \mid \mathbf{z}_d, \mathbf{z}_{d'})$ encourages two images that are positively related to have similar representations
- Option 1: Define the similarity between two image representations as

$$s_{d,d'} = \sum_{k=1}^{K} min(\mathbf{z}_{d,k}, \mathbf{z}_{d',k}),$$

- We model the link with a binary variable $l_{d,d'} \in \{0,1\}$

Positive relation	$\psi(l_{d,d'} = 1 \mid \mathbf{z}_d, \mathbf{z}_{d'}) = s_{d,d'}$
Negative relation	$\psi(l_{d,d'} = 0 \mid \mathbf{z}_d, \mathbf{z}_{d'}) = 1 - s_{d,d'}$

The link probability function

- Option 2: model the link with a multi-valued variable $l_{d,d'} \in \{0, \dots, W\}$

$$l_{d,d'} = \boxed{y_d} \boxed{R} \boxed{y_{d'}^T} \in \Re$$

The tag vector

Correlation matrix

Quantized with W thresholds

$$\Big\downarrow$$

$$l_{d,d'} \in \{0,1,2, \dots, W\}$$

- ***A Binomial function*** is adopted as the link probability function

$$\boxed{\psi(l_{d,d'} \mid s_{d,d'}) = \binom{W}{l_{d,d'}} s_{d,d'}^{\,l_{d,d'}} \left(1 - s_{d,d'}\right)^{W - l_{d,d'}}}$$

  - The probability will be higher if $l_{d,d'}$ and $s_{d,d'}$ are both larger or smaller.
  - If W=1, multi-valued links are reduced to binary links
- In the Relation Topic Model (RTM) model [Chang & Blei 2009], the link variable is either 1 or unobserved

Inference

- Objective: estimate $\boldsymbol{Z}, \boldsymbol{\Theta}, \boldsymbol{\Phi}$ given all observed $\boldsymbol{W}$ and $\boldsymbol{L}$
- Collapsed Gibbs sampling method for VTN
  - Given the hyper-parameters, compute the posterior distribution of the latent variables via Collapsed Gibbs sampler

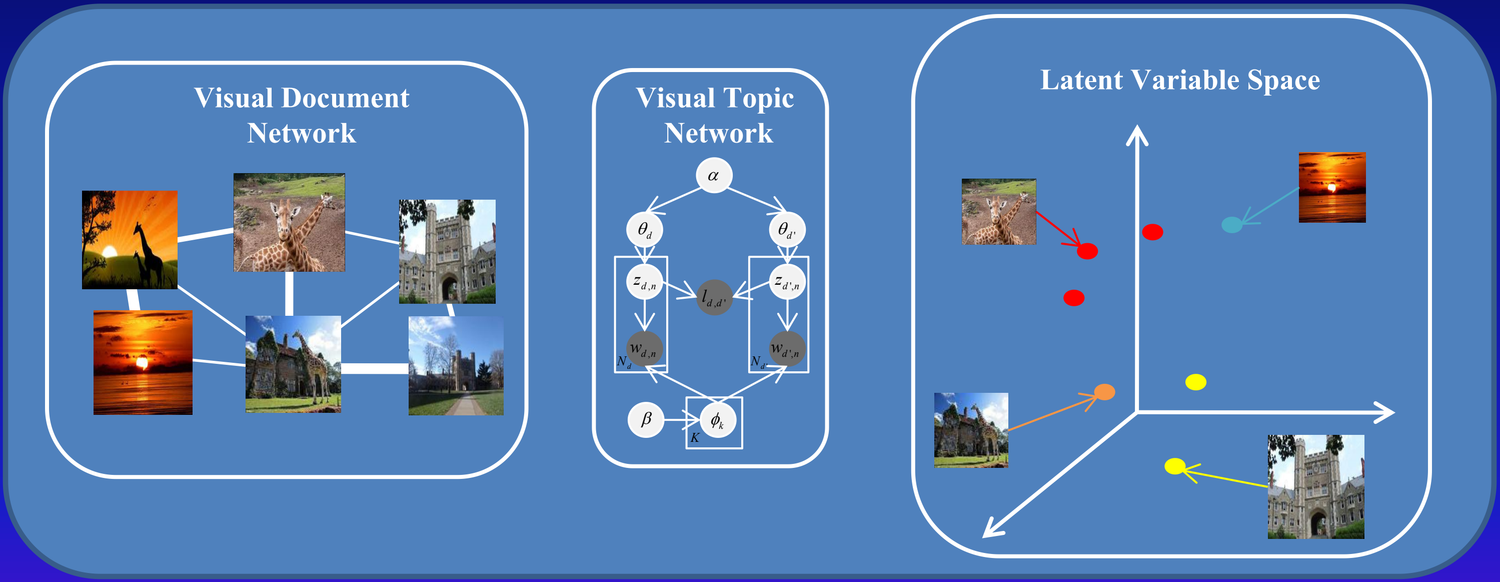
$$p\left(z_{dn} = k \middle| \boldsymbol{Z}^{-dn}, \boldsymbol{W}, \boldsymbol{L}, \alpha, \beta, \tau\right) \propto \left(\alpha + m_{d,k}^{-dn}\right) \frac{n_{k,w_{dn}}^{-dn} + \beta}{\sum_w n_{k,w_{dn}}^{-dn} + W\beta} \boxed{\prod_{d,d'} \frac{\psi(l_{d,d'} | \mathbf{z}_d, \mathbf{z}_{d'})}{\psi(l_{d,d'} | \mathbf{z}_d^{-dn}, \mathbf{z}_{d'})}}$$

  - Given the per-word topic assignments, estimate the image representation

$$\phi_k(w) = \frac{n_{w,k} + \beta}{\sum_w n_{w,k} + W\beta} \qquad \boxed{\theta_d(k) = \frac{n_{d,k} + \alpha}{\sum_k n_{d,k} + K\alpha}}$$

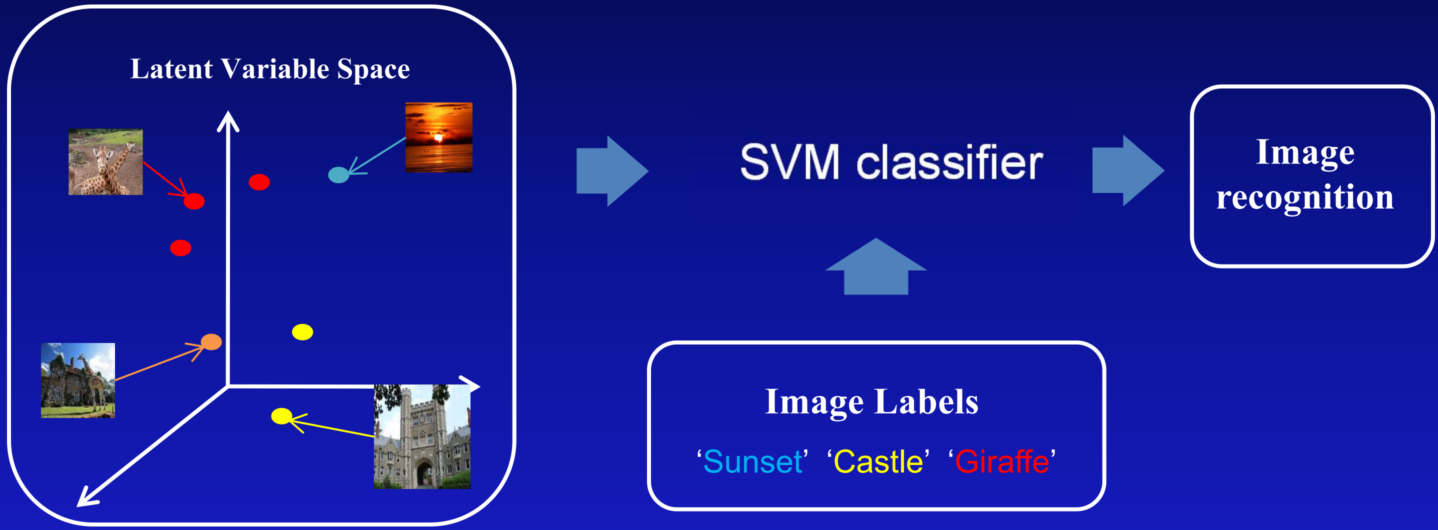
Visual topic network

- Jointly learn the image representations



- Two images are more likely to have similar representations if they have a positive relation

VTN is unsupervised



- Joint representation learning of all images
- Middle-level fusion of visual and textual information
  - Better than Pre-fusion (feature fusion) and Post-fusion (score fusion)

Supervised VTN (sVTN)

- What if we also observe the labels of the training images?

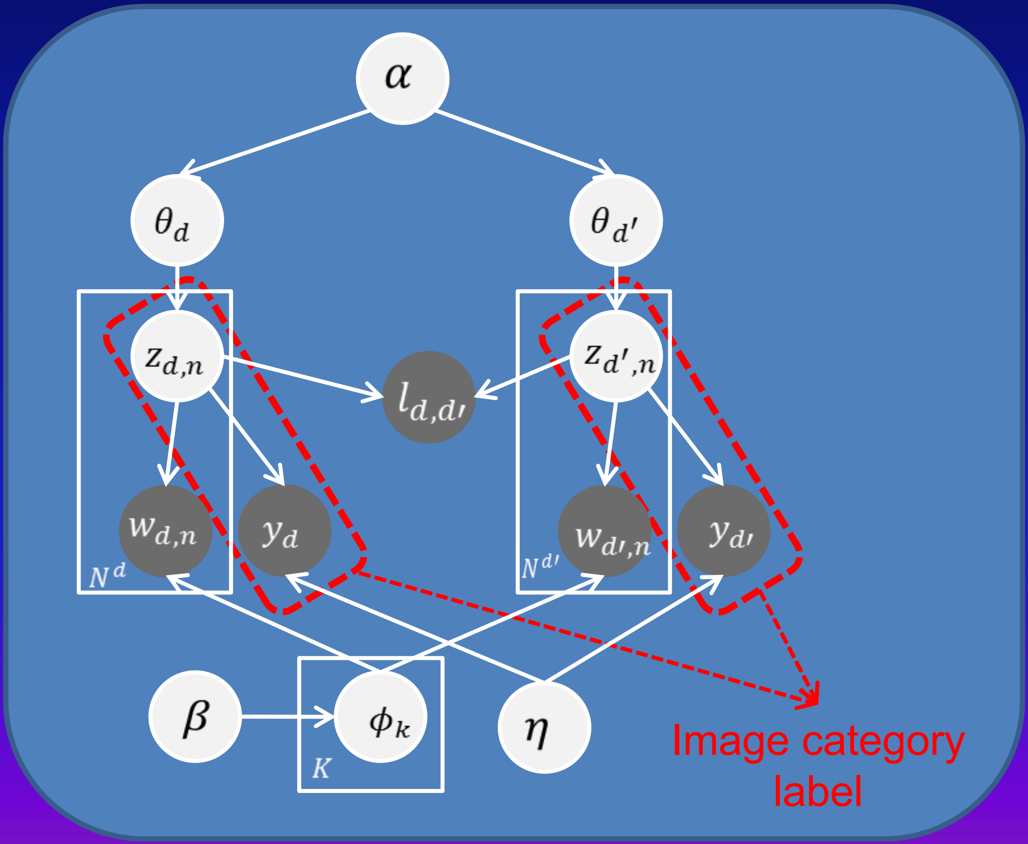
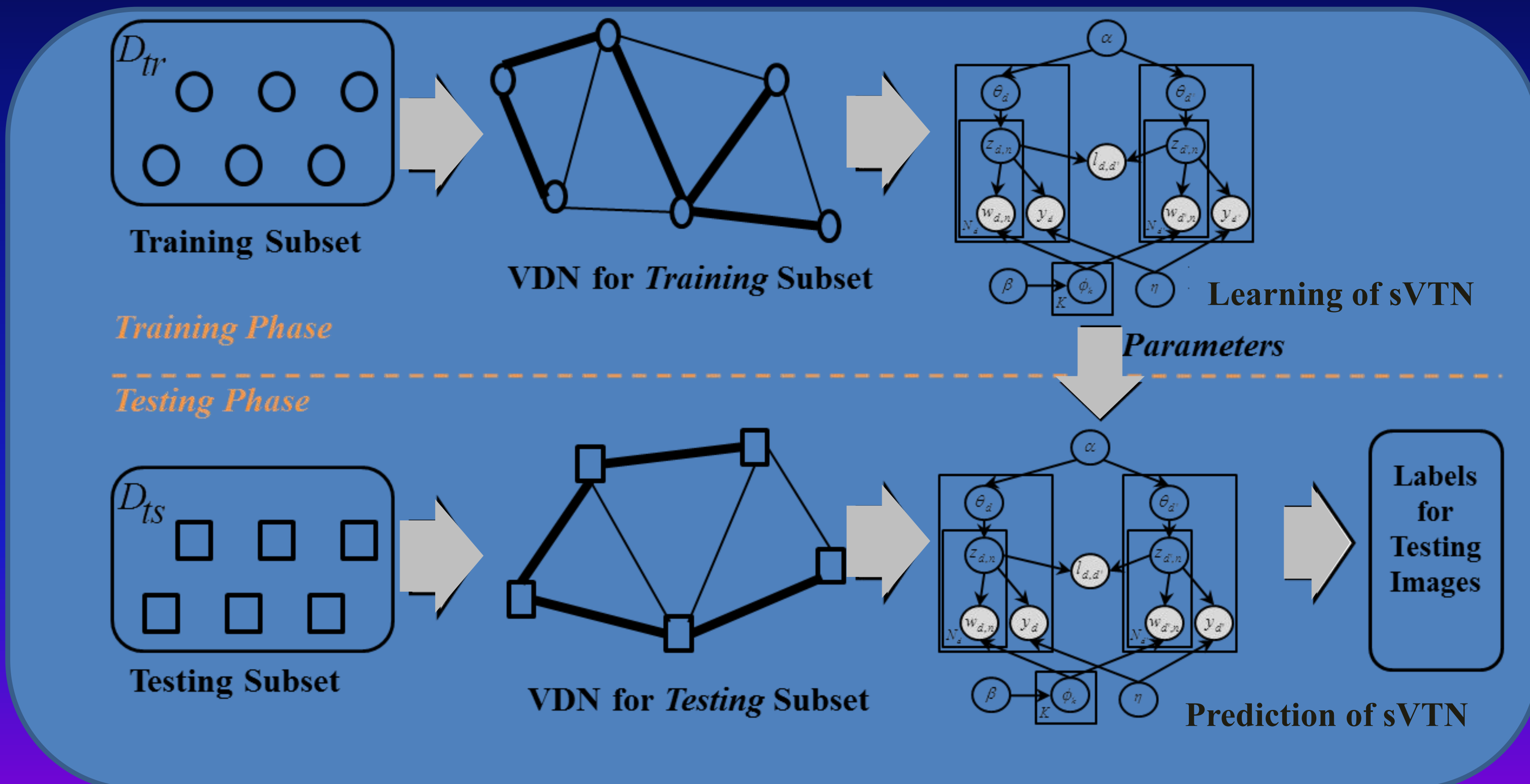
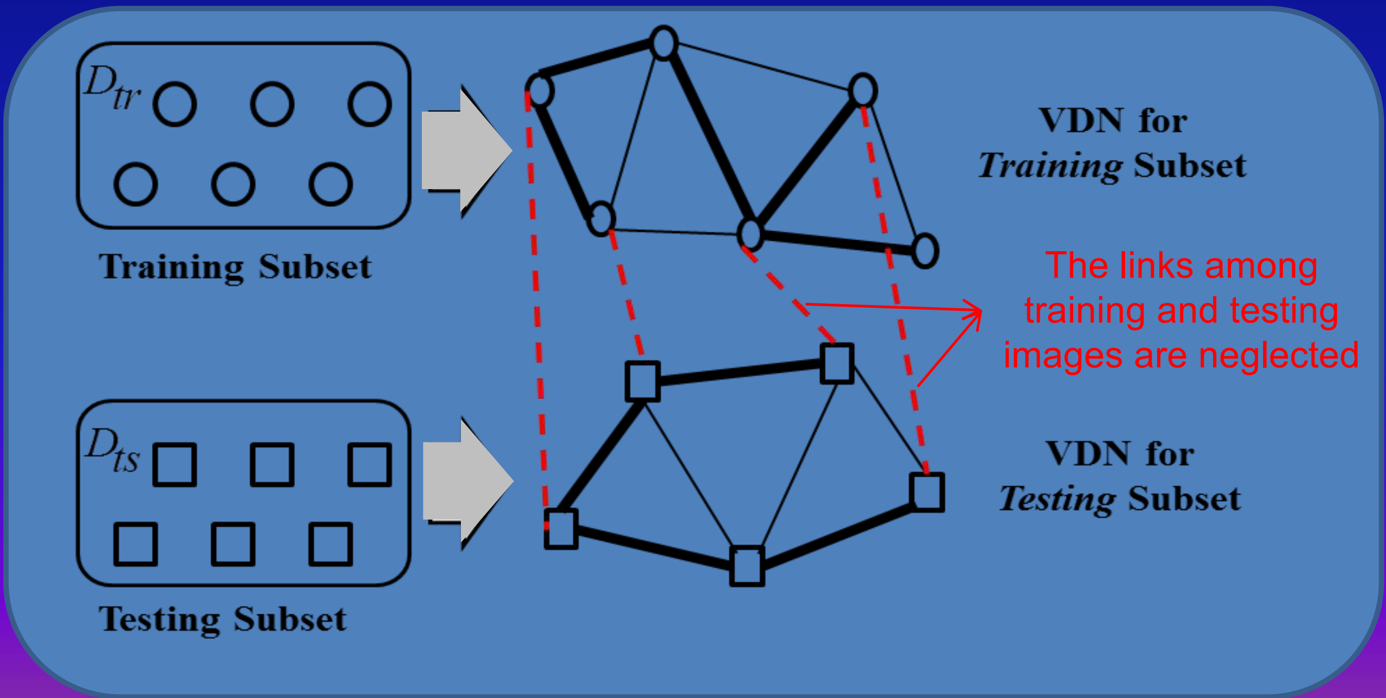


Image recognition with sVTN



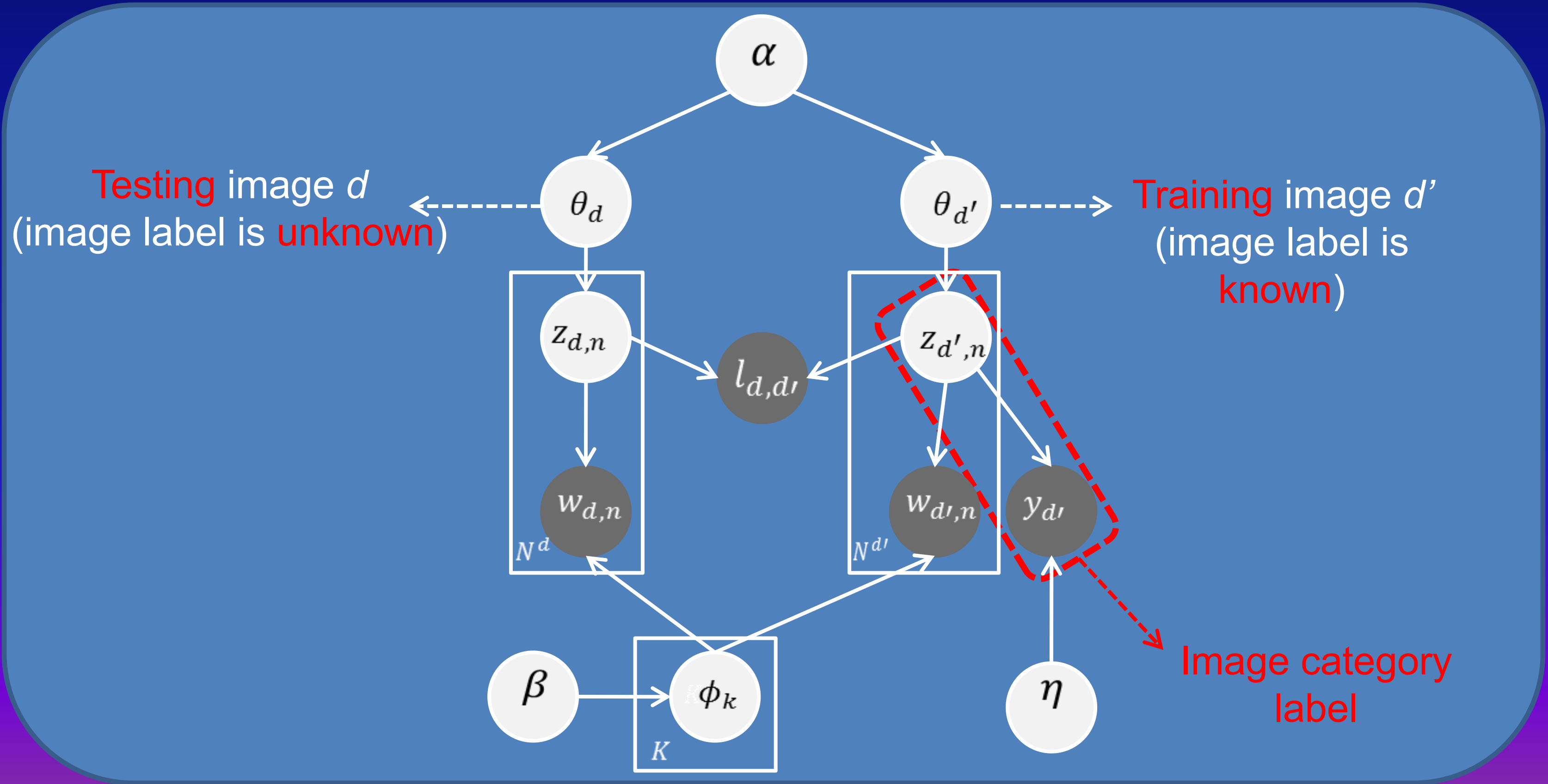
Semi-supervised VTN (ssVTN)

- Image relations within the training and within the testing images are separately modeled in a supervised VTN
- The relations among training and testing images are not leveraged in a sVTN



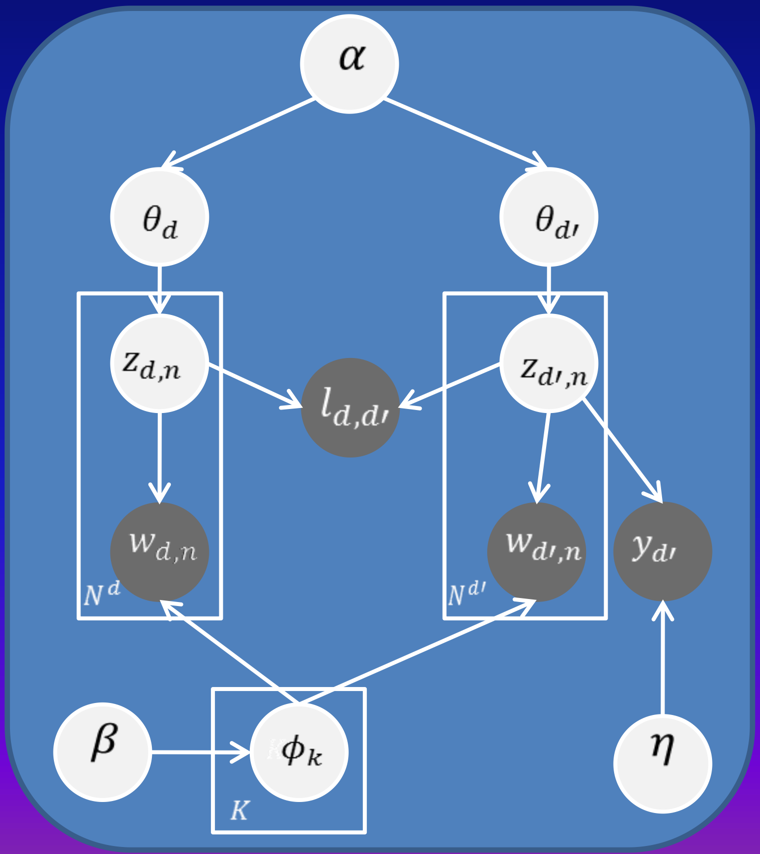
Semi-supervised VTN (ssVTN)

- The relations both within and among the training and testing images are modeled in a ssVTN



Semi-supervised VTN (ssVTN)

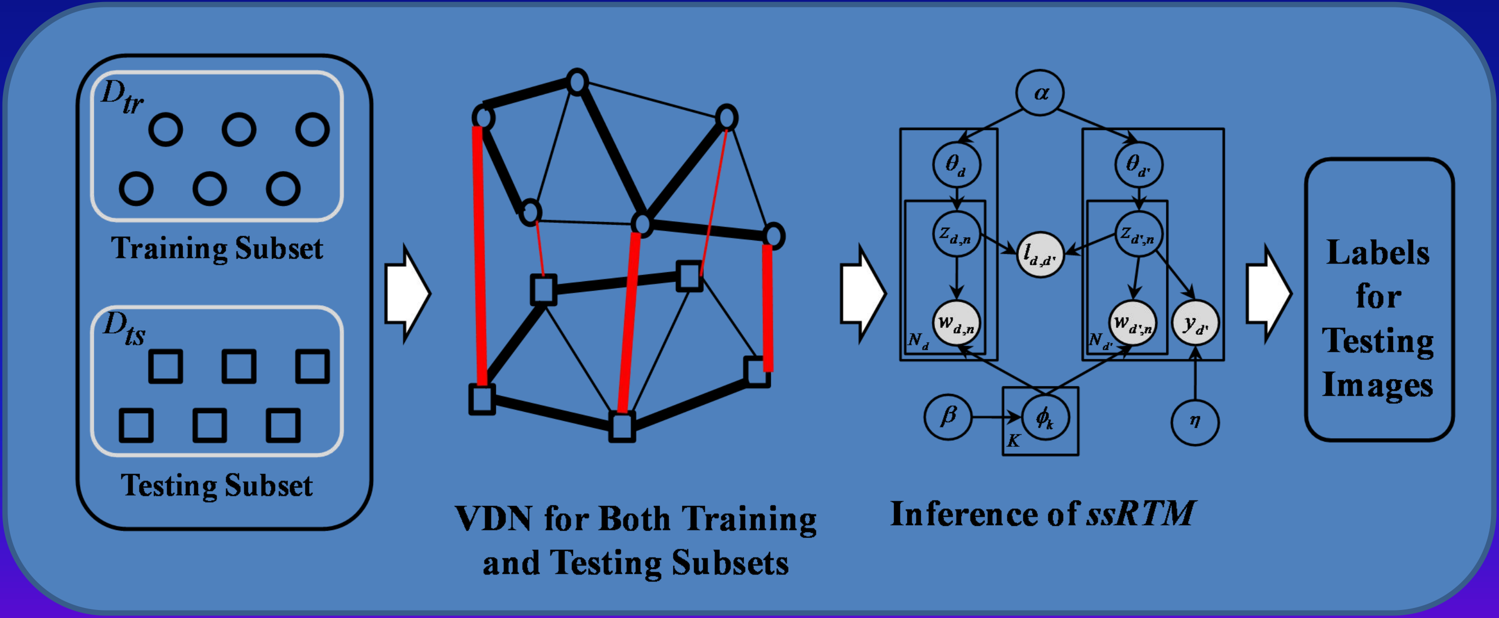
$$p(\boldsymbol{W}, \boldsymbol{L}, \boldsymbol{Y} | \alpha, \beta, \eta) = \prod_{d \in D} p(\theta_d | \alpha) \prod_{n \in N_d} p(z_{dn} | \theta_d) p(w_{dn} | \phi_{z_{dn}}) \prod_{k \in K} p(\phi_k | \beta) \prod_{d,d'} \psi(l_{d,d'} | z_d, z_{d'}) \prod_{y^d} p(y_d | z_d, \eta)$$



- 1. For each topic k :
  - (a) Draw topic distribution over codebook ϕ_k
- 2. For each image d :
  - (a) Draw topic proportions θ_d
  - (b) For each visual word:
    - (i) Select a topic $z_{d,n}$
    - (II) Draw a visual word $w_{d,n}$
  - If y_d is observed, draw image category label $y_d \sim p(y_d | z_d, \eta)$
- 3. For each link l :
  - (a) Draw a link $l_{d,d'}$ from a link probability function $\psi(l_{d,d'} | z_d, z_{d'})$

The flowchart of image recognition

- Transductive learning
  - training and testing images are modeled at the same time



Model learning

- Model definition: the joint distribution of visual words, image relations, and image category label is given by

$$p(\boldsymbol{W}, \boldsymbol{L}, \boxed{\boldsymbol{Y}} | \alpha, \beta, \boxed{\eta}) = \prod_{d \in D} p(\theta_d | \alpha) \prod_{n \in N_d} p(z_{dn} | \theta_d)\, p(w_{dn} | \phi_{z_{dn}}) \prod_{k \in K} p(\phi_k | \beta) \prod_{d, d'} \psi(l_{d,d'} | z_d, z_{d'}) \prod_{y^d} \boxed{p(y_d | z_d, \eta)}$$

Topic proportions

Topic assignments

Visual words

Topic distribution

Image relations

Image label

- Model learning
  - Given the W, L, Y, infer the θ_d of all images, learn the parameter η, and estimate category labels y_d for unlabeled test images

Model learning

- Collapsed Gibbs sampling for sVTN and ssVTN

➢ **Iteratively repeat next two steps**:

➢ **Model inference**: given the hyper-parameters, compute the posterior distribution of the latent variables via Collapsed Gibbs sampler

$$p(z_{dn} = k \mid Z^{-dn}, W, L, \boxed{Y}, \alpha, \beta, \eta) = (\alpha + m_{d,k}^{-dn}) \frac{n_{k,w_{dn}}^{-dn} + \beta}{\sum_w n_{k,w_{dn}}^{-dn} + W\beta} \prod_{d,d'} \frac{\psi(l_{d,d'} \mid z_d, z_{d'})}{\psi(l_{d,d'} \mid z_d^{-dn}, z_{d'})} \boxed{\frac{\rho(y_d \mid z_d, \eta)}{\rho(y_d \mid z_d^{-dn}, \eta)}}$$

Difference from VTN

➢ **Parameter estimation**: given the per-word topic assignments, conduct logistic regression to obtain η according to

$$\rho(y_d = 1 \mid z_d, \eta) = \frac{1}{1+\exp(-\eta^T z_d)}$$

➢ **Obtaining image representations and category labels**: given the per-word topic assignments, estimate image representations and category labels $\theta_d(k) = \frac{n_{d,k} + \alpha}{\sum_k n_{d,k} + K\alpha}$ $\rho(y_d = 1 \mid z_d, \eta) = \frac{1}{1+\exp(-\eta^T z_d)}$

Evaluation datasets

- Two social media datasets
  - NUS-WIDE: *269,648* images, *1,000* tags, and *81* concepts.

building, sky, clouds

car, tree, villages

statue, sky, horse

elder, chair, beach, sky,

  - MIRFLICKR-25k: *25,000* images, *1,386* tags, and *23* labels.

sunset, sky, clouds, flowers

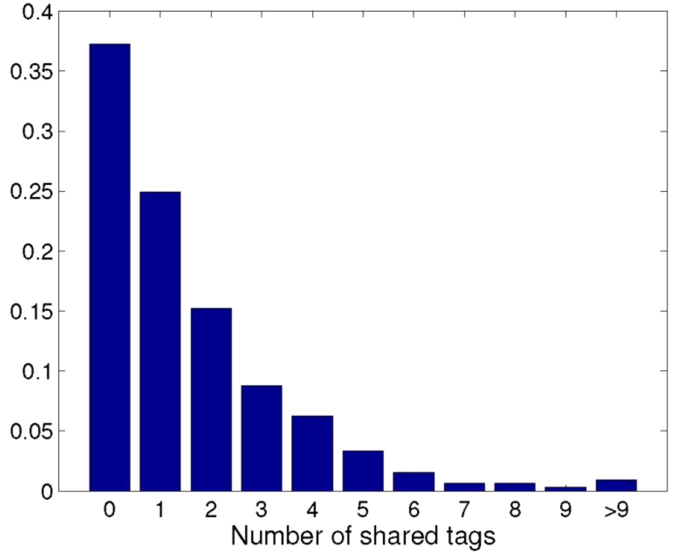
flowers

people, girl

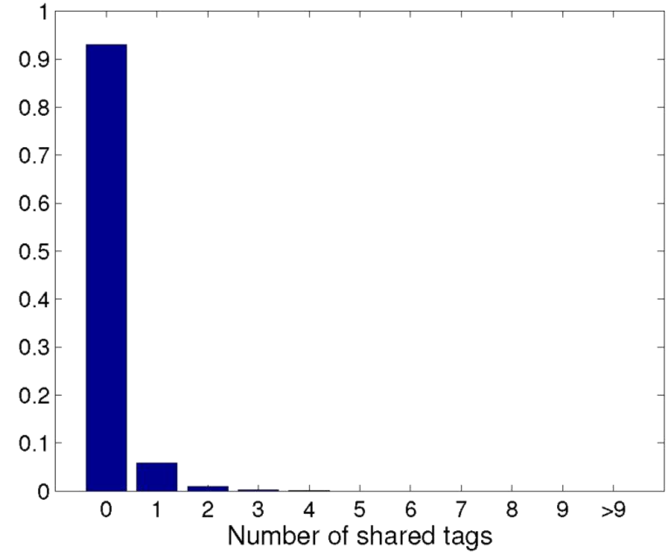
building, flowers

Evaluation datasets

- Verify that images with shared tags have close relations in semantics



For images from a specific category



For images from all the categories

Experimental results

- Image recognition

Methods	NUS-WIDE dataset	MIRFLICKR-25k dataset
BoW+SVM	70.8%	72.9%
Tag+SVM	74.4%	73.8%
BoW+Tag+SVM	75.1%	74.2%
BoW+Tag+MKL	76.2%	77.4%
LDA+SVM	72.3%	73.1%
RTM+SVM	74.1%	75.1%
sLDA	72.8%	73.8%
VTN+SVM	76.5%	78.7%
sVTN	84.2%	80.3%
ssVTN	**87.1%**	**83.5%**

Discriminative Methods

Probabilistic Model

Pre-fusion

Post-fusion

Unsupervised

Supervised

Experimental results

- Detailed comparison per concepts

Methods	NUS-WIDE (81)		MIRFLICKR-25k (23)	
	1st	2nd	1st	2nd
BoW+Tag+MKL	18	10	5	3
RTM+SVM	1	20	0	5
sLDA	2	18	0	2
VTN+SVM	23	2	6	2
sVTN	0	30	0	10
ssVTN	37	1	12	1

- ➢ For some concepts, semi-supervised model is better than supervised model
  - ➢ where the ssVTN is consistently better than sVTN over such concepts
- ➢ For some other concepts, unsupervised model is better than supervised
  - ➢ where the VTN is consistently better than the RTM over such concepts

Experimental results

- The modeling of image relation
  - For $l_{d,d'} = \boldsymbol{y}_d^T \boldsymbol{y}_{d'}$, and quantizing it as a binary variable (VTN-B)
  - For $l_{d,d'} = \boldsymbol{y}_d^T \boldsymbol{y}_{d'}$, and quantizing it as a multi-valued variable (VTN-M)
  - For $l_{d,d'} = \boldsymbol{y}_d \boldsymbol{R} \boldsymbol{y}_{d'}$, and quantizing it as a binary variable (VTN-CB)
  - For $l_{d,d'} = \boldsymbol{y}_d \boldsymbol{R} \boldsymbol{y}_{d'}$, and quantizing it as a multi-valued variable (VTN and ssRTM)

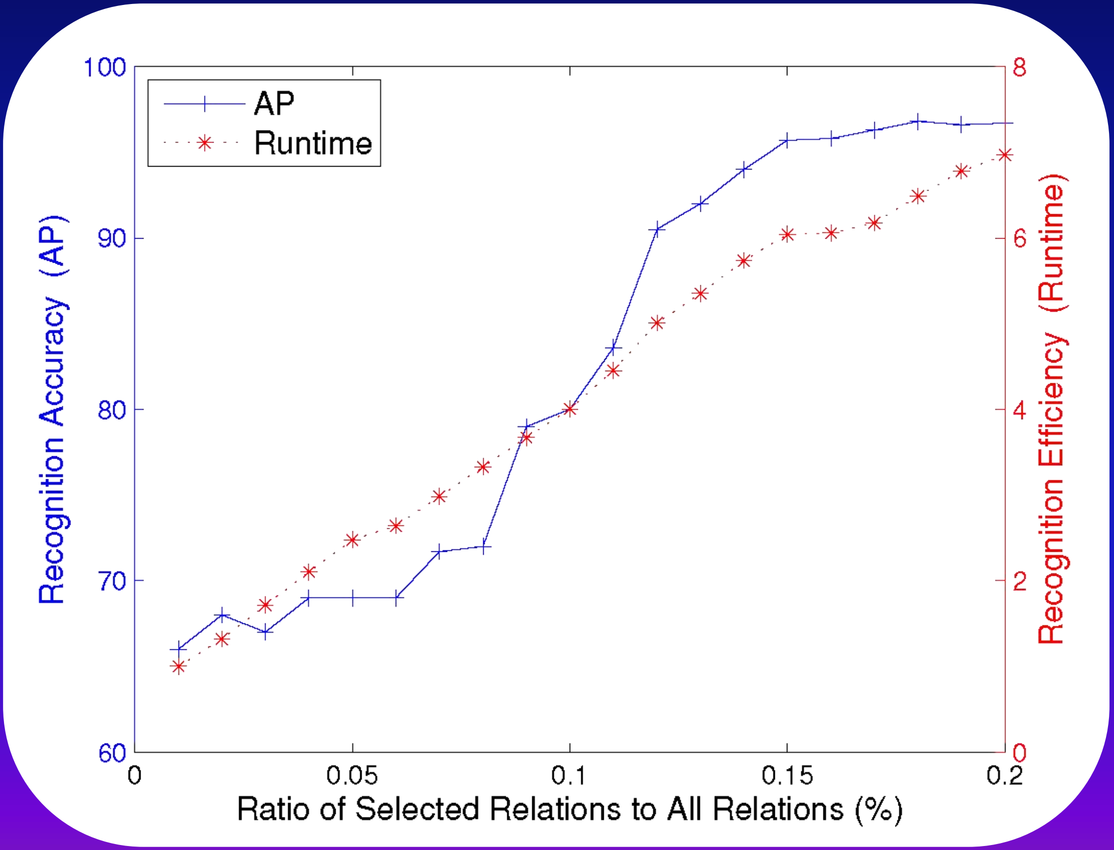
Methods	NUS-WIDE dataset	MIRFLICKR-25k dataset
VTN-B + SVM	74.6%	75.6%
VTN-M + SVM	72.8%	73.2%
VTN-CB + SVM	74.5%	75.5%
VTN + SVM	76.5%	78.7%
ssVTN	**87.1%**	**83.5%**

Unsupervised

Semi-supervised

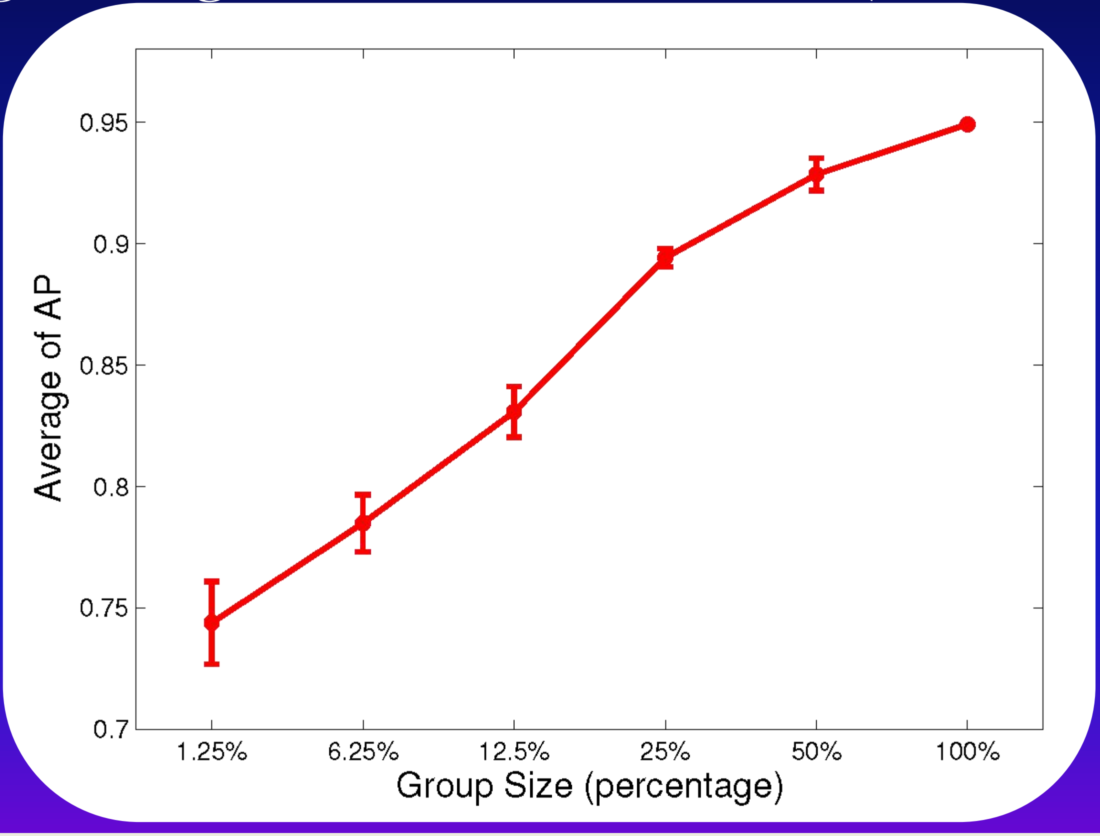
Experimental results

- The selection of image relations (NUS-WIDE)



Experimental results

- Image recognition in batch mode (NUS-WIDE)



Group size (percentage)	0%	6.25%	12.5%	25%	50%	100%
mAP	0	0.32	0.51	0.81	0.91	1

Conclusion remarks

- Be cautious when you use “end-to-end” deep learning
- Prior knowledge is important in solving computer vision
- It is all about context
  - “Semantics without context are meaningless” [Quote from Prof. Ramesh Jain]

The VC Group is recruiting

- We are hiring for both FTEs and interns

Email: ganghua@microsoft.com

CVPR2019 Bid

- Location: Long Beach, Los Angeles, CA
- Team:
  - General Chairs:
    - Song-Chun Zhu, Philip Torr, Larry Davis
  - Program Chairs:
    - Gang Hua, Abhinav Gupta, Two to be confirmed
- We hope to have your support and especially

Your Vote!

References

- [1] T. Hofmann, “Probabilistic latent semantic indexing,” *SIGIR*, 1999.
- [2] D. Blei, A. Ng and M. Jordan, “Latent Dirichlet allocation,” *J. Mach. Learn. Res*, 2003.
- [3] D. Blei and M. Jordan, “Modeling annotated data,” *SIGIR*, 2003.

- d J. McAuliffe, “Supervised topic models,” *NIPS*, 2007.
- nd D. Blei, “Relational topic models for document networks,”
.
- s and M. Steyvers, “Finding scientific topics,” *PNAS*, 2004.
- umin, J. Verbeek and C. Schmid, “Multimodal semi-supervised
mage classification,” *CVPR*, 2010.
- [8] Z. Niu, G. Hua, et al, “Semi-supervised relational topic model for weakly annotated image recognition in social media,” *CVPR*, 2014.
- [9] Z. Niu, G. Hua, et al, “Visual Topic Network: Building better image representations for images in social media,” *CVIU*, 2015.