Attribute-Based Transfer Learning for Object Categorization with zero/one Training Samples

Outline

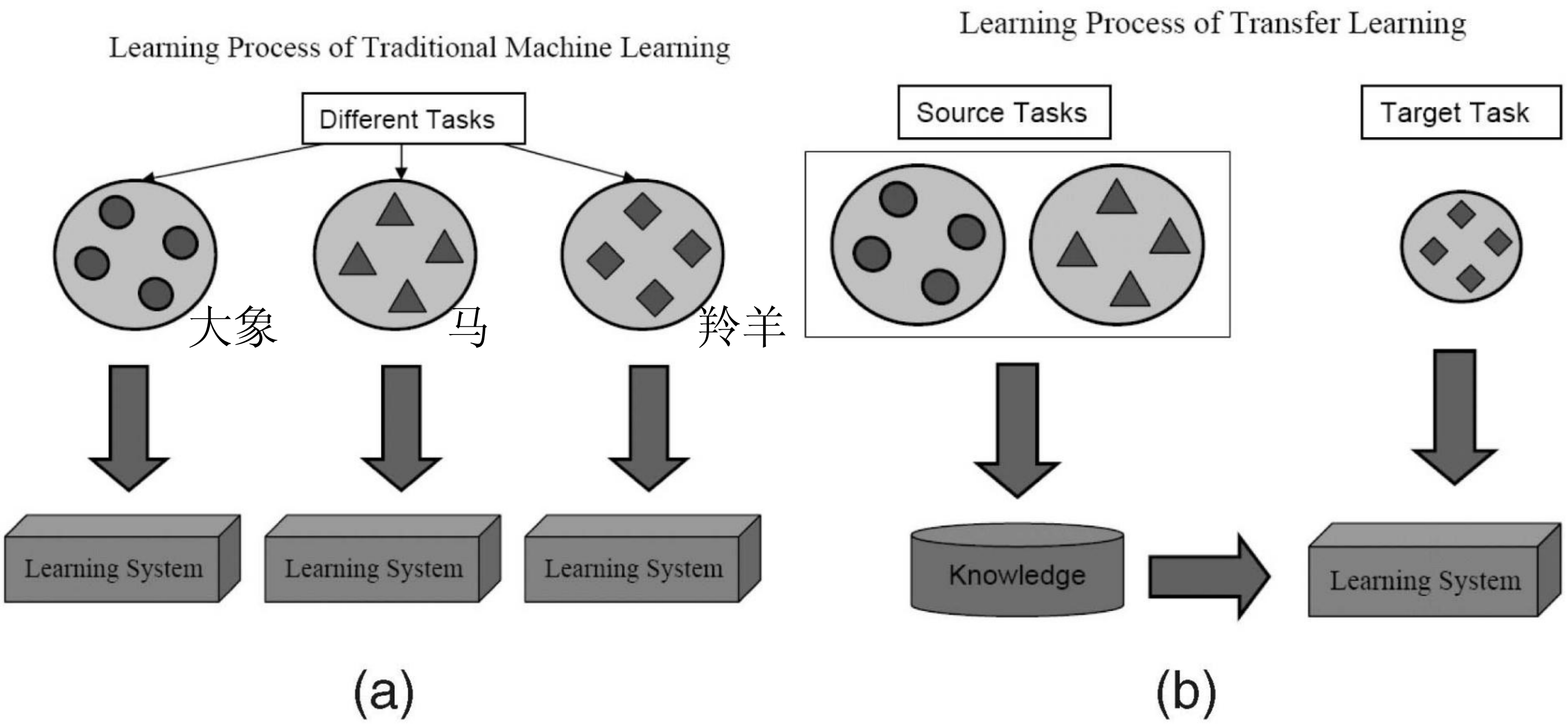
- 背景介绍
  - Transfer Learning
- Attribute-Based Transfer Learning
  - Attribute model: Author-Topic Model
  - Target Classifier: Category-Topic Model
  - Method to Transfer Learning
    - Knowledge Transfer by Synthesis of Training Examples
    - Knowledge Transfer by Information Parameter Priors
- 实验结果

摘要

- 研究one-shot 和zero-shot learning问题。
  - One-shot learning：每类只有一个训练样本
  - Zero-shot learning：没有训练样本
- 通过迁移学习解决这个问题
  - 通过物体属性(object attribute)将源类别(or known)上得到的知识迁移到目标类别上
  - 物体属性是对物体类别的高级描述，比如颜色，纹理，形状等。他们是不同类别的共有属性，可以用来将源类别的信息迁移到目标类别上
- 提出了一个基于属性的迁移学习框架
  - 首先建立一个产生式模型，针对每个属性学习它对应的图像特征的概率分布，这将被视作先验
  - 属性先验可以用来
    - □ 解决没有训练样本的分类问题(zero-shot learning)
    - □ 促进只有一个训练样本的分类问题(one-shot learning)
- 方法在Animal with Attributes 数据集的zero-shot 和one-shot任务上取得了state-of-the-art的性能

背景介绍—Transfer Learning

- Different learning processes between traditional machine learning and transfer learning



(a) traditional machine learning (b) transfer learning

背景介绍—Transfer Learning

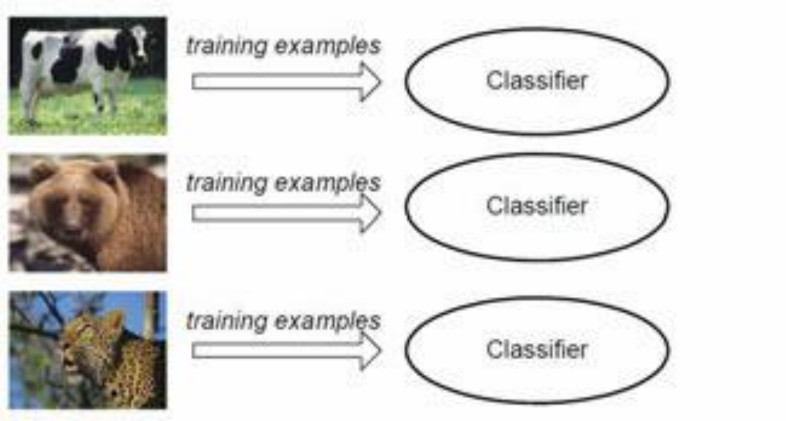
- One-Shot Learning
  - Only one training example per category
- Zero-Shot Learning
  - No training example for the target category
- Solutions
  - 将在源类别上得到的知识迁移到目标类别上
  - 相当于增加了目标类别的训练样本

背景介绍—Transfer Learning

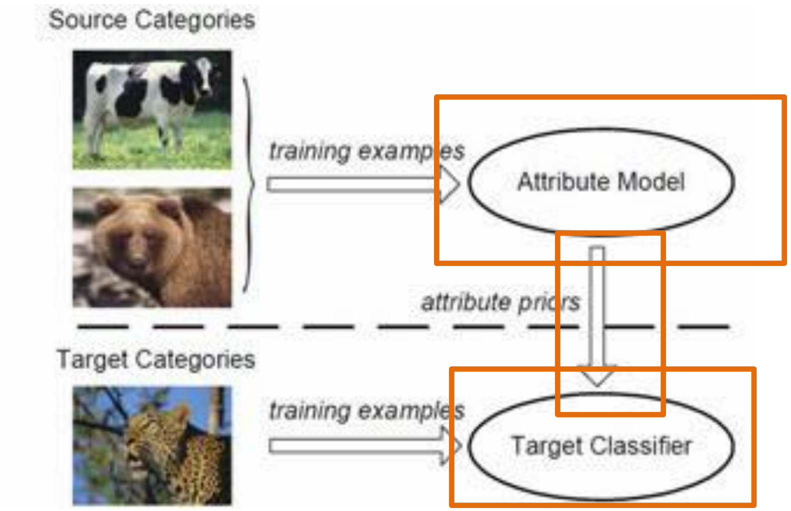
- 物体分类上的knowledge transfer 方法可以分成三类
  - By sharing either features or model parameters, or context information
  - Considering ontological knowledge of object similarity
  - Employing the object attributes
    - □ Semantic knowledge: high-level descriptions about properties of object categories such as color, texture, shape, parts, context.

Attribute-Based Transfer Learning Framework

- Different learning processes between
  - (a) traditional machine learning
  - (b) transfer learning



(a)



(b)

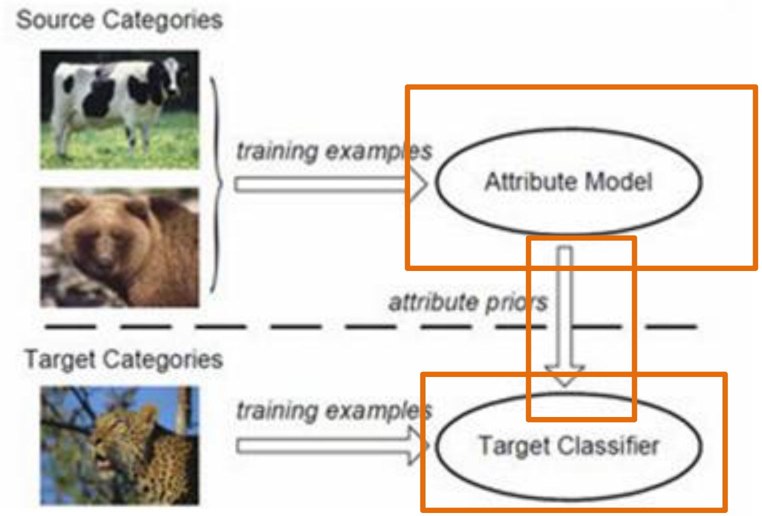
Attribute: represent common properties across different categories

Attribute-Based Transfer Learning Framework

- An example
  - people who have never seen a zebra (斑马) still could reliably identify an image of zebra if we tell them that “a zebra is a wild quadrupedal (四脚的) with distinctive white and black strips (黑白相间条纹) living on African savannas (非洲草原)”.
  - Since they have prior knowledge about the related object attributes, e.g., ***quadrupedal, white and black strips, African savannas***, they can transfer them to facilitate prediction of unseen categories.

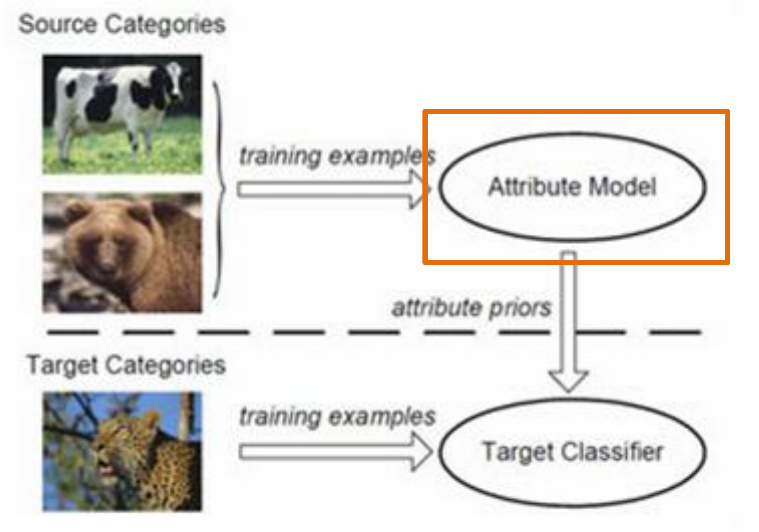
Attribute-Based Transfer Learning Framework

- 3 key components in an attribute-based transfer learning system
  - Attribute model
    - 属性和图像特征之间的联合概率密度分布
  - Target classifier
  - Method to transfer attribute prior



Attribute-Based Transfer Learning

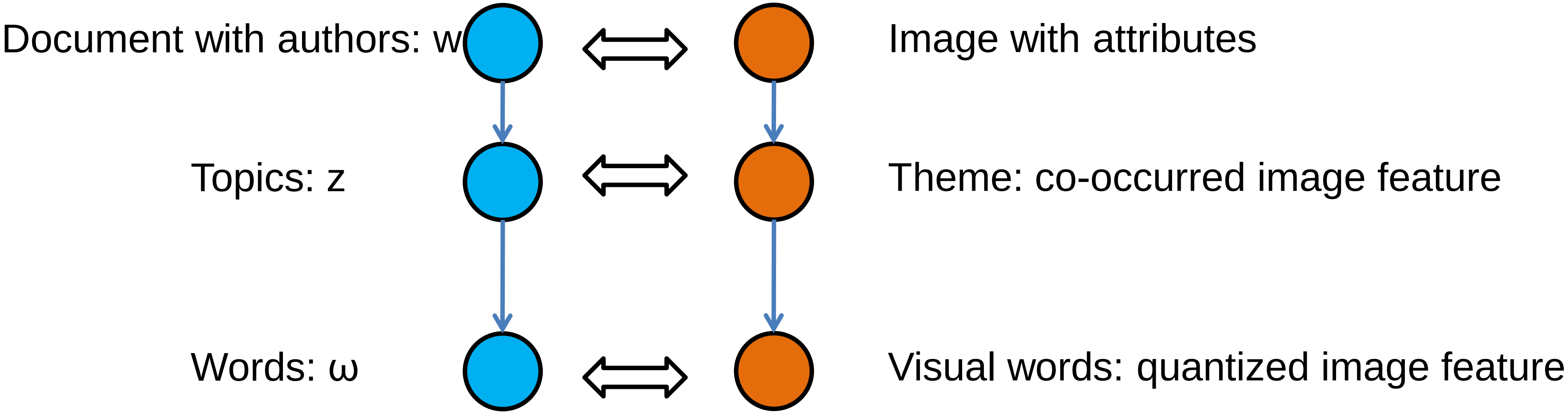
- Framework



(a) Attribute Model: Author-Topic Model
(b) Target Classifier: Category-Topic Model
(c) Target Classifier with transfered knowledge

Attribute Model

- Employ **Author-Topic** model as attribute model
  - Relationship of Category-Attribute is similar to that of Document-Author.



Document with authors: w

Image with attributes

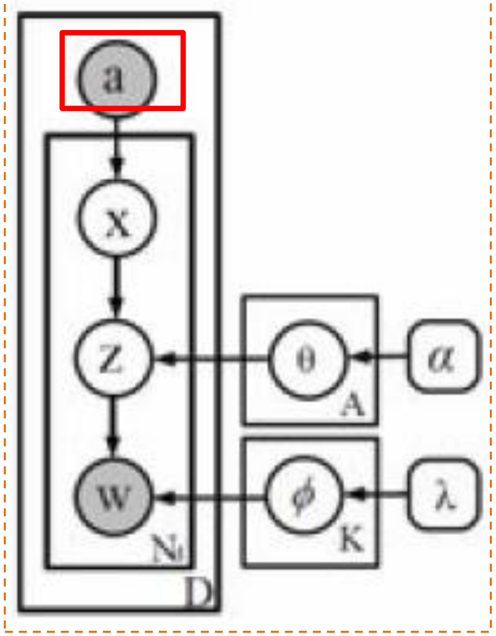
Topics: z

Theme: co-occurred image feature

Words: ω

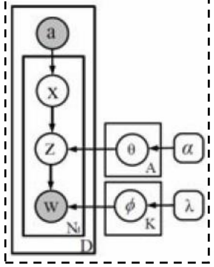
Visual words: quantized image feature

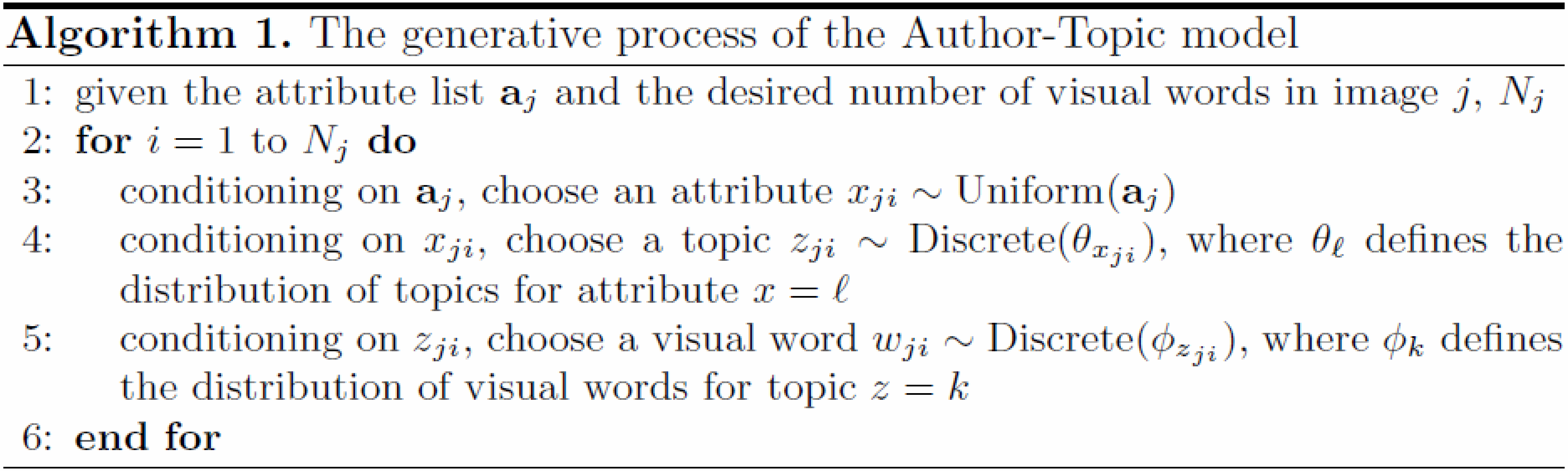
Attribute Model



- AT model is a generative model
  - **a**$_j$: a list of attributes of Image j
  - x: one attribute
    - □ modeled by a discrete distribution of *K* topics, which parameterized by a *K*-dim vector $\theta_\ell = (\theta_{\ell 1}, ..., \theta_{\ell K})$
  - z: one topic
    - □ modeled by a discrete distribution of *W* codewords, which parameterized by a *W*-dim vector $\phi_k = (\phi_{k1}, ..., \phi_{kW})$
  - w: one word
  - Symmetric Dirichlet priors are placed on θ and ϕ, with $\theta_\ell \sim \text{Dirichlet}(\alpha)$ and $\phi_k \sim \text{Dirichlet}(\lambda)$

Attribute Model



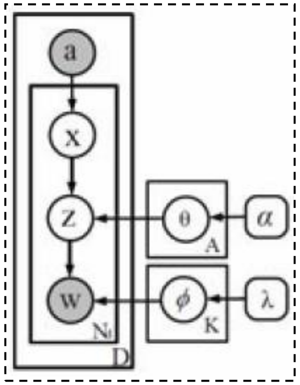


Algorithm 1. The generative process of the Author-Topic model

```
given the attribute list $\mathbf{a}_j$ and the desired number of visual words in image $j$, $N_j$
for $i = 1$ to $N_j$ do
    conditioning on $\mathbf{a}_j$, choose an attribute $x_{ji} \sim \text{Uniform}(\mathbf{a}_j)$
    conditioning on $x_{ji}$, choose a topic $z_{ji} \sim \text{Discrete}(\theta_{x_{ji}})$, where $\theta_\ell$ defines the distribution of topics for attribute $x = \ell$
    conditioning on $z_{ji}$, choose a visual word $w_{ji} \sim \text{Discrete}(\phi_{z_{ji}})$, where $\phi_k$ defines the distribution of visual words for topic $z = k$
end for
```

- 离散概率密度函数 $\theta_\ell = (\theta_{\ell 1}, ..., \theta_{\ell K})$ 和 $\phi_k = (\phi_{k1}, ..., \phi_{kW})$ 本身服从Symmetric Dirichlet分布
- 我们需要求解 $\theta_\ell \sim \text{Dirichlet}(\alpha)$ and $\phi_k \sim \text{Dirichlet}(\lambda)$
  - $\theta_\ell = (\theta_{\ell 1}, ..., \theta_{\ell K})$
  - $\phi_k = (\phi_{k1}, ..., \phi_{kW})$

Attribute Model



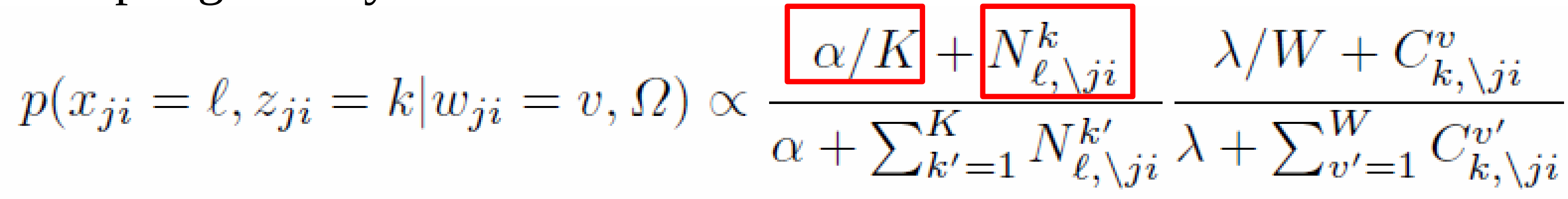
- Goal
  - Joint Distribution

    $$p(x,z,\omega) = p(x)p(z|x)p(\omega|x,z) = p(x)p(z|x)p(\omega|z)$$

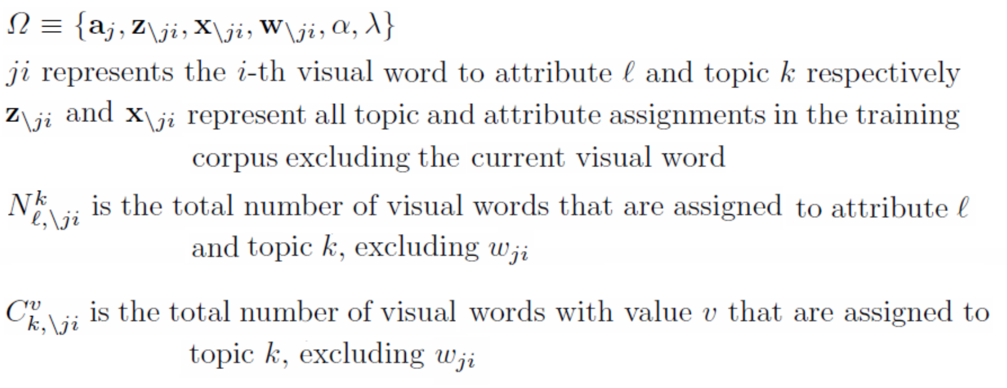
  - For discrete we need
    - □Identify the values of θ and ϕ given training corpus
- How?
  - Gibbs sampling
    - □Sampling (*x*, *z*, *ω*), since *ω* are observed data, they can be used directly, rather than sampling

Attribute Model

- Gibbs sampling
  - Sampling (*x*, *z*, *ω*)
    - □Only need to sample (*x*, *z*)
    - □since *ω* are observed data, they can be used directly, rather than sampling
    - □Sampling density:



$$p(x_{ji}=\ell, z_{ji}=k|w_{ji}=v, \Omega) \propto \frac{\boxed{\alpha/K}+\boxed{N^{k}_{\ell,\backslash ji}}}{\alpha+\sum_{k'=1}^{K} N^{k'}_{\ell,\backslash ji}} \frac{\lambda/W+C^{v}_{k,\backslash ji}}{\lambda+\sum_{v'=1}^{W} C^{v'}_{k,\backslash ji}}$$



$\Omega \equiv \{\mathbf{a}_j, \mathbf{z}_{\backslash ji}, \mathbf{x}_{\backslash ji}, \mathbf{w}_{\backslash ji}, \alpha, \lambda\}$

ji represents the i-th visual word to attribute ℓ and topic k respectively

$\mathbf{z}_{\backslash ji}$ and $\mathbf{x}_{\backslash ji}$ represent all topic and attribute assignments in the training corpus excluding the current visual word

$N^{k}_{\ell,\backslash ji}$ is the total number of visual words that are assigned to attribute ℓ and topic k, excluding w_{ji}

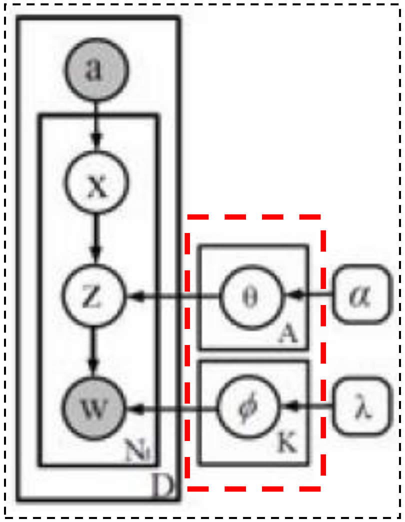
$C^{v}_{k,\backslash ji}$ is the total number of visual words with value v that are assigned to topic k, excluding w_{ji}

More details found in: **Learning author-topic models from text corpora**. TIS 09.

Attribute Model

- Gibbs sampling
  - Sampling (*x*, *z*, *ω*)
  - Posterior mean

$$\hat{\theta}_{\ell k} = \frac{\alpha/K + N_\ell^k}{\alpha + \sum_{k'=1}^{K} N_\ell^{k'}}, \quad \hat{\phi}_{kv} = \frac{\lambda/W + C_k^v}{\lambda + \sum_{v'=1}^{W} C_k^{v'}}$$



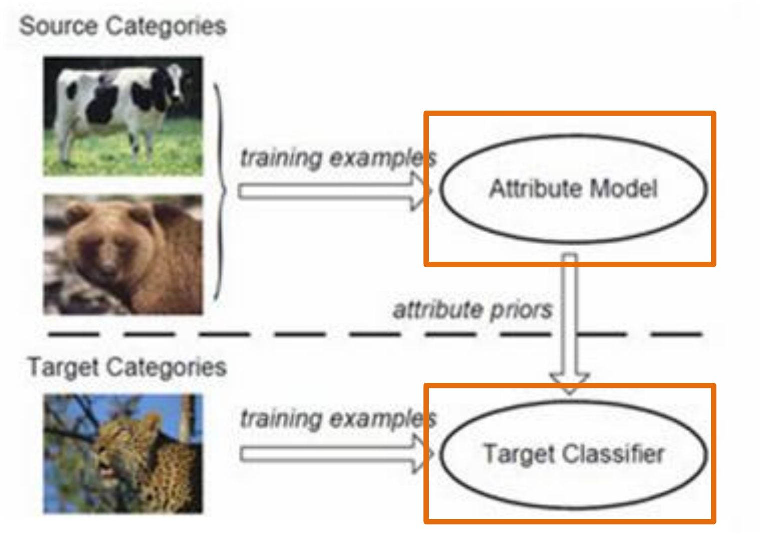
corresponding to the attribute model

$\hat{\theta}_{\ell k}$: probability of topic *k* on attribute *l*

$\hat{\phi}_{kv}$: probability of word *v* on topic *k*

Attribute-Based Transfer Learning

- Framework



(a) Attribute Model: Author-Topic Model
(b) Target Classifier: Category-Topic Model
(c) Target Classifier with transfered knowledge

Target Classifier

- Target Classifier
  - SVM
  - AT model if attribute list is unique in each category
  - **Category-Topic Model**

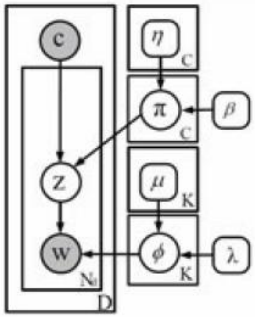
Target Classifier

- Category-Topic Model
  - 只有一个属性的AT model(label information)

$$p(z_{ji} = k | w_{ji} = v, c_j = m, \Omega) \propto \frac{\beta/K + M^{k}_{m,\backslash ji}}{\beta + \sum_{k'=1}^{K} M^{k'}_{m,\backslash ji}} \frac{\lambda/W + C^{v}_{k,\backslash ji}}{\lambda + \sum_{v'=1}^{W} C^{v'}_{k,\backslash ji}}$$

π: $$\hat{\pi}_{mk} = \frac{\beta/K + M^{k}_{m}}{\beta + \sum_{k'=1}^{K} M^{k'}_{m}}$$

Φ: $$\hat{\phi}_{kv} = \frac{\lambda/W + C^{v}_{k}}{\lambda + \sum_{v'=1}^{W} C^{v'}_{k}}$$



  - For a test image $\mathbf{w}_t = \{w_{t1}, ..., w_{tN_t}\}$, by choosing the target classifier that yields the highest likelihood

$$p(\mathbf{w}_t | c = m, \mathcal{D}^{\text{train}}) \approx \prod_{i=1}^{N_t} \sum_{k=1}^{K} \hat{\phi}_{kw_{ti}} \hat{\pi}_{mk}$$

Target Classifier

- Author-Topic model
  - 如果每类的属性列表是唯一的(unique)，AT model也可以用作Target Classifier
    - □classify a new image by maximum likelihood criterion
  - 假设我们从源类别学习到了A个属性

$$\theta_\ell \text{ for every } \ell = 1, \ldots, A$$

  - 分类

$$p(\mathbf{w}_t | c = m, a_m, \mathcal{D}^{\text{train}}) \approx \prod_{i=1}^{N_t} \sum_{k=1}^{K} \hat{\phi}_{kw_{ti}} \left(\frac{1}{A_m} \sum_{\ell \in a_m} \hat{\theta}_{\ell k} \right) \equiv \prod_{i=1}^{N_t} \sum_{k=1}^{K} \hat{\phi}_{kw_{ti}} \tilde{\pi}_{mk}$$

$\mathbf{a}_m$ is the attribute list associated to a target category $c = m$

A_m the length of $\mathbf{a}_m$

Pseudo weight: $\tilde{\pi}_{mk} \equiv \left(\frac{1}{A_m} \sum_{\ell \in a_m} \hat{\theta}_{\ell k} \right)$

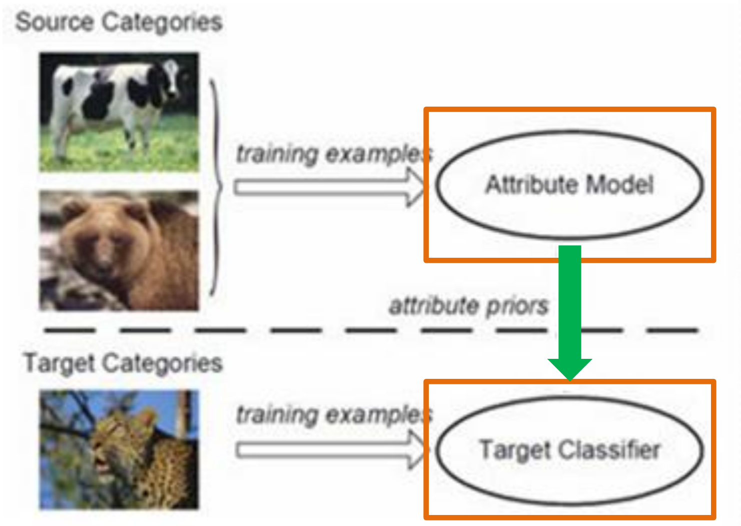
This pseudo weight can be viewed as the prior of *π*π_m
before we see the real training examples of the new category

Target Classifier

- Target Classifier
  - SVM
    - □Have to see some training examples of target categories
  - AT model if attribute list is unique in each category
    - □Can be used to deal with zero-shot learning
    - □Ineffective for one-shot learning
  - Category-Topic Model
    - □Have to see some training examples of target categories
    - □Can not be used in zero-shot learning

Attribute-Based Transfer Learning

- Framework



(a) Attribute Model: Author-Topic Model
(b) Target Classifier: Category-Topic Model
(c) Target Classifier with transferred knowledge

Method to Transfer

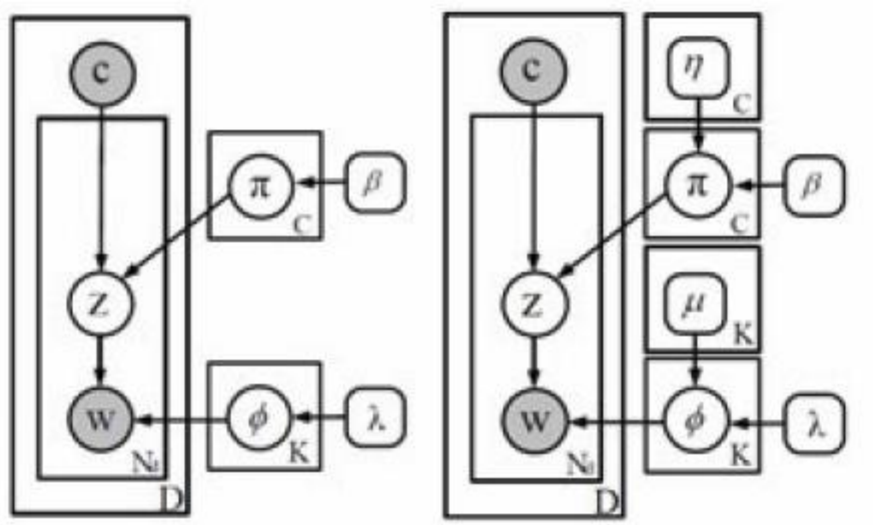
- Problems
  - Cannot be applied in zero-shot learning
  - Ineffective in one-shot learning
- Method to Transfer
  - Knowledge Transfer by Synthesis of Training Examples
  - Knowledge Transfer by Information Parameter Priors

Method to Transfer

- Knowledge Transfer ： 合成训练样本
  - 合成目标类别的训练样本
    - 首先，通过源类别学习属性模型
    - 然后，对于每个目标类别，用前面的产生过程产生S个训练样本。所利用的参数$\hat{\theta}$ *and* $\hat{\varphi}$ 为与目标类别相关属性相应的估计参数
    - S反映了对attribute prior的confidence，可以用来调节attribute prior和目标类别的新样本之间的平衡

Method to Transfer

- Knowledge Transfer : Informative Parameter Priors
  - Give parameters of the CT models in the target classifiers informative priors



(b)CT

(c)CT+P

(b): π and ϕ are given symmetric Dirichlet Distribution as prior. The base measure is uniform distribution and be neglected

$$control\ vector = \left(\frac{1}{K}\beta, \frac{1}{K}\beta, \cdots, \frac{1}{K}\beta \right)$$

(c): π and ϕ are given Dirichlet Distribution as prior. μ and η are the base measure

$$control\ vector = \eta_m \beta = (\eta_{m1}\beta, \eta_{m2}\beta, \cdots, \eta_{mK}\beta)$$

Method to Transfer

- Knowledge Transfer : Informative Parameter Priors

Since $E(\phi_k) = \mu_k$ and $E(\pi_m) = \eta_m$ we can set

$$\mu_k = \hat{\phi}_k \text{ and } \eta_m = \tilde{\pi}_m$$

The basic equation of Gibbs sampling of the CT model with informative prior the becomes

$$p(z_{ji} = k | w_{ji} = v, \Omega) \propto \frac{\beta \tilde{\pi}_{mk} + M^{k}_{m,\backslash ji}}{\beta + \sum_{k'=1}^{K} M^{k'}_{m,\backslash ji}} \frac{\lambda \hat{\phi}_{kv} + C^{v}_{k,\backslash ji}}{\lambda + \sum_{v'=1}^{W} C^{v'}_{k,\backslash ji}}$$

Target Classifier with transferred knowledge

Method to Transfer

- Importance of informative priors for zero-shot learning
  - Category-Topic Model

$$p(z_{ji}=k|w_{ji}=v,c_j=m,\Omega) \propto \frac{\beta/K}{\beta} \frac{\lambda/W}{\lambda}$$

□ Have to see some training samples of target category

□ Otherwise, can only give symmetric Dirichlet prior $\eta_{mk} = 1/K$

□ Cannot be used in zero-shot learning

  - Category-Topic Model with informative priors

$$p(z_{ji}=k|w_{ji}=v,\Omega) \propto \frac{\beta\tilde{\pi}_{mk}}{\beta} \frac{\lambda\tilde{\phi}_{kv}}{\lambda}$$

□ give Dirichlet prior $\eta_{mk} = \tilde{\pi}_{mk}$

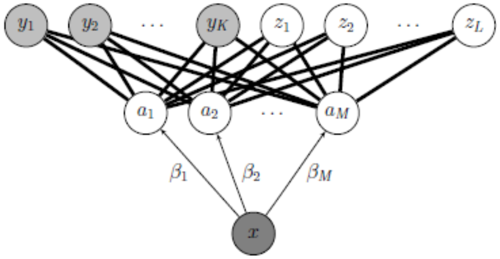
□ Can be used in zeros-shot & one-shot learning

实验

- 数据库和图像特征
  - Animals with Attributes (AwA)
    - □This dataset provides a plattform to benchmark transfer-learning algorithms, in particular *attribute base classification*
    - □30475 images from 50 animal categories
    - □85 attributes per category, 38 non-visual, 47 visual
    - □Category-attribute relationship: 50*85 matrix M
  - 特征
    - □4 types of features: SIFT, rgSIFT，Local color histogram, local self-similarity histogram, about 5000 features per image
    - □Each type of feature: 1000 words by k-mean clustering
    - □Features in images are quantized into one of the codeword

实验

- 实验设置
  - – Tasks
    - □Source categories: 40, Target categories: 10
    - □Zero-shot learning, One-shot Learning
  - – Baseline Algorithm
    - □Direct Attribute Prediction (DAP)
      - ➢ State-of-the-art method for zero-shot learning on AwA
      - ➢ For zero-shot learning : DAP + MAP
      - ➢ For one-shot learning : DAP + 1NN
    - □SVM
      - ➢ SVM + synthesized training samples



(b) Direct attribute prediction (DAP)

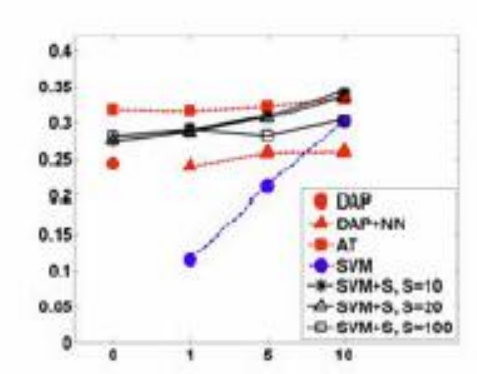
  - – 参数
    - □AT model: K_0=10 unshared topics per attribute
    - □CT model: 100 topics
    - □SVM: C-SVC with χ2 kernel

实验

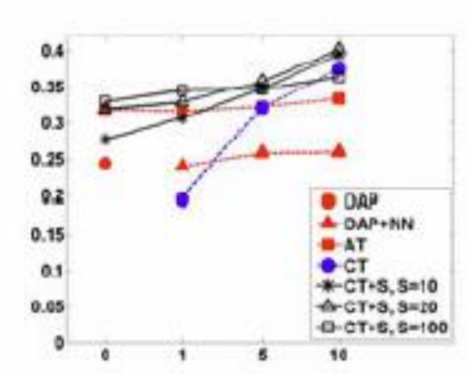
- 实验设置
  - 评测
    - □Zero-shot learning
      - ➢ AT & DAP are trained using the first 100 images of each source category
      - ➢ CT & SVM+S are trained on these synthesized samples
        - » Use AT model to generate S={10,20,100} synthesized examples for each target category
      - ➢ CT with informative prior (CT+P)
        - » ϕ and π learned in AT model as prior in CT+P
    - □One-shot learning
      - ➢ CT & SVM are trained with synthesized samples /informative prior+ first M={1,5,10} images of each target category
      - ➢ AT model is trained with 100 images of each source category + first M images of each target category
      - ➢ DAP+NN: attributes of first M images for NN classifiers
    - □In both Zeros-shot and One-shot tests, all classifiers are tested over the last 100 images of each target category

实验:Test 1

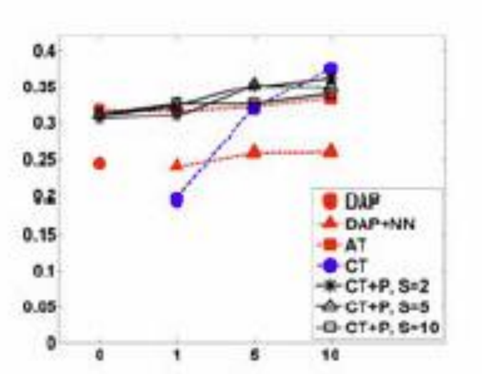
- using all attributes for SVM+S, CT+S, CT+P



(1) SVM+S



(2)CT+S



(3)CT+P

- 结论
  - A better attribute model
  - A better method of knowledge transfer for one-shot
  - A better target classifier
  - Performance of CT+S > CT+P
    - CT+P can be viewed as online version of CT+S
  - More real training samples, less improvement due to prior knowledge

实验

- Illustrations of attribute model

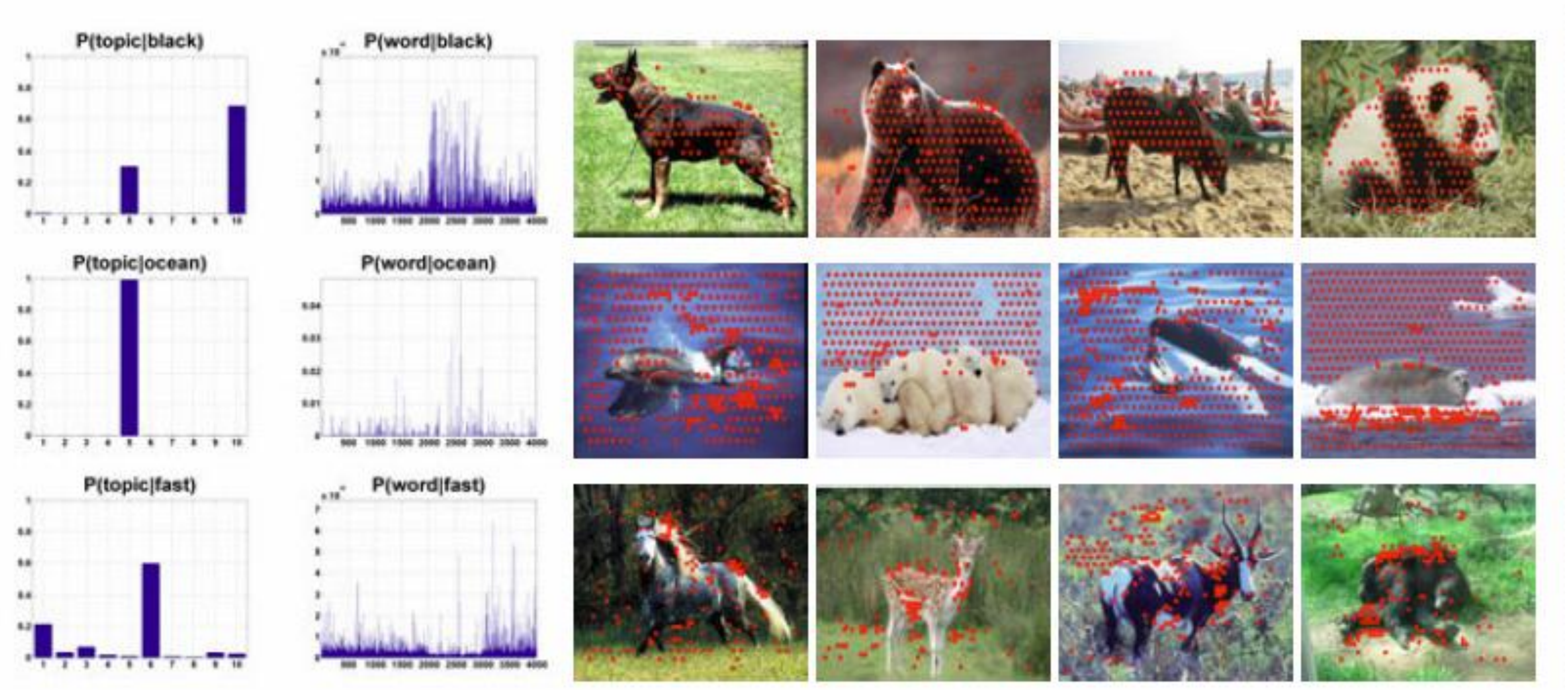
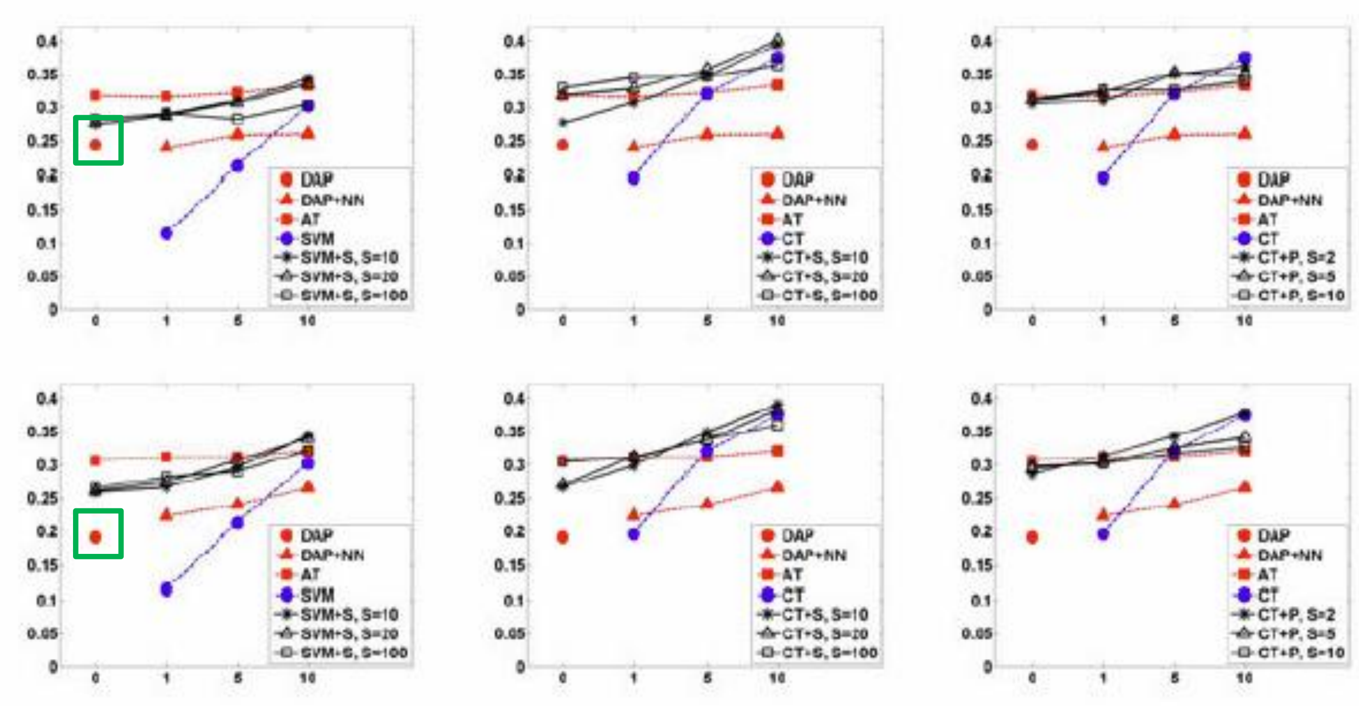


Fig. 5. Illustrations of three attribute models for *black*, *ocean* and *fast* from the top to the bottom. Column 1: the distribution of the 10 topics assigned to a particular attribute; Column 2: the distribution of codewords for a particular attribute; Column 3-6: examples of images from source categories (Column 3-5) and target categories (Column 6), superposed with the top 100 most likely codewords (solid red dots) for the attributes of the same row. Figures are best viewed in color.

实验:Test 2

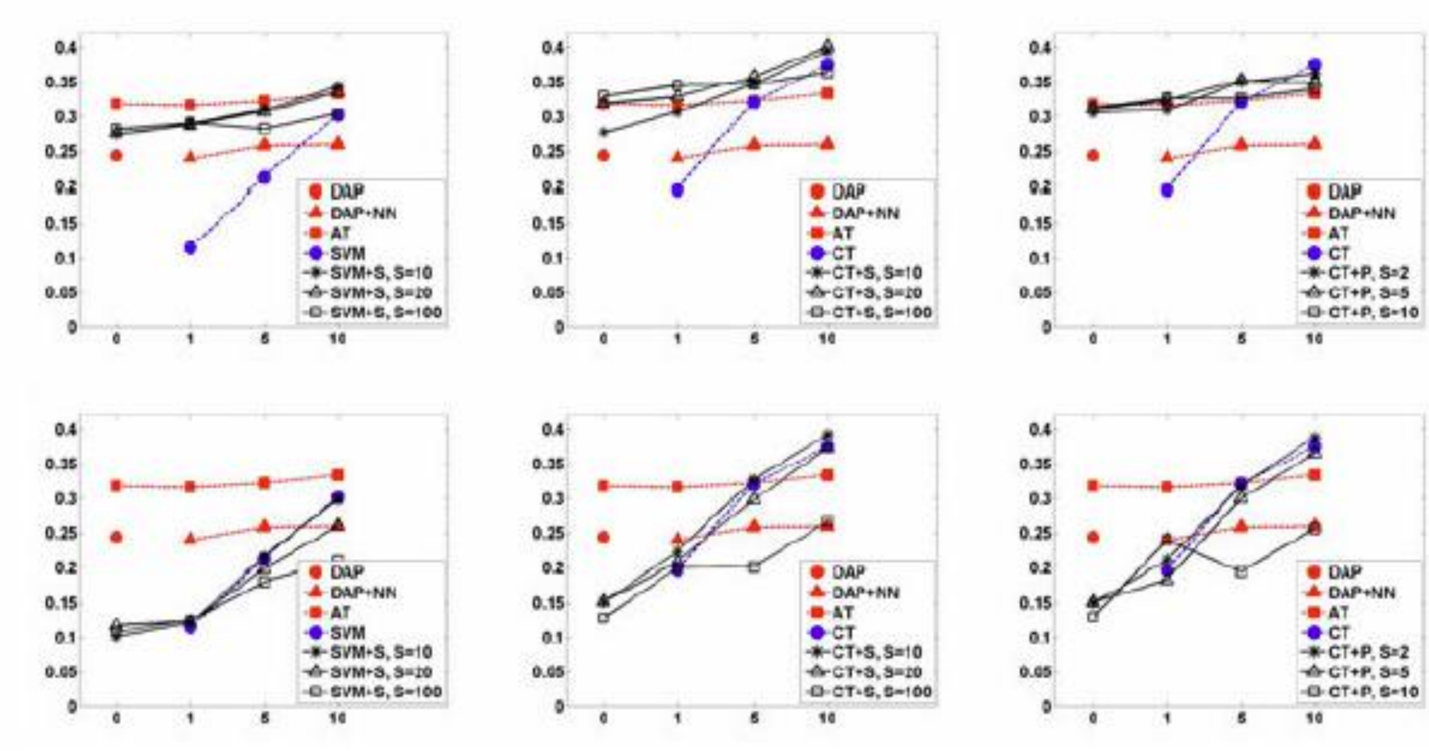
- Non-visual Attributes is important in the transfer learning



Top: with all attributes
Bottom: with visual attributes only

实验:Test 3

- Effectiveness of the Knowledge of attribute in the transfer learning



Top: with ground truth attributes

Bottom: with attributes selected randomly for each target category

Conclusion

- 提出了一种基于物体属性的迁移学习框架
  - 使用产生式模型来描述属性相关的特征的分布
  - Category-Topic model作为Target Classifier
  - 两种迁移attribute prior的方法
- Personal View
  - 用object attribute作为迁移知识的载体
  - 将object attribute的概率分布作为先验加入到target classifier
- Future work
  - More evaluations
  - Compare with methods not using attribute
  - Employ spatial constraints
  - Select informative attributes for particular category

Thanks for Attention

附录

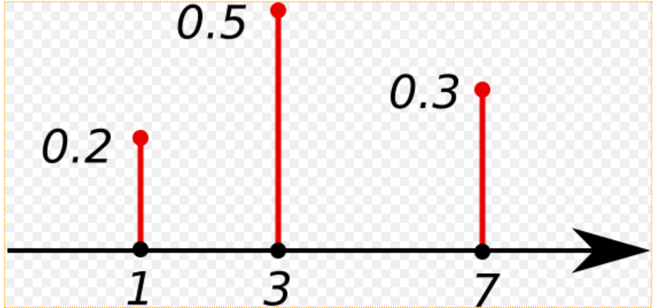
- Discrete Distribution
- Dirichlet Distribution
- Gibbs Sampling

Discrete probability distribution

- **Discrete probability distribution**
  - A probability distribution is called *discrete* if its cumulative distribution function only increases in jumps.
  - More precisely, a probability distribution is discrete if there is a finite or countable set whose probability is 1.
  - Discrete distributions are characterized by a probability mass function *p* such that

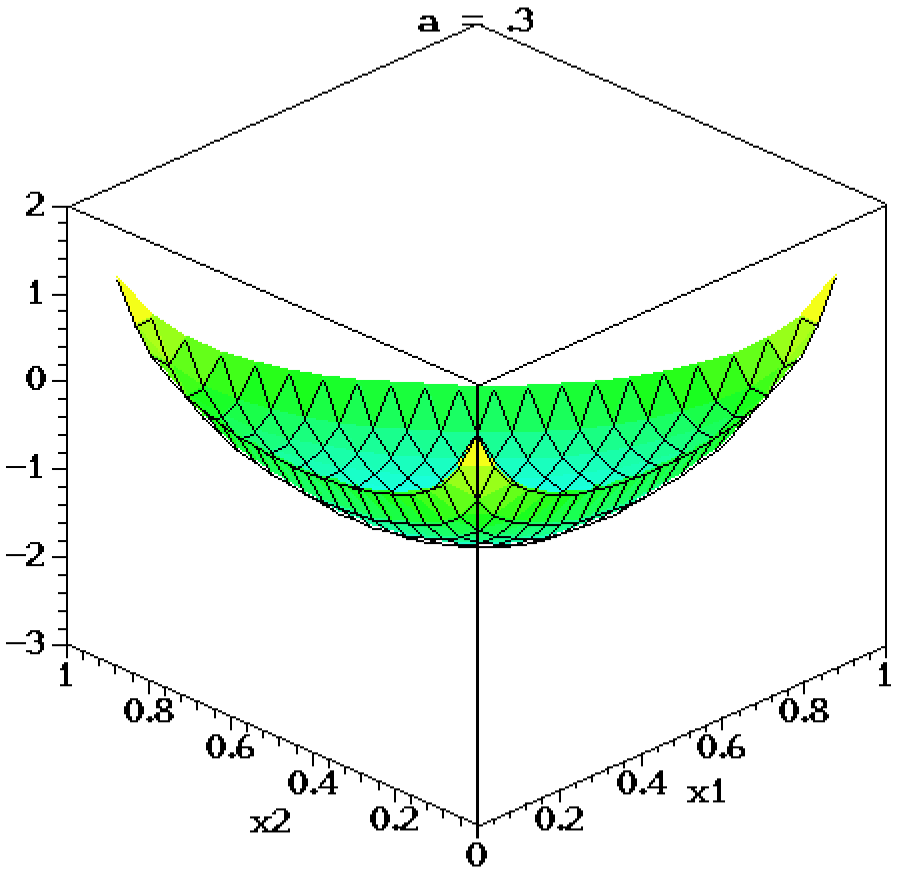
$$\Pr\left[X = x\right] = p(x)$$

$$\sum_u \Pr(X = u) = 1$$



Dirichlet Distribution

- Dirichlet Distribution $f(x_1,\dots,x_K;\alpha_1,\dots,\alpha_K) = \frac{1}{\mathrm{B}(\alpha)}\prod_{i=1}^{K} x_i^{\alpha_i - 1}$
  - often used as prior distributions in Bayesian statistics
  - returns the belief that the probabilities of *K* rival events are x_i given that each event has been observed $\alpha_i - 1$ times.
  - **symmetric Dirichlet distribution**
    - □all of the elements making up the vector have the same value.
    - □often used when there typically is no prior knowledge favoring one component over another.
    - □can be parametrized by a single scalar value α, called the concentration parameter.

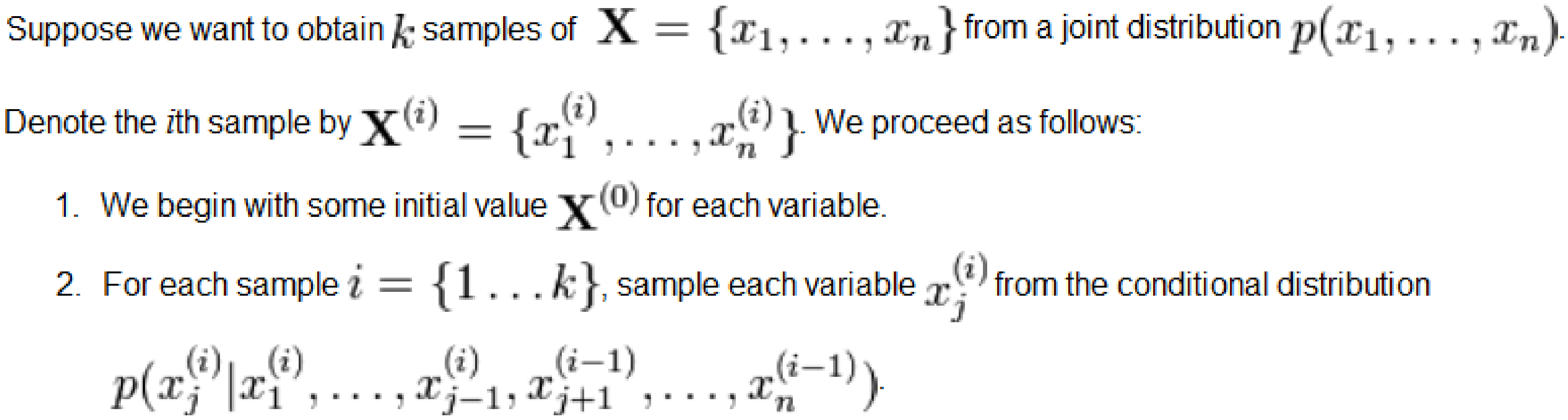

a = .3
2
1
0
−1
−2
−3
1
0.8
0.6
0.4
x2
0.2
0
0.2
0.4
x1
0.6
0.8
1

Gibbs Sampling

- Gibbs Sampling
  - an algorithm to generate a sequence of samples from the joint probability distribution of two or more random variables.
  - The purpose of such a sequence is to approximate the joint distribution; to approximate the marginal distribution of one of the variables…..
  - applicable when the joint distribution is not known explicitly or is difficult to sample from directly, but the conditional distribution of each variable is known and is easy (or at least, easier) to sample from.

Gibbs Sampling

- Gibbs sampling is particularly well-adapted to sampling the posterior distribution of a Bayesian network
- Sampling process



Suppose we want to obtain k samples of $\mathbf{X} = \{x_1, \ldots, x_n\}$ from a joint distribution $p(x_1, \ldots, x_n)$.

Denote the ith sample by $\mathbf{X}^{(i)} = \{x_1^{(i)}, \ldots, x_n^{(i)}\}$. We proceed as follows:

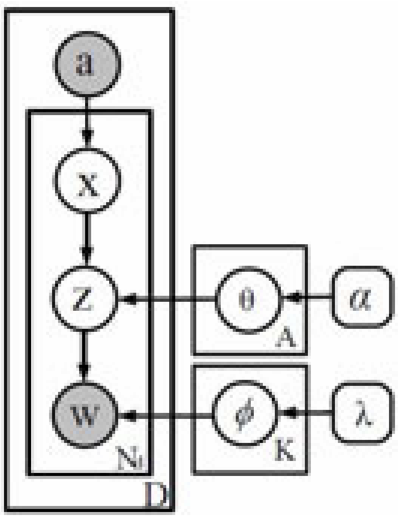
1. We begin with some initial value $\mathbf{X}^{(0)}$ for each variable.
2. For each sample $i = \{1 \ldots k\}$, sample each variable $x_j^{(i)}$ from the conditional distribution

$$p(x_j^{(i)} | x_1^{(i)}, \ldots, x_{j-1}^{(i)}, x_{j+1}^{(i-1)}, \ldots, x_n^{(i-1)}).$$

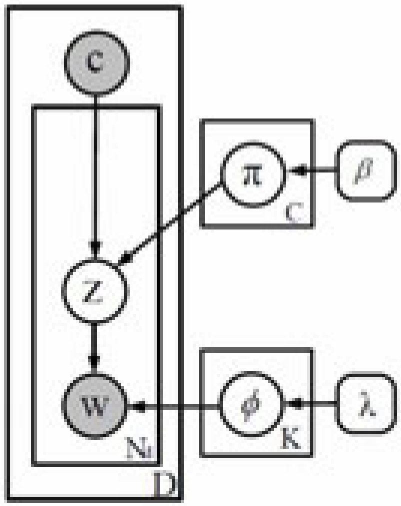
The samples can approximate the joint distribution of all variables

Attribute-Based Transfer Learning

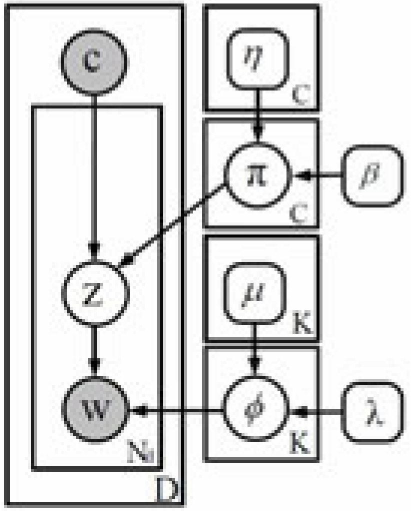
- Framework



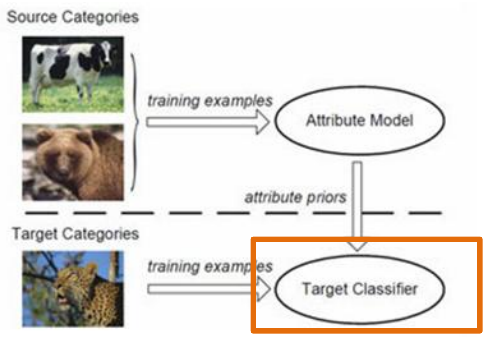
(a)



(b)



(c)



(a) Attribute Model: Author-Topic Model
(b) Target Classifier: Category-Topic Model
(c) Target Classifier with transferred knowledge