Topic models
Generative models

Discriminative methods model the posterior:

$$\frac{P(\text{zebra} \mid \text{image})}{P(\text{non zebra} \mid \text{image})}$$

Generative methods model the likelihood and priors

$$\frac{P(\text{image} \mid \text{zebra})}{P(\text{image} \mid \text{non zebra})} \quad \frac{P(\text{zebra})}{P(\text{non zebra})}$$

From Fei Fei Li
Most known generative models: pLSA, LDA..

- Generative hierarchical models introduce latent variables for general co-occurrence data that represent hidden topics in the document, that must be learned.

- According to the latent variable model, each word-document observation (w,d) is associated with a class variable \( z \in \{z_1,...,z_k\} \) that represents the hidden topic. Most known are pLSA, LDA and their developments.
Probabilistic Latent Semantic Analysis topic model

- Probabilistic Latent Semantic Analysis (pLSA) generative topic model assumes that in a document \( d \), word distributions \( P(w,d) \) are combinations of the factors \( P(w|z) \) (word distribution per topic) and the mixing weights \( P(z|d) \) (topic distribution per document), i.e. the probability that a word \( w \) occurs in a document \( d \) is explained by topic \( z \). The pairs \( (d,w) \) are assumed to be independent.

- A model can be obtained according to the following steps:
  - Select a document \( d \) with probability \( P(d) \)
  - Pick a latent topic class \( z \) with probability \( P(z|d) \)
  - Generate a word \( w \) with probability \( P(w|z) \)

\[
P(w,d,z) = P(w|z)P(z|d)P(d)
\]

\[
P(w,d) = \sum_z P(w,d,z) = P(d)\sum_z P(w|z)P(z|d)
\]

\[
P(w|d) = \sum_z P(w|z)P(z|d)
\]

or, considering the indexes explicitly

\[
P(w_i|d_j) = \sum_{k=1}^{K} P(w_i|z_k)P(z_k|d_j)
\]
The pLSA model with images

D: the number of images in the image collection
n: the number of patches per image
W: the number of visual word in the vocabulary
K: the number of hidden topics

d_j: the j\textsuperscript{th} image in an image collection D
w_i: the visual word of the i\textsuperscript{th} patch
z_k: the k\textsuperscript{th} latent topic of the patch

\[ P(w_i|d_j) = P(d_j) \sum_{k=1}^{K} P(w_i|z_k)P(z_k|d_j) \]
pLSA can be thought of as a matrix decomposition:

\[
P(w_i | d_j) = \sum_{k=1}^{K} P(w_i | z_k) P(z_k | d_j)
\]

- If we assume that in an image the topics correspond to objects, an image will be expressed as a mixture of different objects and backgrounds.

- The goal is to find topic vectors common to all image documents \( d \) (term-document matrix rows) and mixture coefficients specific to each document \( d \) i.e. topic distributions per document (term-document matrix \( z-d \) columns). The PLSA parameters are therefore the two conditional distributions \( P(z|d) \) and \( P(w|z) \). These are all probabilities which sum to one.
EM for pLSA (training on a corpus)

- Following the likelihood principle the fitting model parameters $P(w|z)$ and $P(z|d)$ can be obtained by maximizing the likelihood function:

$$L = \log P(D, W) = \prod_{d} \prod_{w} P(w, d)^{n(d, w)} = \sum_{d} \sum_{w} n(d, w) \log P(w, d)$$

- This is difficult to solve because of the summation in the expression of the log. Expectation Maximization (EM) is used for this purpose.

**E Step**

$$P(z_k \mid d_j w_i) = \frac{P(z_k)P(d_j \mid z_k)P(w_k \mid z_k)}{\sum_{z'} P(z'_k)P(d_j \mid z'_k)P(w_k \mid z'_k)}$$

**M Step**

$$P(w_i \mid z_k) = \sum_{d=1}^{D} n(d_j, w_i)P(z_k \mid d_j w_i)$$

$$P(d_j \mid z_k) = \sum_{i=1}^{W} n(d_j, w_i)P(z_k \mid d_j w_i)$$

$$P(z_k) = \sum_{j=1}^{D} \sum_{i=1}^{W} n(d_j, w_i)P(z_k \mid d_j w_i)$$
Init
Provision random initialization of the unknowns “?”. $P(d)$ and $P(z)$ uniform.

**E Step**

$$P(z_k | d, w_i) = \frac{P(z_k)P(d_j | z_k)P(w_k | z_k)}{\sum_{z_k'} P(z_k')P(d_j | z_k')P(w_k | z_k')}$$

2x3x2 = 12 new values that are used in the M-step

**M Step**

$$P(w_i | z_k) = \sum_{j=1}^{D} n(d_j, w_i)P(z_k | d_j, w_i)$$

4 new values to the next E-step

$$P(d_j | z_k) = \sum_{i=1}^{W} n(d_j, w_i)P(z_k | d_j, w_i)$$

6 new values to the next E-step

$$P(z_k) = \sum_{j=1}^{D} \sum_{i=1}^{W} n(d_j, w_i)P(z_k | d_j, w_i)$$

2 new values to the next E-step
Building the term-document training matrix

- Given an image collection D, from images d and visual words associated to image patches w it is possible to build the term-document matrix (documents, words).

Stack visual word histograms as columns in matrix

<table>
<thead>
<tr>
<th>i</th>
<th>j</th>
<th>Number of times word i appears on image j</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

From B. Russell & J. Sivic
Two-stages training:

- **Stage 1:** The topic specific distributions $P(w|z)$ are learnt from the set of training images. Determining both $P(w|z)$ and $P(z|d_{\text{train}})$ involves fitting the pLSA model to the entire set of training images. It is not necessary to provide the category of each image or any region segmentation. EM is used to maximize the likelihood (MLE). Each training image $d_{\text{train}}$ is then represented by a vector $P(z|d_{\text{train}})$ of dimension $Z$ where $Z$ is the number of topics learnt.

- **Stage 2:** as $P(z|d)$ is obtained for each training image, then the image is classified as containing object (topic) $k$ according to the max of $P(z_k|d_j)$ over $k$.
Learnt parameters

Image collection $D$

$P(w_i \mid d_j) = \sum_{k=1}^{K} P(w_i \mid z_k) P(z_k \mid d_j)$

Codeword distributions per theme (topic)

Topic distributions per image
Recognition using pLSA

- After pLSA, an image is a mixture of learned topic vectors and mixing weights. Recognition is performed by using the learned mixing coefficients to classify an image. In particular:
  - in order to find the most likely topic (class) for an image:
    \[ z^* = \arg \max_z P(z \mid d) \]
  - in order to find the most likely topic (class) for a visual word in a given image:
    \[ z^* = \arg \max_z P(z \mid w, d) = \arg \max_z \frac{P(w \mid z)P(z \mid d)}{\sum_{z'} P(w \mid z')P(z' \mid d)} \]
Recognition examples

From Bryan Russell & Josef Sivic
Words shown are not the most probable words for a topic, but instead they are words that have a high probability of occurring in a topic AND high probability of occurring in the image. Find words with $P(z_k|d_j, w_i) > 0.8$ is a good heuristics

$$z^* = \arg \max_z P(z|w, d) = \arg \max_z \frac{P(w|z)P(z|d)}{\sum_{z'} P(w|z')P(z'|d)}$$
Image as a mixture of topics (objects)

From Sivic, Russell, Efros, Zisserman, Freeman
pLSA-based image topics discovery

The general schema for image topics discovery using pLSA

LEARNING STAGE

CLASSIFICATION STAGE

P(w|z) is the same as in the training stage

Mixing coefficients learnt: the two unknown parameters are computed by the Expectation-Maximization algorithm on the set of training documents

K-most similar images after K-Nearest Neighbour clustering on the training set

From Bosch, Zisserman, Munoz
Classification of unseen images

• Classification of unseen images proceeds in two stages as well:

  – **Stage 1:** the document specific mixing coefficients $P(z_k | d_{test})$ are sought such that the Kullback-Leibler divergence is minimized between the measured distribution $P(w|d_{train})$ in the training stage and

    $$ P(w | d_{test}) = \sum_z P(w | z) p(z | d_{test}) $$

    EM is run as in the training stage but only coefficients $P(z_k | d_{test})$ are updated in each M step keeping fixed the learned value of $P(w | z)$. In the end the test image is represented by a Z-vector

  – **Stage 2:** the test image is then classified using a K-Nearest Neighbours classifier on the Z-vectors of the training images using an Euclidean distance function.

    The KNN selects the K nearest neighbours of the new image within the training database. Then it assigns to the new picture the label of the category which is most represented within the K nearest neighbours.
Performance under variation in various parameters for the 8 category OT classification. Top: example visual words and performance for dense colour SIFT $M = 10$, $r = 4, 8, 12$ and $16$ (each column shows the HSV components of the same word). Lower example visual words and performance for grey patches with $N = 5$ and $M = 3$. (a) Varying number of visual words, $V$, (b) Varying number of topics, $Z$, (c) Varying number $k$ (KNN).
Advantages and problems of pLSA

• Pros:
  – With pLSA images are not related to a single cluster, i.e. topic, as with clustering. $P(z|d)$ defines a specific mixture of factors and this offers more flexibility, and produces more effective modeling
  – pLSA generates a topic model and maximizes its predictive power

• Cons:
  – pLSA is prone to overfitting: in the presence of many parameters it is likely that noise is estimated. Number of parameters to be estimated grows with size of training set.
  – pLSA is not a well-defined generative model of unseen documents: there is no natural way to assign probability to a previously unseen document

Latent Dirichlet Allocation (LDA) overcomes some of the problems
Latent Dirichlet Allocation (LDA)

- Latent Dirichlet Allocation (LDA) treats the topic mixture weights as a \textit{k-parameter hidden random variable} and places a \textit{Dirichlet prior} on the multinomial mixing weights. $\alpha$ and $\beta$ must be estimated before we can find the topic mixing proportions belonging to a previously unseen document.

For each of the N patches
- Choose theme $Z_n \sim \text{mult}(\pi)$
- Choose patch $X_n \sim p(X_n|Z_n, \beta)$
  $\beta$ is matrix of size $K \times T$ (# themes x # words)
LDA topic model

- Category label
- Multinomial distribution over themes
- Visual words

LDA topic model diagram from Fei-Fei Li