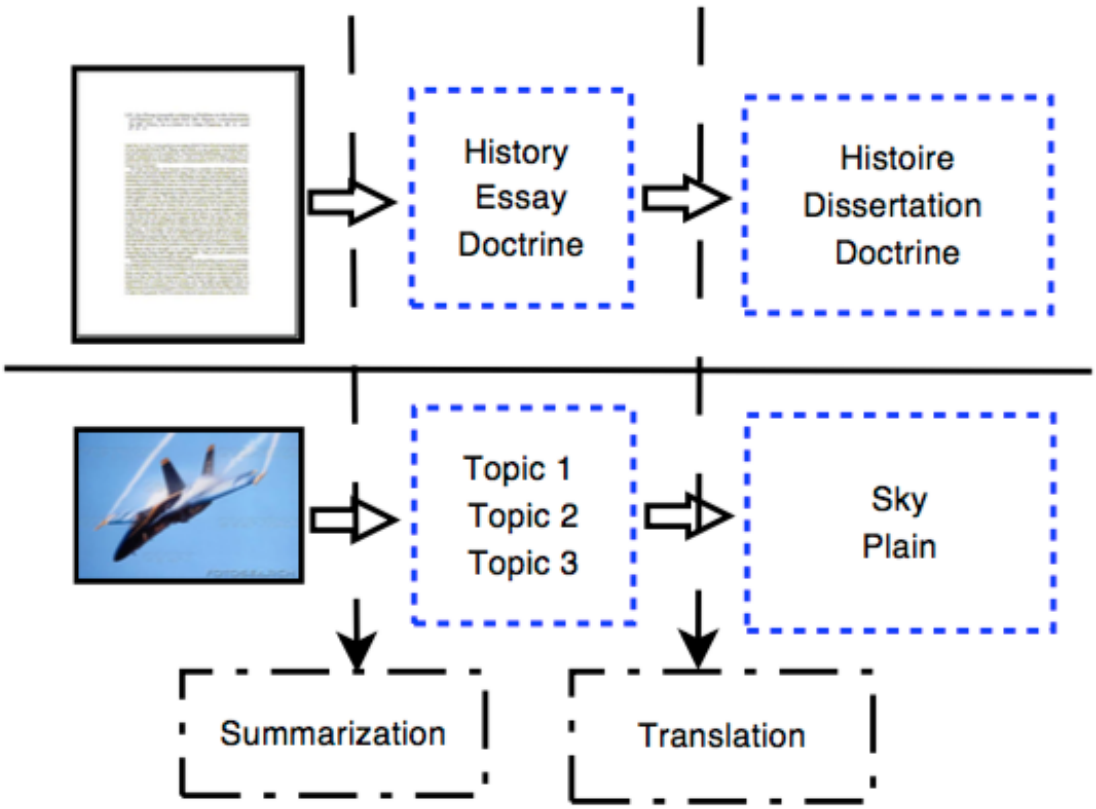
Translating Topics to Words for Image Annotation

`Image annotation/tagging`

Novel ideas

1. **view the annotation processes as two consecutive processes, document summarization and translation**
2. **Compared to the original translation model, our visual topics learned by the probabilistic latent semantic analysis (PLAS) approach, `provide an intermediate abstract level of visual description.`**



Pipeline

1. **PLSA:** learn topics from text documents, Let's see how the PLSA works
    - a set of text D = $\{d_1, d_2, d_3, \ldots, d_n\}$
    - each of which above is represented by a frequency vector

$d_i = [n(d_i, w_1), n(d_i, w_2), \ldots, n(d_i, w_m)]$

- assumption: each word in a document is generated by a specific hidden topic z_k, where $z_k \in \mathcal{Z}$
- $$P(w_j|d_i) = \sum_k^K P(w_j|z_k, d_i)P(z_k|d_i)$$

  where K is the number of hidden topics.
- Since the conditional probability of generating a word by a specific topic is independent from the document, so

$$P(w_j|d_i) = \sum_k^K P(w_j|z_k)P(z_k|d_i)$$

, we can use EM algorithm to compute $P(w_j|z_k), P(z_k|d_i)$

2. **Learning Visual Topics**
   1. **each image is partitioned into a number of small patches using a regular grid**
   2. `From each patch, extract a 128D SIFT and 6D Color descriptor separately.` **The Color descriptor is a concatenation of the mean and the variance values of the R,G,B channels in a patch.**, use clustering to get the subset of SIFT and Color descriptors.
   3. `How to transform the image to text document?` The whole visual vocabulary is obtained by the combination of texture words and color words, we can transform an image into a text document by **assigning a visual word label to each image patch**
   4. Then we can use the PLSA to learn a number of visual topics. After a PLSA model is learned from the training images, we can obtain the topic labeling o of a visual word v in a specific document d.

$$P(o|v, d) = \frac{P(v|o)P(o|d)}{P(v|d)}$$

   5. Finally we make each image patch have a topic label

3. **Machine Translation Model**
   - We have finished the summary of visual topic. Next we train a language translation model from the training data **which can map the visual topics to textual models**
   - We use **IBM model** to translate. The feature of the image J_i is

$$J_i = \{\bar{O}_i; \bar{W}_i\} = \{b_{i,1}, b_{i,2}, b_{i,3}, \ldots, b_{i,m}; a_{i,1}, a_{i,2}, \ldots, a_{i,n}\}$$

where n means the number of visual topics, m means the size of textual vocabulary.

- **How to translate the visual topic to the textual word?** $t_{j,k}$ is the probability of translating the k^{th} visual topic to the j^{th} textual word.
- We employ the EM algorithm to train the image to get the $L(\mathcal{J})$, we can use the EM to find the optimal translation probability table $\Phi^* = \{t_{i,j}\}^*$
- Finally the probability of annotating a test image $\mathcal{J}$ with a word w_j given $\bar{O}$ is

$$P(w_j|\bar{O}; \Phi^*) \propto \sum_{k=1}^{m} (t^*_{j,k} b_k)$$

Experiment Details

1. use a regular grid with 13X13 pixels for each patch
2. use k-means clustering to form vocabularies of texture words and color words and size of vocabularies is 500
3. After obtaining the parameters of PLSA, we estimate the topic representation of each test image.
4. The numbers of learned texture topic and colour topic are both chosen as 50
5. The annotation of test image is achieved by the translation from the visual topics to annotation words. The top five words are selected as the final annotation

文章核心思想解读

$$P(topic|w,d) = \frac{P(w|topic)P(topic|d)}{P(w|d)}$$

$$Due\ to \quad P(topic|d) \propto P(d|topic)P(topic)$$

$$Hence \quad P(topic|w,d) \propto P(w|topic)P(d|topic)P(topic)$$

$topic \in Topic$

$$P(topic|w,d) = \sum_{topic \in Topic} P(w|topic)P(d|topic)P(topic)$$

求上式的MLE，然后取log得到

$$L = \prod_{i=1}^{N} \prod_{j=1}^{M} P(topic|w_j, d_i)^{n(w_j, d_i)}$$

$$logL = \sum_{i=1}^{N} \sum_{j=1}^{M} n(w_j, d_i) logP(topic|w_j, d_i)$$

EM to find optimal solution

其中对应到Image Annotation这个任务上，相应的变更为

1. word变成visual word，通过一个slide window，来取image patch，然后将相似的patch进行一个聚类，相似性度量是从sift和color两方面来度量的，这样来生成相应的visual-vocabulary list
2. document 就是相应的image。