Optimal Tag Sets for Automatic Image Annotation

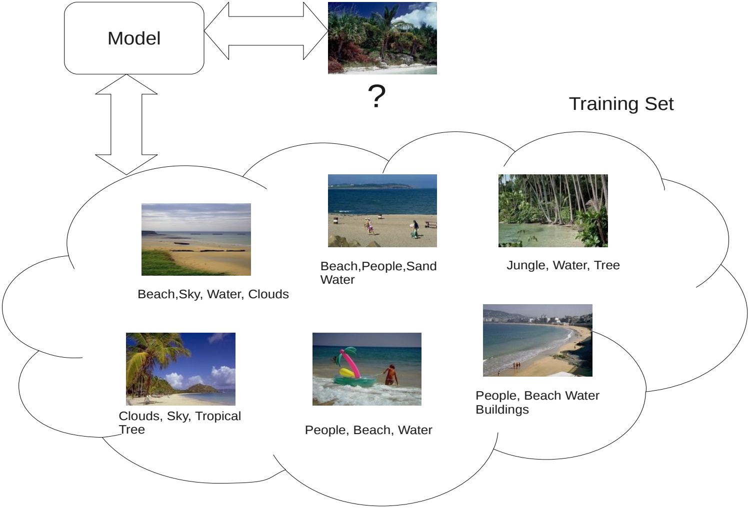
Sean Moran, Victor Lavrenko

University of Edinburgh, UK

BMVC 2011

Overview/Motivation

Image Annotation: Given an un-annotated test image and a training set of annotated images select tags that reflect the content of the test image.



Overview/Motivation

- Popular field of research:
    - Annotation as machine translation [Duygulu et al. '02]
    - Continuous Relevance Model (CRM) [Lavrenko et al. '03]
    - Label diffusion over a similarity graph [Liu et al. '09]
    - Supervised multiclass labeling [Carneiro et al. '07]
- Limiting assumptions across a broad class of models:
    - Gaussian kernel: Standard workhorse of many models. But is it necessarily the most accurate default kernel?
    - Tags independent: Leads to incohesive and contradictory tags e.g. {tropical, blizzard, supernova}.
- BS-CRM: principled framework for solving both limitations in a *generative model* of image annotation.

Outline

- Continuous Relevance Model (CRM)
- Capturing Feature Covariance with Minkowski Kernels
- Capturing Keyword Correlation through Beam Search
- Experiments
- Discussion

Continuous Relevance Model [Lavrenko et al. 2003]

- Statistical generative model for automatic image annotation.
- Estimates joint distribution of visterms and tags [De Finetti'31]:

$$P(\mathbf{w}, \mathbf{f}) = \sum_{J \in T} P(J) \prod_{j=1}^{n} P(w_j | J) \prod_{i=1}^{m} P(\vec{f_i} | J)$$

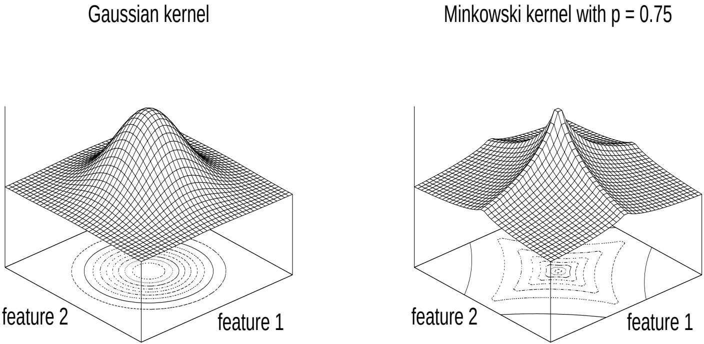
  - $P(J)$: Uniform prior
  - $P(\vec{f_i}|J)$: Gaussian non-parametric kernel density estimate
  - $P(w_i|J)$: Dirichlet prior for word smoothing

- Estimate marginal probability distribution over individual tags:

$$P(w|\mathbf{f}) = \frac{P(w, \mathbf{f})}{\sum_w P(w, \mathbf{f})}$$

- Top e.g. 5 words used as annotation of image.

Capturing Feature Covariance with Minkowski Kernels



$$P(\vec{f}_i|J) = \frac{1}{n}\sum_{j=1}^{n} c_{\mathrm{p}}\, exp\left\{\frac{-|\vec{f}_i - \vec{f}_j|^{\mathrm{p}}}{\beta}\right\}$$

Capturing Feature Covariance with Minkowski Kernels

Sensing subtle changes

Minkowski kernel much more sensitive to subtle feature changes. Known to be an important facet of human vision [Howarth'05].

Conjunction of features

Minkowski kernel mimicks logical AND of variations in feature values whilst Gaussian kernel is closer to a logical OR.

Capturing Keyword Correlation with Beam Search

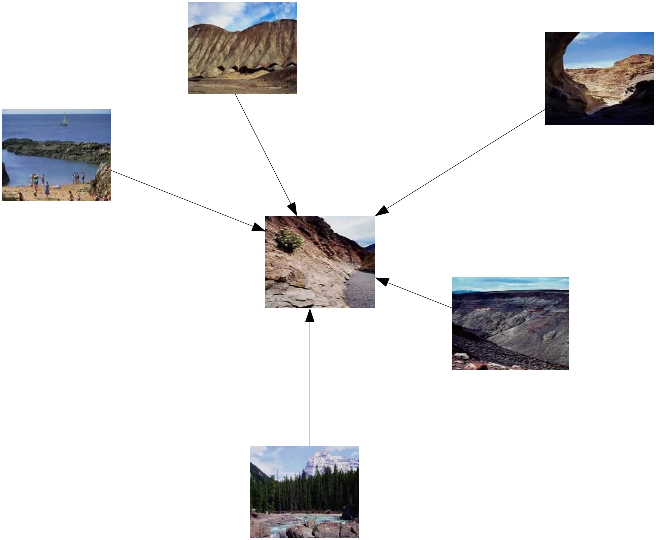
- CRM computes a set to set mapping of tags to visterms $P(\mathbf{w}, \mathbf{f})$
- Add measure to penalize frequent words $I(\mathbf{w})$.

$$I(\mathbf{w}) = P(\mathbf{w}|\mathbf{f}) \cdot \log \frac{P(\mathbf{w}|\mathbf{f})}{P_0(\mathbf{w})}$$

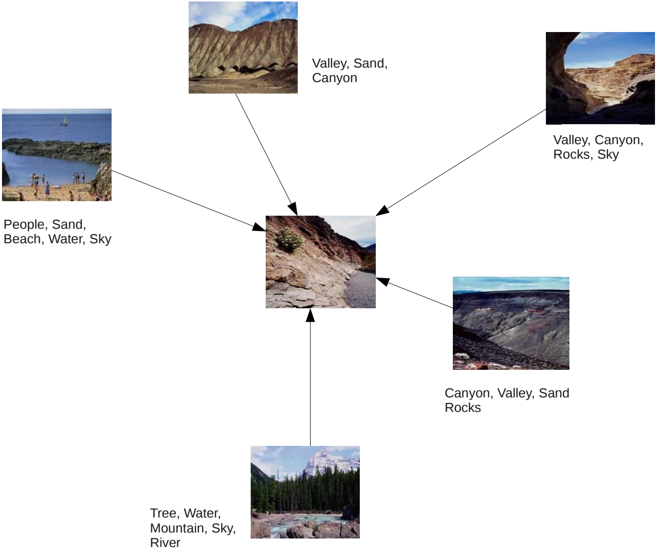
  - $P(\mathbf{w}|\mathbf{f})$: Dependence model between a tag set and image features.
  - $P_0(\mathbf{w})$: Background model that treats every tag as an isolated event.
- Goal: Find optimal tag set maximizing $S_k^* = argmax_{S_k \subset V} I(S_k)$
- Optimisation over universe of all possible tag sets: use efficient approximation via *Beam Search*.

Capturing Keyword Correlation with Beam Search

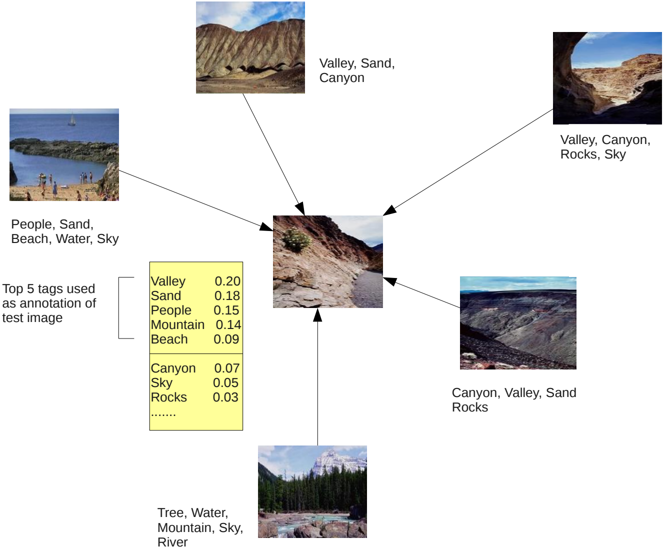
Capturing Keyword Correlation with Beam Search



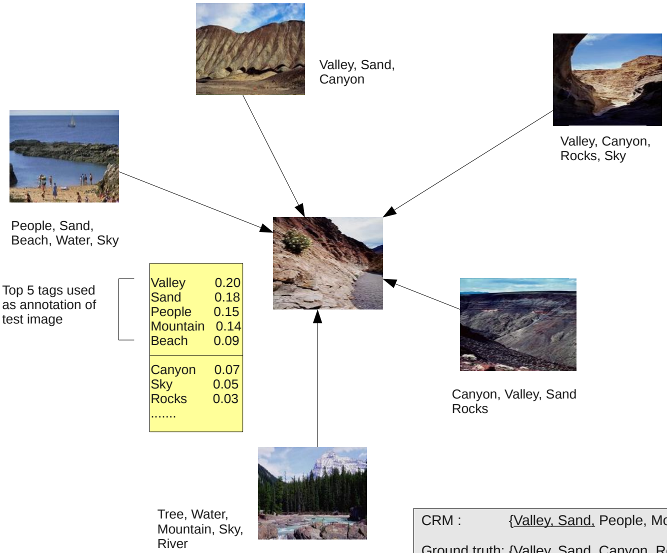
Capturing Keyword Correlation with Beam Search



Capturing Keyword Correlation with Beam Search

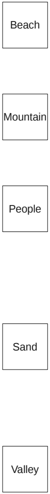


Capturing Keyword Correlation with Beam Search



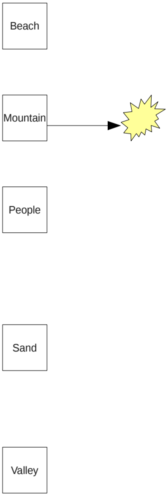
CRM :	{Valley, Sand, People, Mountain, Beach}
Ground truth:	{Valley, Sand, Canyon, Rocks }

Capturing Keyword Correlation with Beam Search

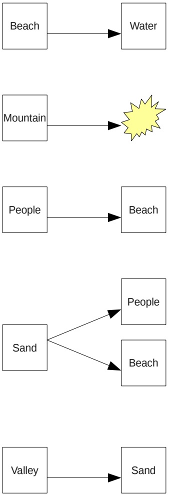


Initialise beam with top B=5 words from CRM

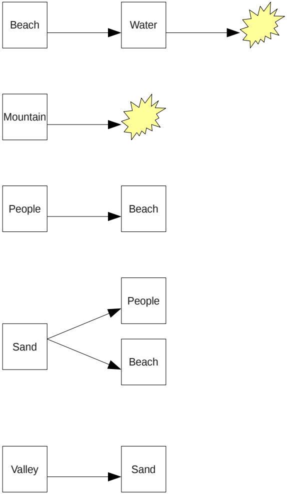
Capturing Keyword Correlation with Beam Search



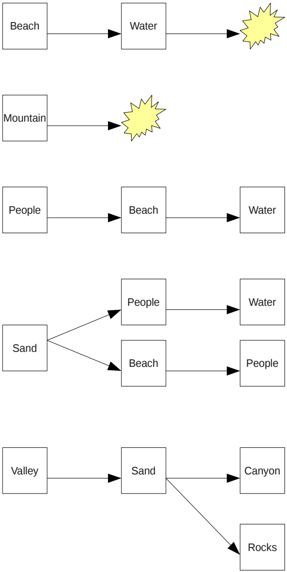
Capturing Keyword Correlation with Beam Search



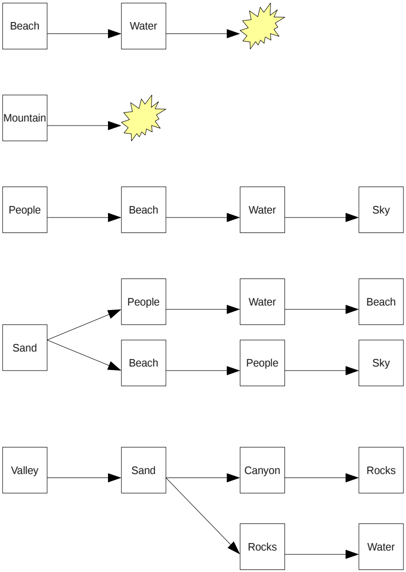
Capturing Keyword Correlation with Beam Search



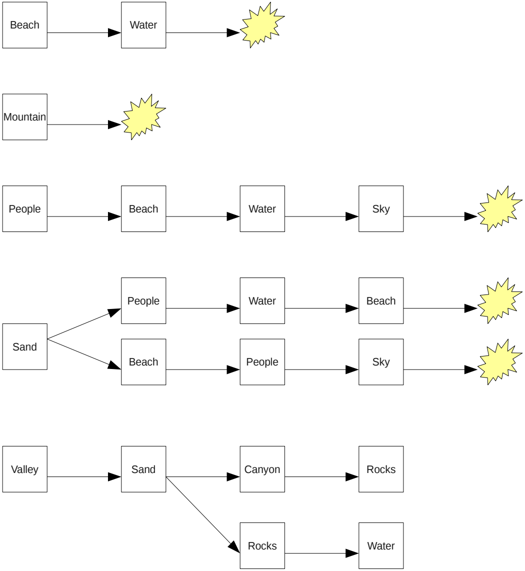
Capturing Keyword Correlation with Beam Search



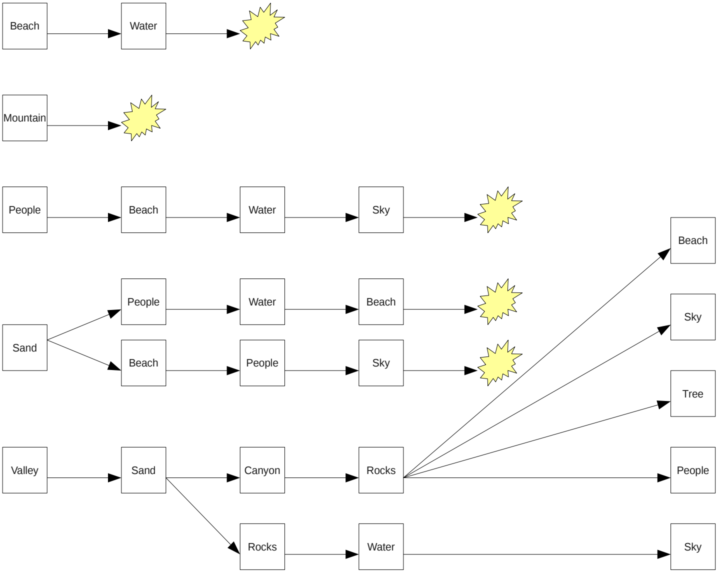
Capturing Keyword Correlation with Beam Search

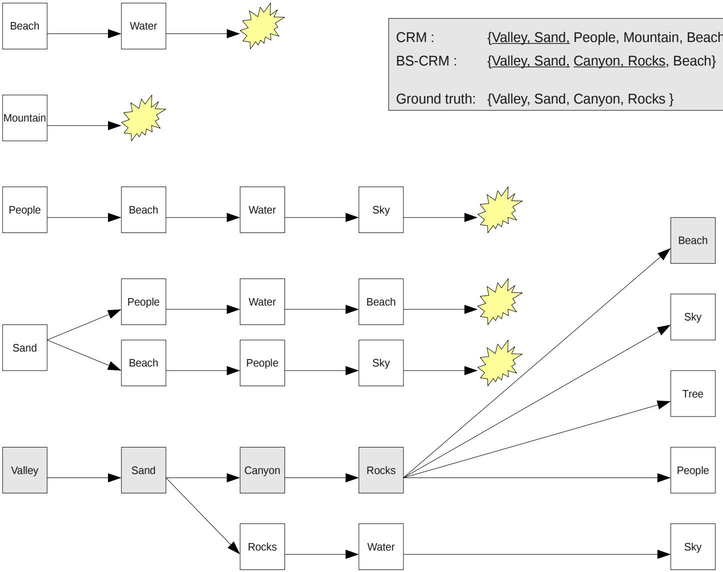


Capturing Keyword Correlation with Beam Search



Capturing Keyword Correlation with Beam Search





CRM :	{Valley, Sand, People, Mountain, Beach}
BS-CRM :	{Valley, Sand, Canyon, Rocks, Beach}
Ground truth:	{Valley, Sand, Canyon, Rocks }

Setup of Experiments - Data

- Corel 5K:
  - 5000 images: landscape, animals, cities
  - Vocabulary of 260 words

Setup of Experiments - Data

- IAPR TC-12:
  - 20,000 images: touristic photos, sports
  - Vocabulary of 291 words
  - Annotations extracted from descriptive text (nouns)

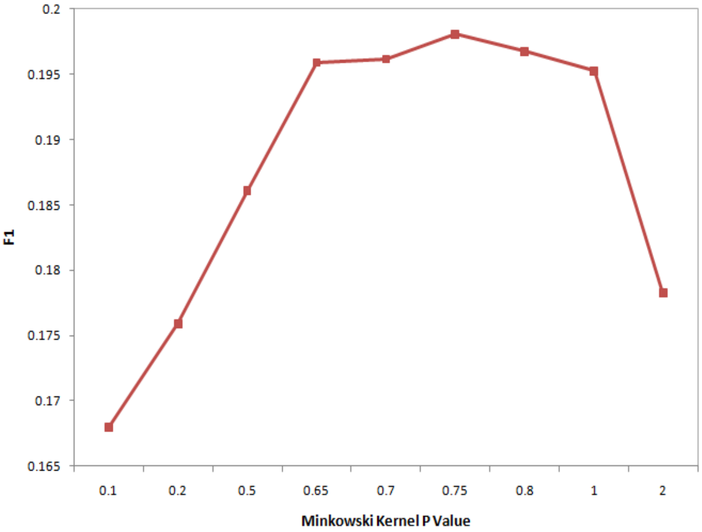
Setup of Experiments - Data

- University of Washington (UW):
  - 1109 images: natural scenes, sports
  - Vocabulary of 158 words
  - Manually removed function words and morphological variants

Setup of Experiments - Data

- Colour and texture based features:
  - Region colour average, standard deviation, skewness
  - Gabor mean orientated energy in 30 degree increments
- Model parameters optimized on a held out validation set:
  1. Grid search over the β and μ for standard CRM model.
  2. Hold β and μ constant: optimize B for varying beam widths.
  3. Hold β, μ and B constant: optimize p for Minkowski density.
- Evaluation metrics (for fixed annotation length):
  - Mean per word Recall
  - Mean per word Precision
  - F1 Measure
  - Number of words with Recall > 0

Setup of Experiments - Optimisation



Setup of Experiments - Evaluation

Model	R	P	**F1**	N^+
COREL				
CRM (p=2)	19	16	**17**	106
Zhou et al.	20	19	**19**	...
Liu et al.	24	19	**21**	125
CRM (p=0.75)	25	21	**23**	119
Wang et al.	23	23	**23**	123
BS-CRM (p=0.75)	27	22	**24**	130

- **Corel 5k:**
    - CRM (p=2) vs CRM (p=0.75): 35% increase in F1
    - T-test: $p \leq 0.00004$
    - CRM (p=2) vs BS-CRM (p=0.75): 41% increase in F1
    - T-test: $p \leq 0.00001$

Setup of Experiments - Evaluation

Model	R	P	**F1**	N^+
UW				
CRM (p=0.70)	36	36	**36**	86
BS-CRM (p=0.70)	46	42	**44**	106
IAPR TC-12				
CRM (p=0.70)	15	23	**19**	202
BS-CRM (p=0.70)	22	24	**23**	250

- **UW:**
  - CRM (p=0.70) vs BS-CRM (p=0.70): 22% increase in F1
  - T-test: $p \leq 0.001$
- **IAPR TC-12:**
  - CRM (p=0.70) vs BS-CRM (p=0.70): 19% increase in F1
  - T-test: $p \leq 2 \times 10^{-9}$

Summary & Conclusions

- **Contributions**
  - BS-CRM: a considerably more powerful model of image annotation:
    - *Minkowski kernel* to model the covariance of image features.
    - *Beam search* to select the optimal mutually correlated tag set.
  - Consistent performance gains on standard evaluation datasets.
  - Much greater recall of the more rarer words in the vocabularly.
  - Ideas could be used to improve the performance of other models.
- **Future Work**
  - Datasets with larger number of average tags per image: more correlations for set based model.
  - Dynamically adapt model parameters, beam width B, μ and β.

Thank you for listening