Multimodal Poisson Gamma Belief Network

Chaojie Wang*, Bo Chen*, Mingyuan Zhou#

* National Laboratory of Radar Signal Processing, Xidian University, Xi'an, Shaanxi, China
* Collaborative Innovation Center of Information Sensing and Understanding at Xidian University
# IROM Department, The University of Texas at Austin, Austin, TX, USA

西安电子科技大学 XIDIAN UNIVERSITY

THE UNIVERSITY OF TEXAS AT AUSTIN McCOMBS SCHOOL OF BUSINESS

Motivation

Multimodal learning prefers models can

- extracting a joint representation
- filling in missing modality
- exploiting the connections between different data modalities

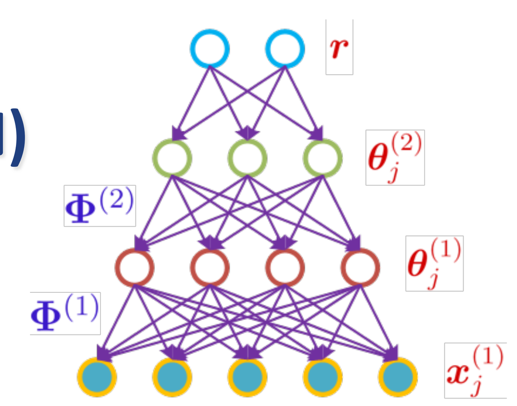
However, most existing methods

- fall short of extracting interpretable multilayer hidden structures
- have trouble visualizing the relationships between modalities
- need to normalize scales of input data

Thus, we propose a novel deep multimodal model whose latent multilayer network can be easily interpreted based on Poisson Gamma Belief Network which can be represented as deep LDA equivalently.

Background

Poisson Gamma Belief Network (PGBN)



$$\boldsymbol{\theta}_j^{(L)} \sim \mathrm{Gam}\left(\boldsymbol{r}, 1/c_j^{(L+1)}\right),$$
$$\cdots$$
$$\boldsymbol{\theta}_j^{(l)} \sim \mathrm{Gam}\left(\boldsymbol{\Phi}^{(l+1)}\boldsymbol{\theta}_j^{(l+1)}, 1/c_j^{(l+1)}\right),$$
$$\cdots$$
$$\boldsymbol{x}_j^{(1)} \sim \mathrm{Pois}\left(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}\right),\ \boldsymbol{\theta}_j^{(1)} \sim \mathrm{Gam}\left(\boldsymbol{\Phi}^{(2)}\boldsymbol{\theta}_j^{(2)}, \frac{p_j^{(2)}}{1-p_j^{(2)}}\right),$$

Priors: $\phi_k^{(l)} \sim \mathrm{Dir}(\eta^{(l)}\mathbf{1}_{K_{l-1}}),\ \boldsymbol{r} \sim \mathrm{Gam}(\gamma_0/K_L, 1/c_0),\ p_j^{(2)} \sim \mathrm{Beta}(a_0, b_0),\ c_j^{(l)} \sim \mathrm{Gam}(e_0, 1/f_0)$

Different Data Formulation

If the observations are high-dimensional sparse binary vectors $\boldsymbol{b}_j^{(1)} \in \{0,1\}^V$, they are factorized as

$$\boldsymbol{b}_j^{(1)} = \mathbf{1}(\boldsymbol{x}_j^{(1)} \geq 0),\ \boldsymbol{x}_j^{(1)} \sim \mathrm{Pois}(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}).$$

If the observations are high-dimensional nonnegative real-value vector $\boldsymbol{y}_j^{(1)} \in \mathbb{R}_+^V$, they are factorized as

$$\boldsymbol{y}_j^{(1)} \sim \mathrm{Gam}(\boldsymbol{x}_j^{(1)}, 1/a_j),\ \boldsymbol{x}_j^{(1)} \sim \mathrm{Pois}(\boldsymbol{\Phi}^{(1)}\boldsymbol{\theta}_j^{(1)}).$$

Contributions

We construct a novel multimodal PGBN that well captures **the correlations between different modalities** at multiple levels of abstraction and these coupled topics visualized by our structure exhibit an increasing level of abstraction when **moving towards a deeper hidden layer**.

Multimodal PGBN

From the top to the bottom, the generative model is expressed as

$$\boldsymbol{\theta}_{share\_j}^{(L)} \sim \mathrm{Gam}(\boldsymbol{r}_{share}, 1/c_{share\_j}^{(L+1)}),$$
$$\cdots$$
$$\boldsymbol{\theta}_{share\_j}^{(l)} \sim \mathrm{Gam}(\boldsymbol{\Phi}_{share}^{(t+1)}\boldsymbol{\theta}_{share\_j}^{(l+1)}, 1/c_{share\_j}^{(l+1)}),$$
$$\cdots$$
$$\boldsymbol{x}_{img\_j}^{(1)} \sim \mathrm{Pois}(\boldsymbol{\Phi}_{img}^{(1)}\boldsymbol{\theta}_{share\_j}^{(1)}),\ \boldsymbol{x}_{txt\_j}^{(1)} \sim \mathrm{Pois}(\boldsymbol{\Phi}_{txt}^{(1)}\boldsymbol{\theta}_{share\_j}^{(1)}).$$

Priors: $\phi_{txt\_k}^{(1)} \sim Dir(\eta^{(1)}\mathbf{1}_{K_{txt}})$, $\phi_{img\_k}^{(1)} \sim Dir(\eta^{(1)}\mathbf{1}_{K_{img}})$, $\phi_{share\_k}^{(l)} \sim Dir(\eta^{(l-1)}\mathbf{1}_{K_{l-1}})$
$r \sim Gam(\gamma_0/K_L, 1/c_0)$, $p_j^{(2)} \sim Beta(a_0, b_0)$, $c_j^{(l)} \sim Gam(e_0, 1/f_0)$

The upward-downward sampler can be applied to train the hidden layers of mPGBN jointly, with the sampling update equation for the first hidden layer replaced as

$$(\theta_{share\_j}^{(1)} \mid -) \sim \mathrm{Gam}(m_{img\_j}^{(1)(2)} + m_{tags\_j}^{(1)(2)} + \Phi_{share}^{(2)}\theta_{share\_j}^{(2)}, [c_j^{(2)} - 2\ln(1-p_j^{(1)})]^{-1}),$$

where the both modalities influence the conditional posteriors of hidden layers.

Adaptive Normalization

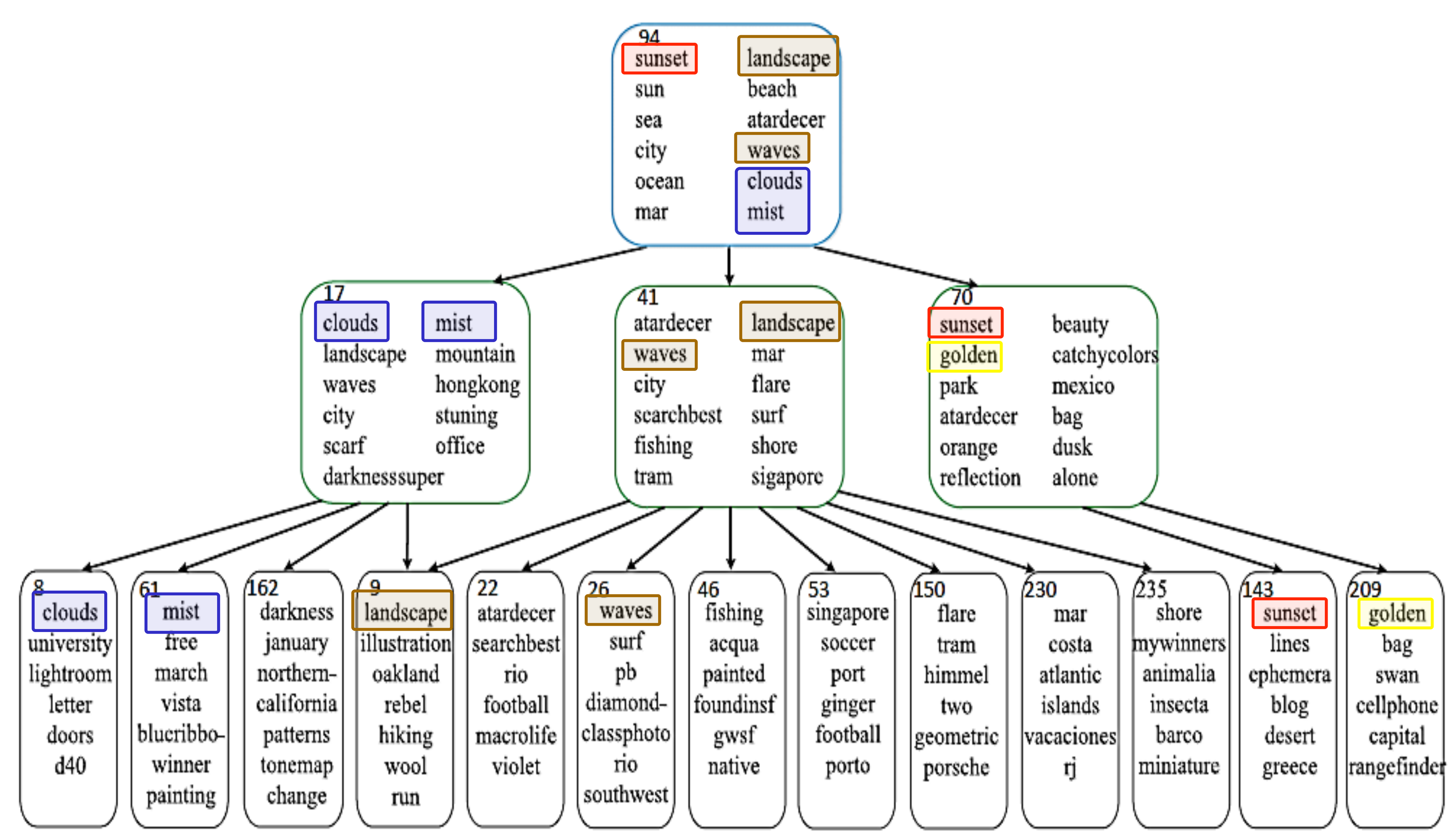
To handle different input data scales, we propose to modify the mPGBN model as following

$$\theta_{img\_j}^{(1)} = k_{img\_j}\theta_{share\_j}^{(1)},\ \theta_{txt\_j}^{(1)} = k_{txt\_j}\theta_{share\_j}^{(1)},$$
$$x_{img\_j}^{(1)} \sim \mathrm{Pois}(\Phi_{img}^{(1)}\theta_{img\_j}^{(1)}),\ x_{txt\_j}^{(1)} \sim \mathrm{Pois}(\Phi_{txt}^{(1)}\theta_{txt\_j}^{(1)}),$$

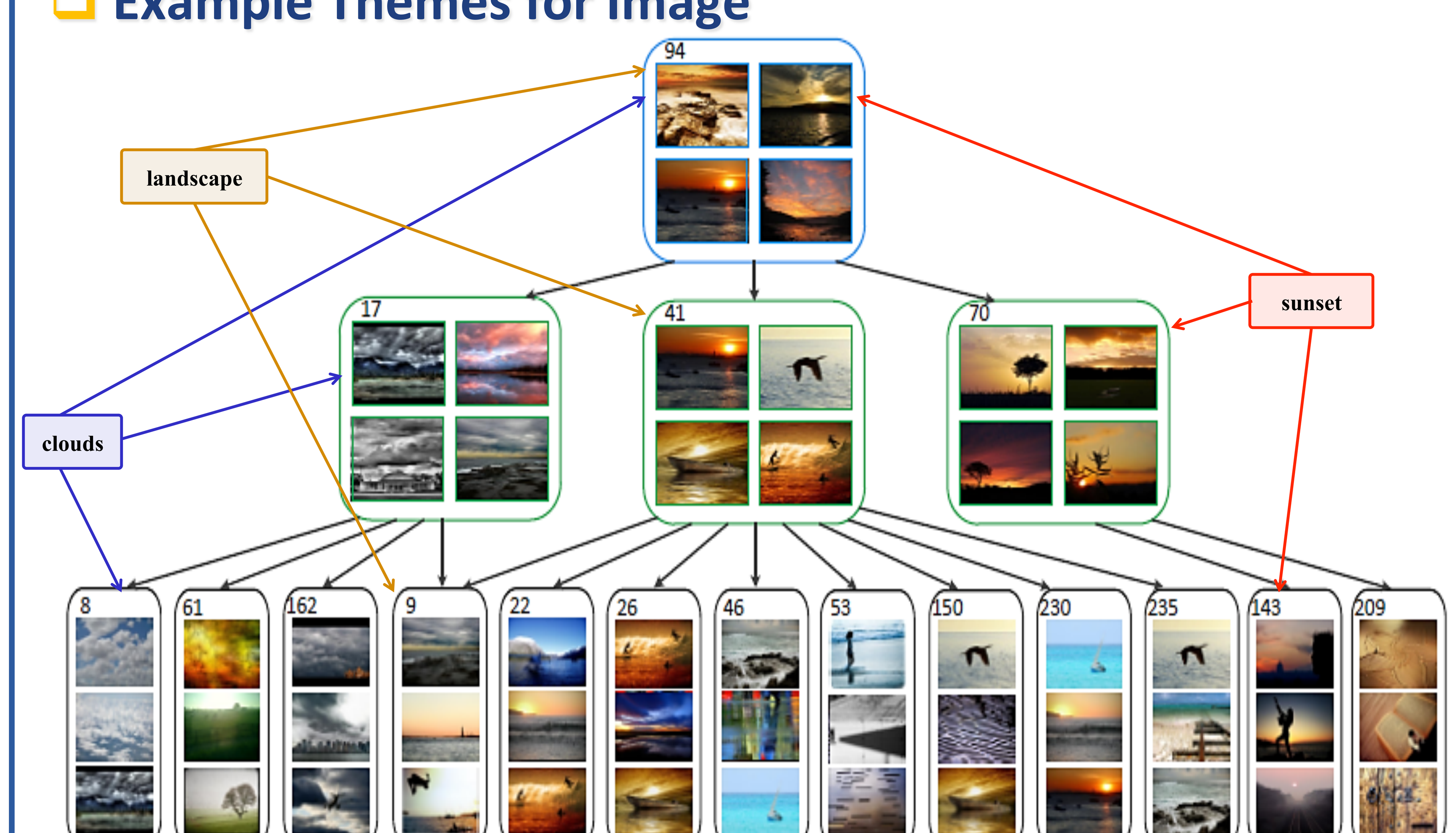
which means that the first hidden layers of both modalities only share their gamma shape parameters but have adaptive scale parameters to suit different input scales.

Experiment Results

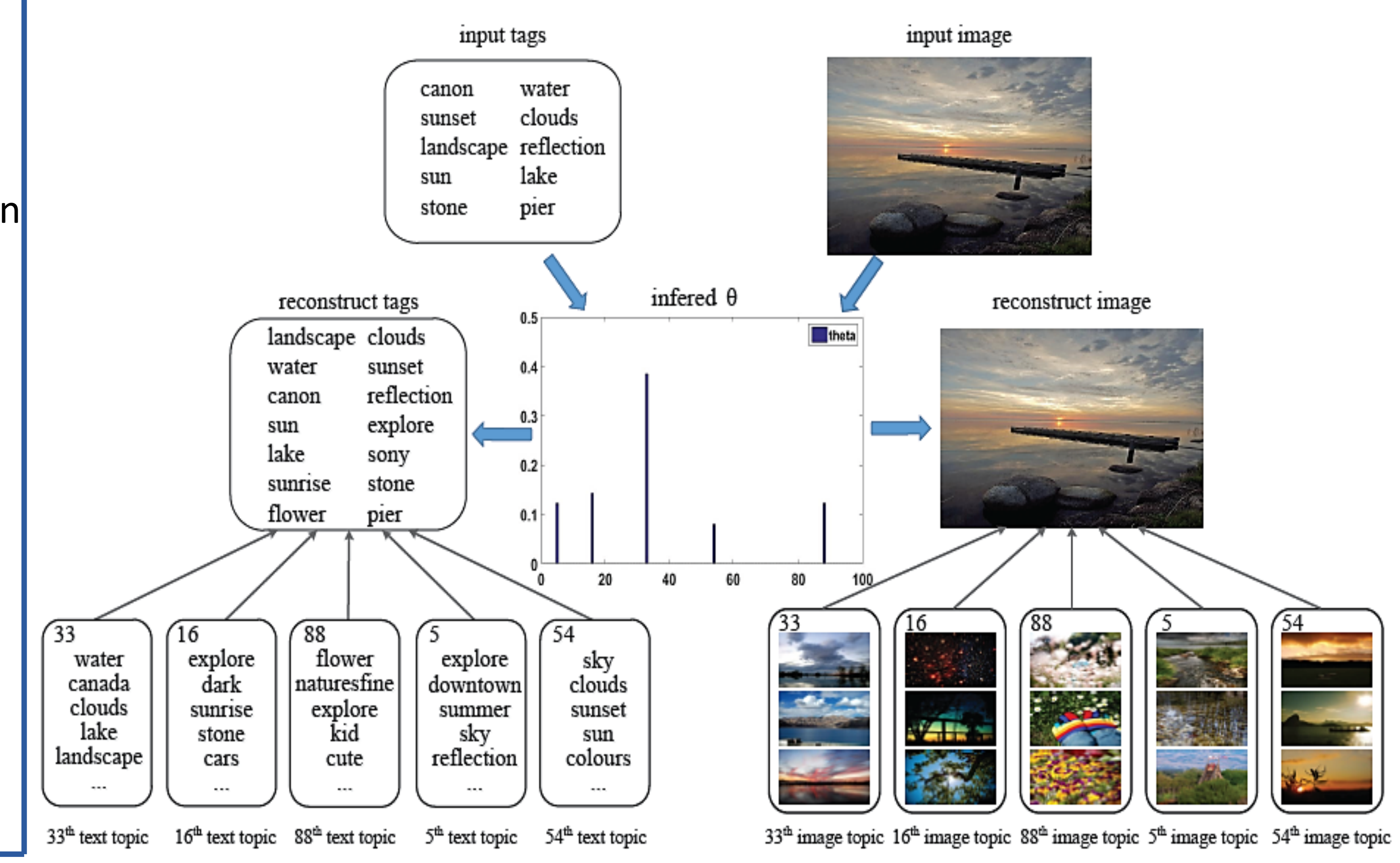
Example Topics for Text



Example Themes for Image



Visualizing the generative process of input image-tags pair



Filling Missing Modality

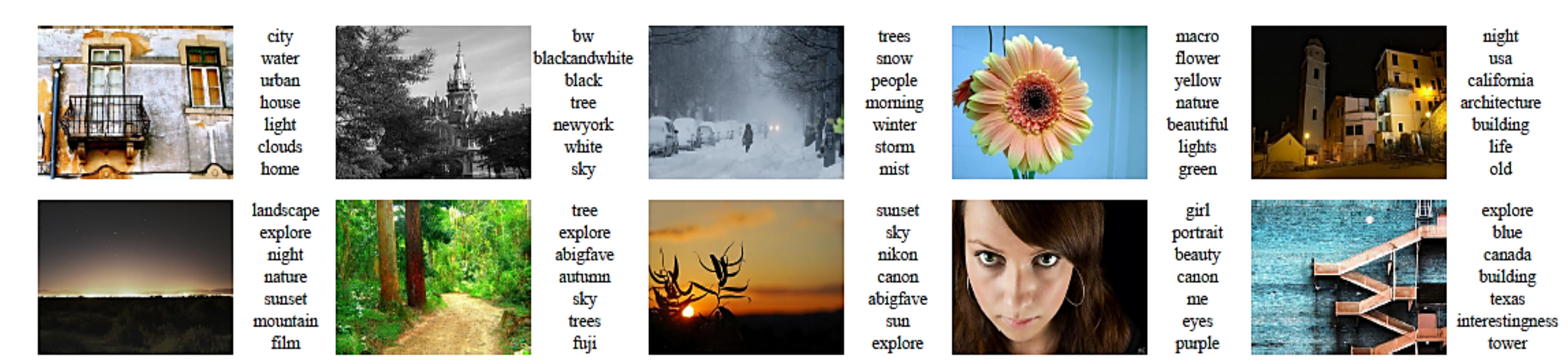


Figure 4: Examples of text generated by Multimodal PGBN conditioned on images

Figure 5: Top-5 nearest images retrieved using the features generated by the multimodal PGBN conditioning on the tags

Performance

Table 1: Comparison of AP scores and Precision@50 of various multimodal models on the MIR-Flicker dataset.

LABELS	ANIMALS	BABY	BABY*	BIRD	BIRD*	CAR	CAR*	CLOUDS	CLOUDS*	DOG
RANDOM	0.129	0.010	0.005	0.030	0.019	0.047	0.015	0.148	0.054	0.027
LDA	0.537	0.285	0.308	0.426	0.500	0.297	0.389	0.654	0.528	0.621
SVM	0.531	0.200	0.165	0.443	0.520	0.339	0.434	0.685	0.434	0.607
DBN	0.498	0.129	0.134	0.184	0.255	0.309	0.354	0.759	0.691	0.342
DBM	0.511	0.139	0.145	0.190	0.253	0.319	0.368	**0.768**	**0.723**	0.351
mPFA	0.603	0.260	0.297	0.487	0.531	0.332	0.496	0.643	0.509	0.601
mPGBN	**0.615**	**0.288**	**0.320**	**0.515**	**0.552**	**0.357**	**0.502**	0.657	0.554	**0.609**
LABELS	DOG*	FEMALE	FEMALE*	FLOWER	FLOWER*	FOOD*	INDOOR	LAKE*	MALE	MALE*
RANDOM	0.024	0.247	0.159	0.073	0.043	0.040	0.333	0.032	0.243	0.146
LDA	**0.663**	0.494	0.454	0.560	0.623	0.439	0.663	0.258	0.434	0.354
SVM	0.641	0.465	0.451	0.480	0.717	0.308	0.683	0.207	0.414	0.335
DBN	0.376	0.540	0.478	0.593	0.679	0.447	0.750	0.262	0.503	0.406
DBM	0.385	0.535	0.493	0.604	0.668	0.462	**0.759**	**0.277**	**0.505**	**0.424**
mPFA	0.650	0.519	0.468	0.605	0.714	0.562	0.678	0.262	0.477	0.382
mPGBN	0.656	**0.551**	**0.497**	**0.614**	**0.736**	**0.579**	0.692	0.268	0.488	0.399
LABELS	NIGHT	NIGHT*	PEOPLE	PEOPLE*	PLANTLIFE	PORTRAIT	PORTRAIT*	RIVER	RIVER	SEA
RANDOM	0.108	0.027	0.415	0.314	0.351	0.157	0.153	0.036	0.006	0.053
LDA	0.615	0.420	0.731	0.664	0.703	0.543	0.541	**0.317**	**0.134**	0.477
SVM	0.588	0.450	0.748	0.565	0.691	0.480	0.558	0.158	0.109	0.529
DBN	0.655	0.483	0.800	0.730	0.791	0.642	0.635	0.263	0.110	**0.586**
DBM	**0.666**	**0.505**	**0.802**	**0.742**	**0.794**	**0.651**	**0.665**	0.274	0.110	0.582
mPFA	0.599	0.373	0.768	0.692	0.744	0.522	0.516	0.299	0.118	0.524
mPGBN	0.625	0.407	0.781	0.719	0.759	0.547	0.541	0.301	0.121	0.533
LABELS	SEA*	SKY	STRUCTURES	SUNSET	TRANSPORT	TREE	TREE*	WATER	MAP	Prec@50
RANDOM	0.009	0.316	0.400	0.085	0.116	0.187	0.027	0.133	0.124	0.124
LDA	0.197	0.800	0.709	0.528	0.411	0.515	0.342	0.575	0.492	0.754
SVM	0.201	0.823	0.695	0.613	0.369	0.559	0.321	0.527	0.475	0.758
DBN	0.259	0.873	0.787	0.648	0.406	0.660	0.483	0.629	0.503	-
DBM	0.260	**0.883**	**0.796**	**0.659**	0.423	**0.668**	**0.492**	0.628	0.513	0.791
mPFA	0.280	0.798	0.748	0.510	0.445	0.520	0.360	0.622	0.515	0.834
mPGBN	**0.343**	0.809	0.764	0.516	**0.455**	0.539	0.377	**0.630**	**0.532**	**0.844**

AAAI 2018: Thirty-Second AAAI Conference on Artificial Intelligence