Large-Margin Determinantal Point Processes

Wei-Lun Chao*[1], Boqing Gong*[1], Kristen Grauman[2], and Fei Sha[1]

[1]University of Southern California, [2]University of Texas at Austin

Highlights

- **Investigate determinantal point processes (DPPs) for discriminative subset selection**
- **Propose *margin* based parameter estimation to explicitly track errors in selecting subsets**
- **Balance different types of evaluation metrics, e.g., precision and recall**
- **Improve modeling flexibility by multiple-kernel based parameterization**
- **Attain state-of-the-art performance on the tasks of video and document summarization**

Background

- A DPP defines a probabilistic distribution over the power set of a ground set: *diverse subsets with large probabilities*

  Ground set of M items, $\mathcal{Y}=\{1,2,...,M\}$

  $\boldsymbol{L}\in\mathbb{S}_+^M$: a kernel matrix of pairwise similarities

  $$P(\boldsymbol{y}\subseteq\mathcal{Y};\boldsymbol{L})=\frac{\det(\boldsymbol{L}_{\boldsymbol{y}})}{\det(\boldsymbol{L}+\boldsymbol{I})}$$

  $$P(\boldsymbol{y}=\{i,j\};\boldsymbol{L})\propto\det\left(\boldsymbol{L}_{\{i,j\}}\right)=L_{ii}L_{jj}-L_{ij}^2$$

- DPPs offer a powerful approach to modeling ***diversity*** in applications where the goal is to select a diverse subset from a ground set of items (e.g., retrieval, summarization)
- MAP inference (NP-hard): $\boldsymbol{y}^{MAP}=\arg\max_{\boldsymbol{y}} P(\boldsymbol{y};\boldsymbol{L})$
- Estimate the kernel $\boldsymbol{L}$ from labeled data $\left\{\left(\boldsymbol{y}^{*(n)},\mathcal{Y}^{(n)}\right)\right\}$
  - ➢ Reparameterization: $\boldsymbol{L}^{(n)}\left(\mathcal{Y}^{(n)};\theta\right)$
  - ➢ Standard method: Maximum likelihood estimation (MLE)

  $$\theta^{MLE}=\arg\max_{\theta}\sum_n \log P\left(\boldsymbol{y}^{*(n)};\boldsymbol{L}^{(n)}\left(\mathcal{Y}^{(n)};\theta\right)\right)$$

Problems with existing methods

Statistical challenges

- ➢ Limited number of training samples

Modelling challenges

- ➢ Limited power in parameterizing kernels with the widely used quality-diversity (QD) decomposition
- ➢ Unable to track discriminative errors in selecting subsets
- ➢ Unable to differentiate different types of metrics (e.g., precision, recall) for complex structured prediction tasks

Contribution I: Multiple-kernel representation (MKR)

- Quality-diversity (QD) decomposition

$$\forall i\in\mathcal{Y}\begin{cases}\boldsymbol{x}_i: \text{quality features}\\ \phi_i: \text{similarity features}\end{cases}\Rightarrow\begin{cases}L_{ij}=q_iq_jS_{ij}=q_iq_j\phi_i^T\phi_j\\ q_i=q(\boldsymbol{x}_i)=\exp\left(\theta^T\boldsymbol{x}_i\right)\end{cases}$$

- Multiple-kernel representation (MKR)

$$S_{ij}=\sum_k \alpha_k \exp\left\{-\|\phi_i-\phi_j\|_2^2/\sigma_k^2\right\}+\ \phi_i^T\phi_j,\quad \text{s.t.}\ \sum_k \alpha_k+\ =1$$

Selected references:

[1] A. Kulesza and B. Taskar. Determinantal point processes for machine learning. 2012.

[2] H. T. Dang. Overview of DUC 2005. In Document Understanding Conf., 2005.

[3] S. E. F. de Avila et al. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recognition Letters, 2011.

Contribution II: Margin based parameter estimation (LME)

- Maintain desired margins between correct/incorrect subsets

$$\log P(\boldsymbol{y}^*;\boldsymbol{L})\ge\max_{\boldsymbol{y}\subseteq\mathcal{Y}}\log\left\{\ell(\boldsymbol{y}^*,\boldsymbol{y})P(\boldsymbol{y};\boldsymbol{L})\right\}$$
$$=\max_{\boldsymbol{y}\subseteq\mathcal{Y}}\log\ell(\boldsymbol{y}^*,\boldsymbol{y})+\log P(\boldsymbol{y};\boldsymbol{L})$$

➢ *The multiplicative margin leads to tractable optimization*

- Measure the subset discrepancy by structured loss functions

$$\ell_\mu(\boldsymbol{y}^*,\boldsymbol{y})=\underbrace{\sum_{i\notin\boldsymbol{y}^*}\mathbb{I}[i\in\boldsymbol{y}]}_{\text{precision}}+\mu\underbrace{\sum_{i\in\boldsymbol{y}^*}\mathbb{I}[i\notin\boldsymbol{y}]}_{\text{recall}}$$

- Optimization: Jensen's inequality (softmax) for tractability

$$\log P(\boldsymbol{y}^*;\boldsymbol{L})\ge\text{softmax}_{\boldsymbol{y}\subseteq\mathcal{Y}}\log\ell_\mu(\boldsymbol{y}^*,\boldsymbol{y})+\log P(\boldsymbol{y};\boldsymbol{L})$$
$$=\log\left(\sum_{i\notin\boldsymbol{y}^*}K_{ii}+\mu\sum_{i\in\boldsymbol{y}^*}(1-K_{ii})\right),\ \text{where } K=\boldsymbol{L}(\boldsymbol{L}+\boldsymbol{I})^{-1}$$

- Objective function: hinge loss $[.]_+=\max(0, .)$

$$\min\sum_n\left[-\log P\left(\boldsymbol{y}^{*(n)};\boldsymbol{L}^{(n)}\right)+{}^{a}\log\left(\sum_{i\notin\boldsymbol{y}^{*(n)}}K_{ii}^{(n)}+\mu\sum_{i\in\boldsymbol{y}^{*(n)}}\left(1-K_{ii}^{(n)}\right)\right)\right]_+$$

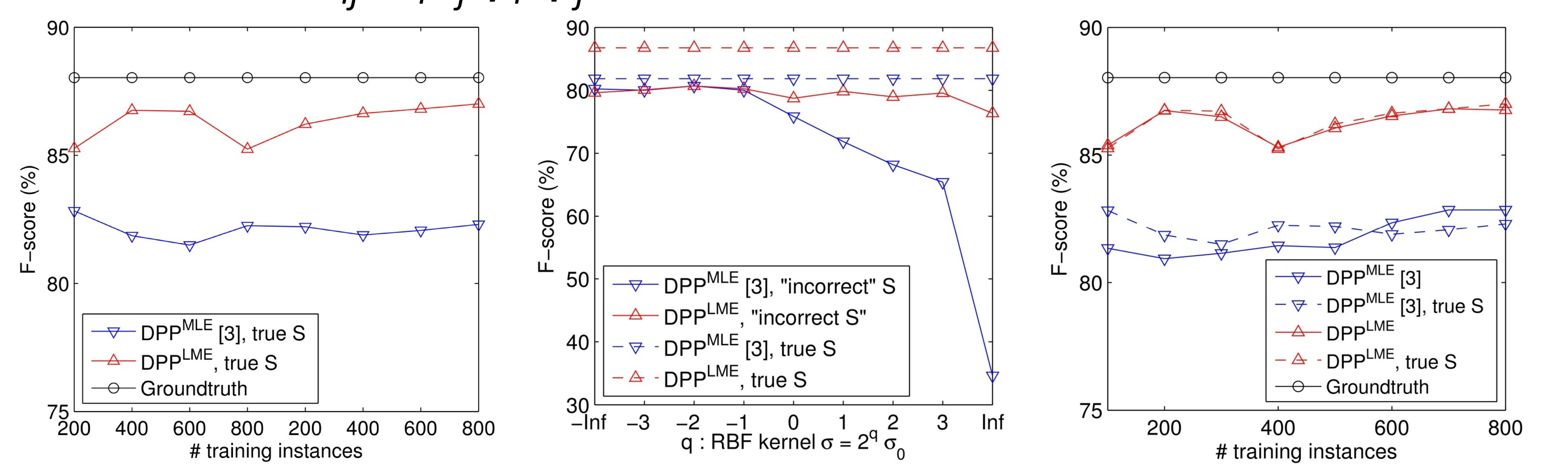
Experiments

Evaluation: **P**recision, **R**ecall, **F**-score (harmonic mean of **P**, **R**)

Inference: Brute-force search, minimum Bayesian risk (MBR)

Synthetic dataset:

- ➢ Generate ground sets and target subsets by sampling $\{\boldsymbol{x}_i=\boldsymbol{\phi}_i\}$, $\boldsymbol{\theta}$, computing $S_{ij}=q_iq_j\,\boldsymbol{\phi}_i^T\boldsymbol{\phi}_j$, brute-force search, and adding noise



learning $\boldsymbol{\theta}$ with true $\boldsymbol{S}$ | learning $\boldsymbol{\theta}$ with wrong $\boldsymbol{S}$ | learning $\boldsymbol{\theta}$ and $\boldsymbol{S}$

- ➢ Our method **(MKR+LME)** is more robust in parameter estimation

Document summarization: Document Understanding Conf. (DUC)

- ➢ Train: DUC 2003 (60 clusters) Test: DUC 2004 (50 clusters)

Method	ROUGE1-F	ROUGE1-P	ROUGE1-R	ROUGE2-F	ROUGE2-P	ROUGE2-R
PEER65 [2]	37.9	37.6	38.2	9.13	--	--
DPP^{MLE}+COS	37.9±0.08	37.4±0.08	38.5±0.08	7.72±0.06	7.63±0.06	7.83±0.06
Ours (DPP^{LME}+COS)	38.4±0.09	37.7±0.10	39.1±0.08	8.20±0.07	8.07±0.07	8.35±0.07
Ours (DPP^{MLE}+MKR)	39.1±0.08	39.0±0.09	39.3±0.09	9.25±0.08	9.24±0.08	9.27±0.08
Ours (DPP^{LME}+MKR)	**39.7**±0.05	**39.6**±0.08	**39.9**±0.06	**9.40**±0.08	**9.38**±0.08	**9.43**±0.08

Video summarization: Open Video Project (OVP)

- ➢ 50 videos from OVP, 5 user annotations, 5-fold cross-validation

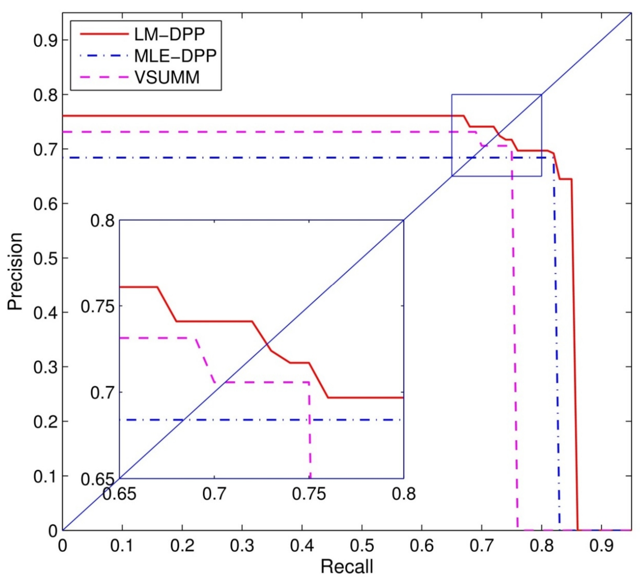
Metric	VSUMM [3]		DPP^{MLE} +MKR	**Ours** (DPP^{LME}+MKR)		
	Type1	Type2		$\omega=2^{-6}$	$\omega=2^{0}$	$\omega=2^{6}$
F	70.25	68.20	72.94	71.25	**73.46**	72.39
P	70.57	73.14	68.40	**74.00**	69.68	67.19
R	75.77	69.14	82.51	72.71	81.39	**83.24**

Oracle (F=88, P=88, R=91)

Ours ($\omega=2^{-6}$) (F=77, P=76, R=81)

Ours ($\omega=1$) (F=67, P=55, R=91)

Ours ($\omega=2^{6}$) (F=63, P=48, R=97)



Our method can balance different types of evaluation metrics and achieve better performance in both summarization tasks

Acknowledgements: F. S., B. G., and W. C. are supported by ARO Award #W911NF-12-1-0241, DARPA Award #D11AP00278, NSF IIS Award #1065243, ONR #N00014-12-1-0066, and Alfred P. Sloan Fellowship. K. G. is supported by ONR YIP #N00014-12-1-0754.