Auto-Summarization: A Step Towards Unsupervised Learning of a Submodular Mixture

Chandrashekhar Lavania and Jeff Bilmes
MELODI Lab @ University Of Washington, Seattle

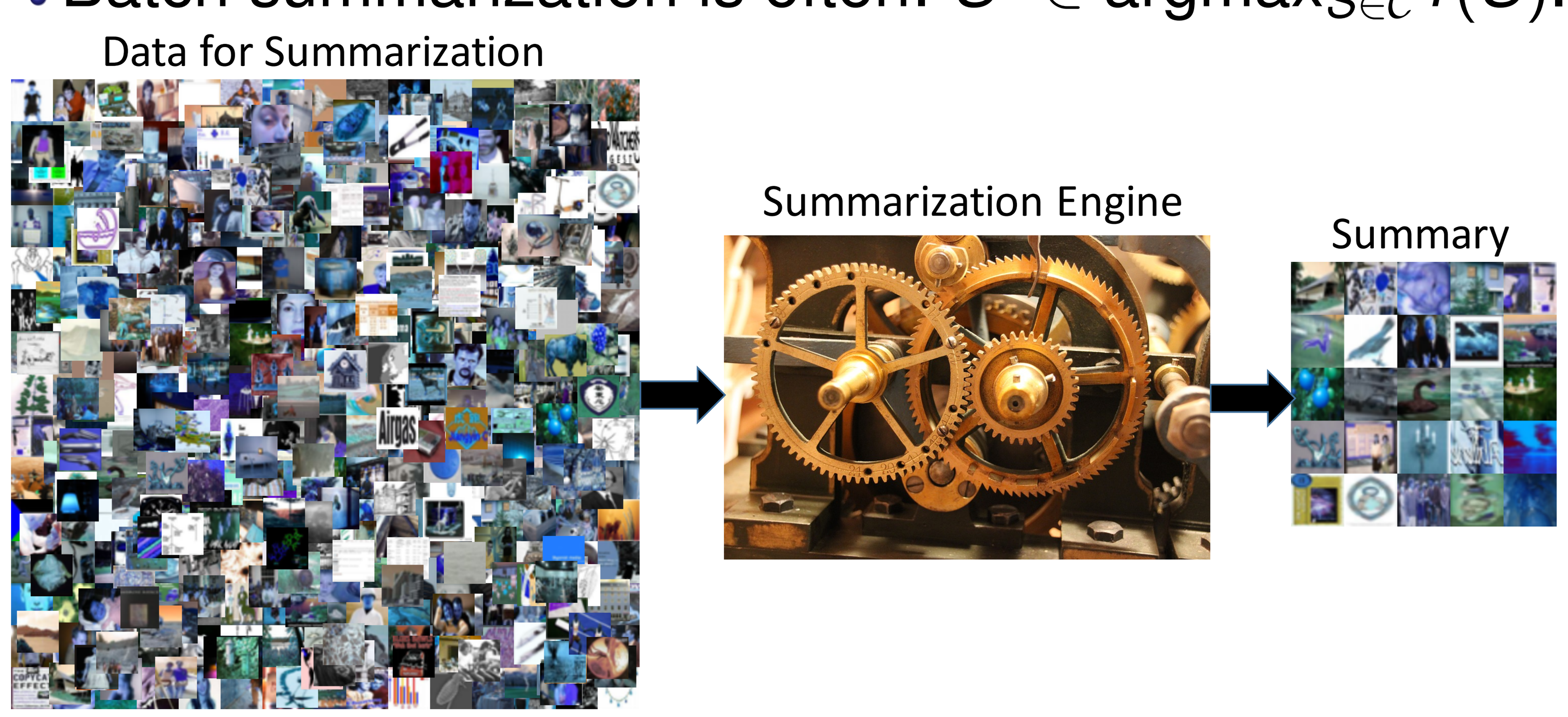
MachinE Learning, Optimization, & Data Interpretation @ UW
http://melodi.ee.washington.edu

Overview

- Massive amounts of data is generated daily via a plethora of sources. (eg: videos, different types sensors, text sources etc.).
- Processing of this data is resource intensive.
- Summarization can help extract essential information from the data.
- Lack of "training summaries prevents supervised learning of summarization objective.
- Unsupervised approaches are thus needed.

Submodular functions for Summarization

- Given $V = \{v_1, v_2, \ldots, v_n\}$. Then $f : 2^V \to \mathbb{R}$ is *submodular* if $f(a|A) \geq f(a|B)\ \ \forall A \subseteq B \subseteq V,\ a \in V \setminus B$, where $f(a|A) \triangleq f(a \cup A) - f(A)$.
- They have shown merit in a variety of summarization and data selection tasks.
- Batch summarization is often: $S^* \in \text{argmax}_{S \in \mathcal{C}} f(S)$.



Aim

- Learn a submodular mixture $F_w(\cdot) = \sum_i w_i f_i(\cdot)$ using minimal hyperparameters and without the incorporation of ground truth information in the learning objective.
- Each $f_i(\cdot)$ is submodular and $||w|| = 1$.
- The mixture is learned using objective $J(w) = \sum_l \lambda_l J_l(w)$ as:

$$\max_{w \geq 0, ||w||=1} J(w)$$

Feature Based Submodular Functions

- Rich and resource efficient class of submodular functions
- Defined in the form $F_w(A) = \sum_{u \in U} w_u \phi_u(m_u(A))$.
- U is a set of features, $w \in \mathbb{R}_+^U$ (for $u \in U$, $w_u \geq 0$ is a feature weight), ϕ_u is monotone non-decreasing concave function, and $m_u : V \to \mathbb{R}_+$ is a non-negative normalized modular function specific to feature u.
- They have been successfully used in a variety of summarization tasks.

Additively Contributive Features (AC)

- Let $\mathfrak{e}(x_v) = \{\mathfrak{e}(x_v)(1), \mathfrak{e}(x_v)(2) \ldots \mathfrak{e}(v)(d)\}$ be the d dimensional representation of an object x_v in feature space U where $v \in V$ is an index of a data item .
- Given samples x_{v_1} and x_{v_2}, if for any $u \in U$, $\mathfrak{e}(x_{v_1})(u) < \mathfrak{e}(x_{v_2})(u)$ implies that object x_{v_2} has more of the property represented by feature u as compared to x_{v_1}, then the representation is AC.
- Moreover, for two or more objects indexed by $A \subseteq V$, then the objects should contribute to property u additively, as in $\sum_{a \in A} \mathfrak{e}(a)(u)$.

Autoencoders for AC Feature Generation

- The encoding-decoding process for item x can be considered as $\mathfrak{d}(\mathfrak{e}(x))$.
- Let $W \in \mathbb{R}^{d' \times d}$ such that $\mathfrak{d}(\mathfrak{e}(x)) = \mathfrak{d}'(W\mathfrak{e}(x))$.
- Restrict W to be non-negative during the training process.

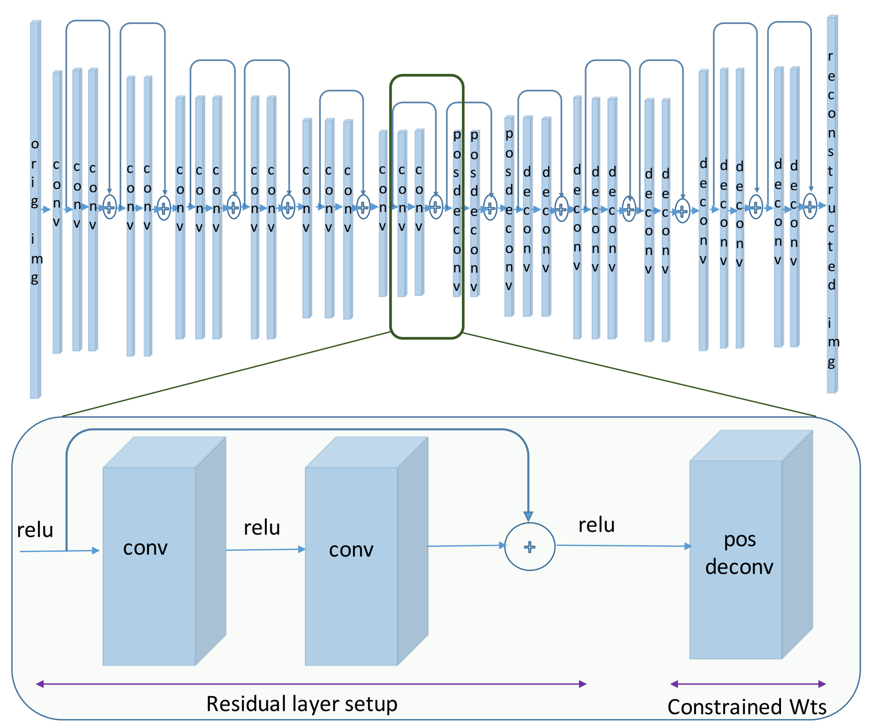


Figure 1: Sample architecture for constructing AC image features. The zoomed view shows that the first layer after the bottleneck is pos deconv. A pos deconv layer's weight matrices are non-negatively constrained during training.

Meta-Objectives

- Confidence

$$J_1(w) = \mathbb{E}_{k \sim p}\left[\left|\max_{S \subseteq V, |S|=k} F_w(S) - \min_{S' \subseteq V, |S'| \geq k} F_w(S')\right|\right]$$

- Entropy

$$J_2(w) = -\sum_{u,\gamma} w_{u,\gamma} log(w_{u,\gamma})$$

- Non-modularity

$$J_3(w) = \mathbb{E}_{k \sim p}\left[\sum_{s \in \hat{S}} F_w(s) \quad - F_w(\hat{S}) \middle| \hat{S} \in \arg\max_{S \subseteq V : |S| \leq k} F_w(S)\right]$$

- Curvature

$$J_4(w) = 1 - \min_{j \in V} \frac{F_w(j|V \setminus j)}{F_w(j)}$$

- Stability

$$J_5(w) = -\mathbb{E}_{p_w}\left[\mathbb{E}_{k \sim p}\left[\max_{S \subseteq V, |S|=k} F_w(S) \quad - \max_{S' \subseteq V, |S'|=k} F_{\hat{w}}(S')\right]^2\right]$$

- Soft De-duplication before saturation

$$J_6(w) = \mathbb{E}_{k \sim p}\left[\sum_{i=1}^{k} \left(1 - \frac{F_w(\hat{S}_i)}{F_w(V)}\right) \frac{F_w(v_i|\hat{S}_{i-1})}{\psi(v_i, \hat{S}_{i-1})} \middle| \hat{S}_i \in \text{gargmax}_{S \subseteq V, |S|=i} F_w(S), v_i \in \hat{S}_i \setminus \hat{S}_{i-1}\right]$$

Results

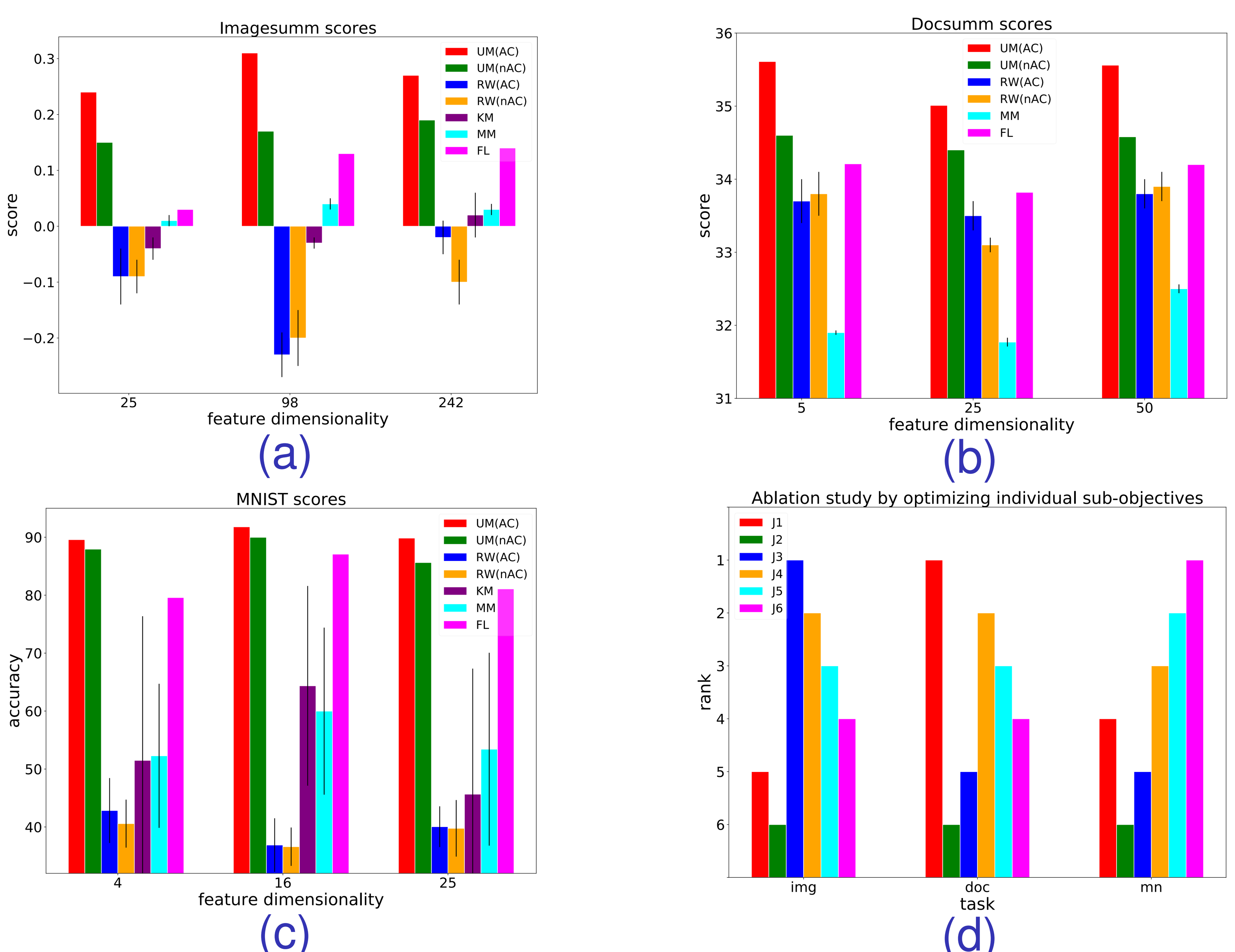


Figure 2: Comparison of Unsupervised Mixture (UM) and other baselines. Rank 1 is best.

Reference:
Lavania, C. and Bilmes, J. Auto-summarization: A step towards unsupervised learning of a submodular mixture. SDM 2019.