Deep Supervised Summarization: Algorithm and Application to Learning Instructions

Chengguang Xu xu.cheng@husky.neu.edu

Ehsan Elhamifar eelhami@ccs.neu.edu

Khoury College of Computer Sciences, Northeastern University, Boston, USA

Motivation

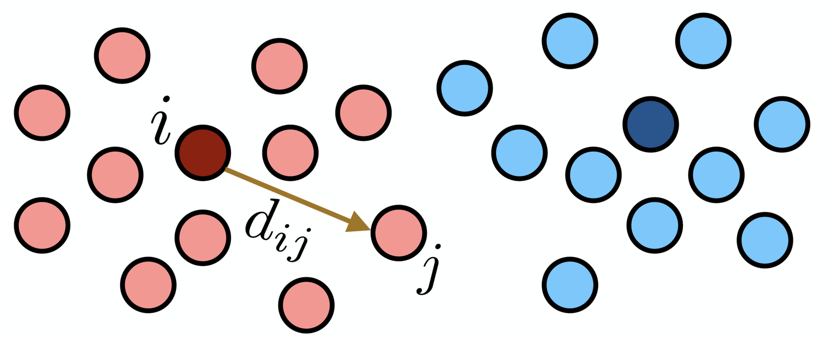
- **Supervised subset selection** aims to learn from ground-truth summaries.
  - Humans perform remarkably well in summarization of video and speech data.
- Supervised subset selection is different and **more challenging than classification**.
  - Label **'rep'** vs **'non-rep'** depends on relationships among entire data.
- Majority of existing work focus on unsupervised subset selection.
  - Few existing supervised methods naively treat the problem as classification.

Contributions

- Develop a **theoretically-motivated supervised subset selection** framework.
- Propose a **representation learning** method using which subset selection **recovers ground-truth summaries**.
  - Investigate **theoretical conditions** under which **facility location** recovers **ground-truth representatives** of a dataset.
  - Use the theory to design a **new loss function** for representation learning.
- Outperforms SOTA on **learning from instructional videos** on two large datasets.

Subset Selection via Facility Location

- Given: dataset $\{\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_N\}$ and pairwise dissimilarities $\{d_{ij}\}_{i,j=1,\dots,N}$.
- d_{ij}: how well $\boldsymbol{y}_i$ represents $\boldsymbol{y}_j$, smaller means better.



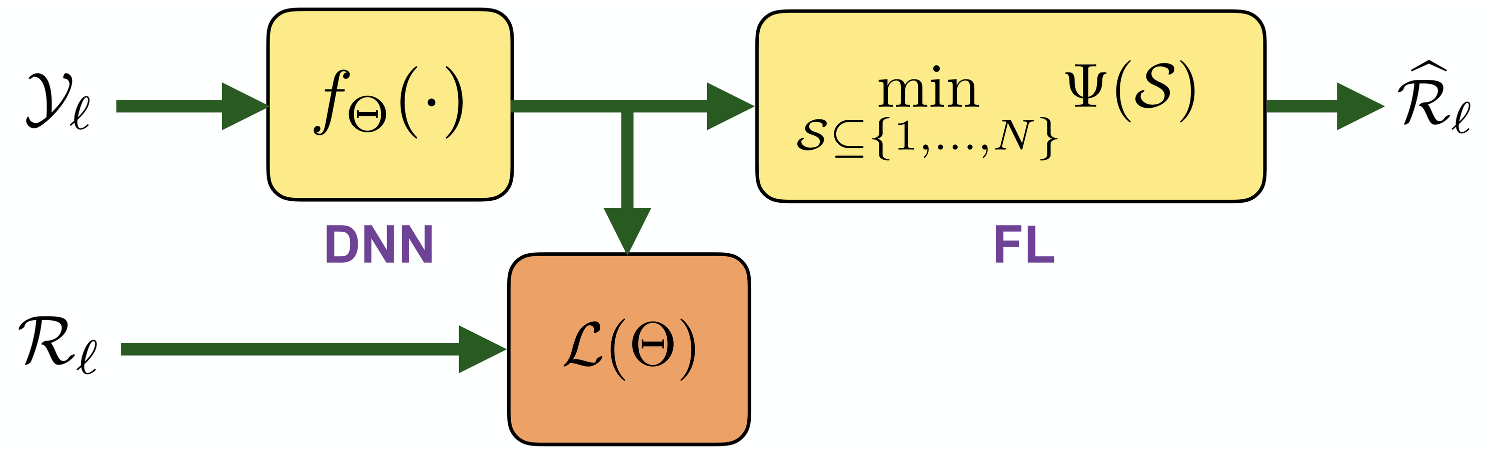
$$\mathcal{S} = \{ \bullet , \bullet \}$$

- Goal: find a small subset $\mathcal{S} \subseteq \{1, \dots, N\}$ to represent the dataset.
- Minimize **cardinality plus encoding quality** cost of the representative set:

$$\min_{\mathcal{S} \subseteq \{1,\dots,N\}} \lambda |\mathcal{S}| + \sum_{j=1}^{N} \min_{i \in \mathcal{S}} d_{ij}$$

Deep Supervised Subset Selection

- Given datasets and their ground-truth representatives, $\{(\mathcal{Y}_\ell, \mathcal{R}_\ell)\}_{\ell=1}^{L}$
  - $\mathcal{Y}_\ell = \{\boldsymbol{y}_{\ell,1}, \dots, \boldsymbol{y}_{\ell,N_\ell}\}$ corresponds to N_ℓ data points in the ℓ-th dataset
  - $\mathcal{R}_\ell \subseteq \{1, \dots, N_\ell\}$ is the set of indices of ground-truth representatives
- **Goal**: learn $f_\Theta(\cdot)$ on input data so that running subset selection on $f_\Theta(\mathcal{Y}_\ell)$ obtains $\mathcal{R}_\ell$,



- Write facility location (FL) as an efficient sparse convex program

$$\min_{\{z_{ij}\}} \lambda \sum_{i=1}^{N} \left\| \left[z_{i1} \cdots z_{iN} \right] \right\|_\infty + \sum_{i,j=1}^{N} d_{ij} z_{ij} \quad \text{s. t.} \quad z_{ij} \geq 0, \quad \sum_{i=1}^{N} z_{ij} = 1, \quad \forall i, j.$$

- Let $\mathcal{G}_i^\ell$ denote the cluster associated with the representative $i \in \mathcal{R}_\ell$, i.e.,

$$\mathcal{G}_i^\ell = \{ j \mid i = \text{argmin}_{i'} d_{i',j}^\ell = \text{argmin}_{i'} \| f_\Theta(\boldsymbol{y}_{\ell,i'}) - f_\Theta(\boldsymbol{y}_{\ell,j}) \|_2 \}.$$

Theorem: FL and its sparse relaxation recover $\mathcal{R}_\ell$ as representatives of $\mathcal{Y}_\ell$, if:

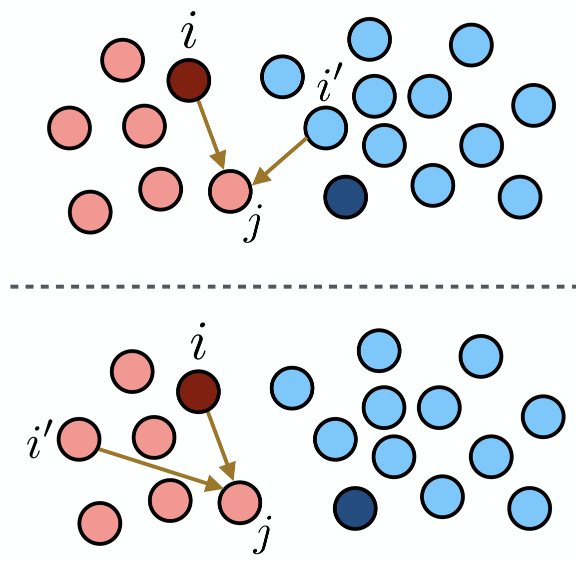
- $\forall i \in \mathcal{R}_\ell, \forall i' \in \mathcal{G}_i^\ell$, we have $\sum_{j \in \mathcal{G}_i^\ell} d_{i,j}^\ell \leq \sum_{j \in \mathcal{G}_i^\ell} d_{i',j}^\ell$;
- $\forall i \in \mathcal{R}_\ell, \forall j \in \mathcal{G}_i^\ell, \forall i' \notin \mathcal{G}_i^\ell$, we have $\frac{\lambda}{|\mathcal{G}_i^\ell|} + d_{i,j}^\ell < d_{i',j}^\ell$;
- $\forall i \in \mathcal{R}_\ell, \forall i', j \in \mathcal{G}_i^\ell$, we have $d_{i',j}^\ell \leq \frac{\lambda}{|\mathcal{G}_i^\ell|} + d_{i,j}^\ell$.

- **Proposed Learning Framework:** Use the theoretical conditions to design a loss whose minimization ensures to recover $\mathcal{R}_\ell$ as representative of $\mathcal{Y}_\ell$

$$\mathcal{L}_{medoid}^\ell(\Theta) \triangleq \sum_{i \in \mathcal{R}_\ell} \sum_{i' \in \mathcal{G}_i^\ell} \Big(\sum_{j \in \mathcal{G}_i^\ell} d_{i,j}^\ell - \sum_{j \in \mathcal{G}_i^\ell} d_{i',j}^\ell \Big)_+,$$

$$\mathcal{L}_{inter}^\ell(\Theta) \triangleq \sum_{i \in \mathcal{R}_\ell} \sum_{j \in \mathcal{G}_i^\ell} \sum_{i' \notin \mathcal{G}_i^\ell} \Big(\frac{\lambda}{|\mathcal{G}_i^\ell|} + d_{i,j}^\ell - d_{i',j}^\ell \Big)_+,$$

$$\mathcal{L}_{intra}^\ell(\Theta) \triangleq \sum_{i \in \mathcal{R}_\ell} \sum_{i', j \in \mathcal{G}_i^\ell} \Big(d_{i',j}^\ell - d_{i,j}^\ell - \frac{\lambda}{|\mathcal{G}_i^\ell|} \Big)_+,$$



$$\min_\Theta \mathcal{L}(\Theta) \triangleq \sum_{\ell=1}^{L} \Big(\mathcal{L}_{medoid}^\ell(\Theta) + \rho_{inter} \mathcal{L}_{inter}^\ell(\Theta) + \rho_{intra} \mathcal{L}_{intra}^\ell(\Theta) \Big).$$

Algorithm 1 : Supervised Facility Location Learning

Input: Datasets $\{\mathcal{Y}_\ell\}_{\ell=1}^{L}$ and ground truth representatives $\{\mathcal{R}_\ell\}_{\ell=1}^{L}$.

```
Initialize Θ by using a pretrained network;
while (Not Converged) do
    For fixed Θ, compute G_1^ℓ, G_2^ℓ, ... for each dataset ℓ;
    For fixed {G_1^ℓ, G_2^ℓ, ...}_{ℓ=1}^L, update Θ by minimizing the loss function;
end while
```

Output: Optimal parameters Θ.

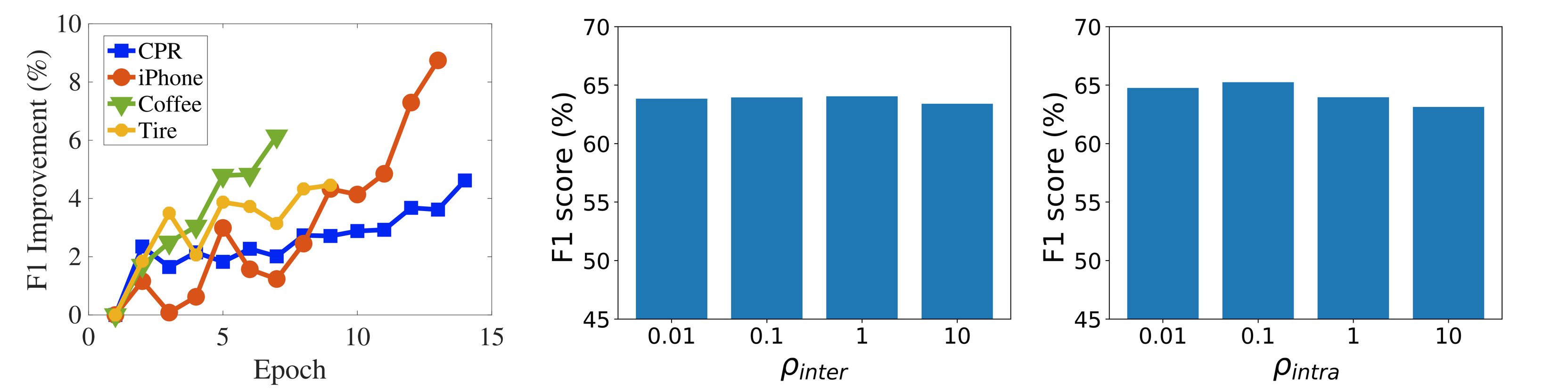
Experiments on Learning Instructions

- **ProceL** [1] (12 tasks, 60 videos/task) and **Breakfast** [2] (10 tasks, 200 videos/task)
  - Measure Precision, Recall and F1 score against ground-truth.

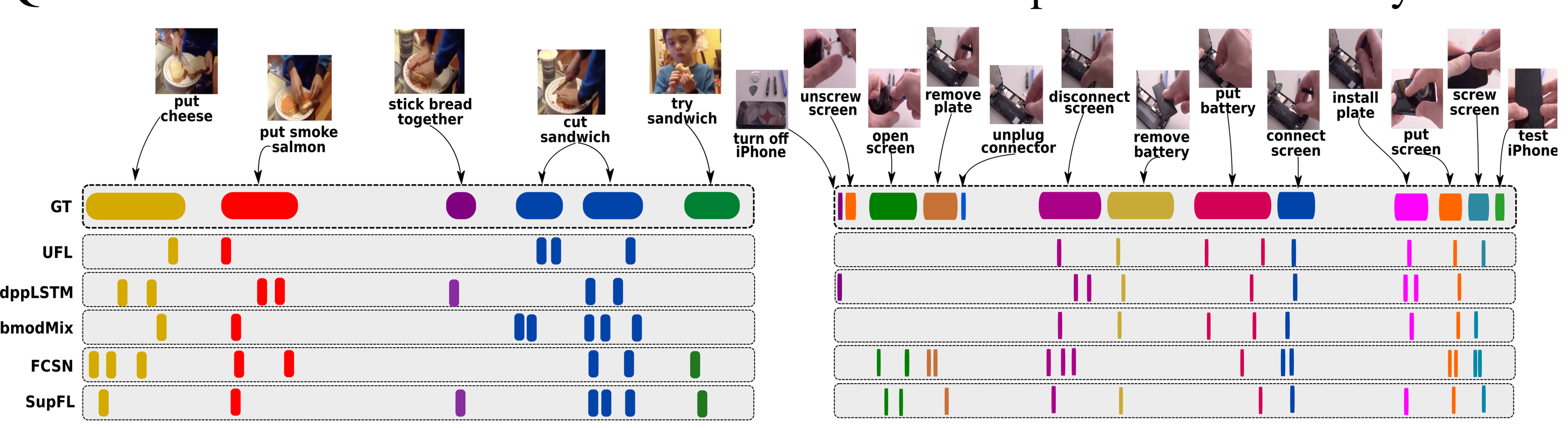
Activity (ProceL)	Uniform	UFL	dppLSTM	SubmodMix	FCSN	SupFL(L)	SupFL(N)
perform CPR	55.7	59.7	53.4	60.0	57.4	63.7	**64.9**
make coffee	57.3	62.6	56.8	62.3	64.2	71.5	**71.6**
jump-start car	57.2	66.0	55.8	67.2	69.6	68.5	**71.4**
repot plant	59.6	67.3	64.7	68.2	69.2	**69.7**	69.1
change tire	54.6	68.4	57.3	65.5	65.7	71.0	**71.2**
tie a tie	44.6	51.6	48.1	53.5	**60.2**	58.5	60.0
setup Chromecast	52.6	61.7	55.5	61.8	56.8	63.7	**66.0**
change iPhone battery	53.0	55.9	53.4	61.2	59.3	62.3	**63.2**
make pbj sandwich	52.7	60.8	53.2	58.0	62.0	**64.9**	64.2
make smoke salmon	59.9	69.4	62.6	71.4	65.3	72.8	**74.3**
change toilet seat	55.5	61.9	56.5	62.7	**68.4**	66.0	67.5
assemble clarinet	57.8	67.2	61.7	66.0	67.8	**72.0**	70.5
Average	55.0	62.7	56.6	63.2	63.8	67.0	**67.8**

Activity (Breakfast)	Uniform	UFL	dppLSTM	SubmodMix	SupFL(L)	SupFL(N)
cereals	58.6	63.8	58.3	64.6	**66.3**	63.4
coffee	73.9	77.7	78.1	79.5	**82.6**	80.5
friedegg	55.2	53.8	**61.2**	53.4	54.9	59.7
juice	61.8	67.9	65.6	67.7	**72.9**	71.9
milk	55.3	63.4	54.9	63.1	**65.8**	63.9
pancake	53.1	53.6	41.0	**54.1**	51.5	53.3
salad	57.5	60.5	59.3	59.4	**64.5**	61.2
sandwich	60.2	65.6	61.7	65.0	**69.1**	67.0
scrambledegg	56.8	61.9	57.9	61.6	**63.6**	59.6
tea	69.2	76.8	72.6	76.1	**78.1**	76.3
Average	60.2	64.5	61.1	64.4	**66.9**	65.7

- Effect of training epochs and hyper-parameters on performance



- Qualitative results on 'make salmon sandwich' and 'replace iPhone battery' tasks



- Ablation studies for SupFL(N) on ProceL

SupFL	Precision	Recall	F1 score
medoid loss	68.1	61.4	61.6
inter-cluster loss	66.2	60.2	59.9
intra-cluster loss	67.2	57.5	59.3
medoid + inter-cluster loss	68.2	60.6	61.2
medoid + intra-cluster loss	68.4	60.4	61.8
inter-cluster + intra-cluster loss	64.7	57.5	58.2
medoid + inter-cluster + intra-cluster loss	**72.8**	**66.3**	**67.8**

[1] E. Elhamifar, Z. Naing, Unsupervised Procedure Learning via Joint Dynamic Summarization, ICCV, 2019.

[2] H. Kuehne, A. Arslan, T. Serre, Language of Actions: Recovering Syntax and Semantics of Goal-Directed Human Activities, CVPR, 2014.