Submodular Span, with Applications to Conditional Data Summarization

Lilly Kumari, Jeff Bilmes
Department of Electrical & Computer Engineering
University of Washington, Seattle
{lkumari, bilmes}@uw.edu

Abstract

As an extension to the matroid span problem, we propose the *submodular span problem* that involves finding a large set of elements with small gain relative to a given query set. We then propose a two-stage *Submodular Span Summarization* (S3) framework to achieve a form of conditional or query-focused data summarization. The first stage encourages the summary to be relevant to a given query set, and the second stage encourages the final summary to be diverse, thus achieving two important necessities for a good query-focused summary. Unlike previous methods, our framework uses only a single submodular function defined over both data and query. We analyze theoretical properties in the context of both matroids and polymatroids that elucidate when our methods should work well. We find that a scalable approximation algorithm to the polymatroid submodular span problem has good theoretical and empirical properties. We provide empirical and qualitative results on three real-world tasks: conditional multi-document summarization on the DUC 2005-2007 datasets, conditional video summarization on the UT-Egocentric dataset, and conditional image corpus summarization on the ImageNet dataset. We use deep neural networks, specifically a BERT model for text, AlexNet for video frames, and Bi-directional Generative Adversarial Networks (BiGAN) for ImageNet images to help instantiate the submodular functions. The result is a minimally supervised form of conditional summarization that matches or improves over the previous state-of-the-art.

1 Introduction

Conditional data summarization involves extracting, from a large dataset, a subset that is both relevant to a given query set and representative of the large dataset. Multiple applications in machine learning and information retrieval are related to this task. For example, in query based extractive Multi-Document Summarization (MDS), given a large collection of text documents, the aim is to produce a short human-readable summary that is not only relevant to the query, but also representative of the information in the full suite of documents. Similarly, query based image summarization aims at retrieving a subset of diverse images which are similar to the query images, given a large image dataset. In fact, general web search can be cast in this framework, where the end result is a succinct summary of the web that is relevant to the user query(s).

Conditional data summarization is related to generic summarization (where there is no query to which relevance is preferred), and is formulated as selecting a representative subset of a large dataset, often based on maximizing a utility function. The utility function captures properties such as informativeness, diversity, and coverage and often satisfies a submodularity property. Submodular functions possess a natural diminishing returns property i.e., the incremental value of a new element is less in a larger than in a smaller context. Mathematically, a set function $f : 2^V \rightarrow \mathbb{R}$ is submodular (Fujishige 2005) if for subsets $S, T \subseteq V$ such that $S \subseteq T$ and $j \notin T$, $f(j|S) \geq f(j|T)$ where $f(j|S) = f(S \cup j) - f(S)$ is the marginal gain of adding j to S. Given a submodular function f, generic summarization can be addressed via cardinality constrained submodular maximization $\max_{A \subseteq V : |A| \leq k} f(A)$, which is solvable with a constant factor i.e., $(1 - 1/e)$ guarantee using the simple greedy algorithm (Nemhauser, Wolsey, and Fisher 1978).

In machine learning, and fields such as natural language processing (NLP) and information retrieval (IR), various approaches have been used to solve this problem. Query-based MDS can be in either supervised where labels are available and a training phase occurs, for example (Lin and Bilmes 2011, 2012) or unsupervised where there are no target labels to train on as in (He et al. 2012; Yao, Wan, and Xiao 2015; Feigenblat et al. 2017). In query-based extractive video summarization, recent methods include snippet selection using sequential and hierarchical Determinantal Point Processes (DPP) (Sharghi, Gong, and Shah 2016; Sharghi, Laurel, and Gong 2017). Although applicable, these methods are supervised and consider the query to be extraneous to the data/corpus. Given the tremendous growth in data and expensive task-based data annotation, there is a pressing need for an unifying conditional data summarization framework that (a) generalizes to different queries i.e., query independent function formulation, (b) supports multiple query summarization i.e., does not limit the query size to one, (c) considers the query to be intrinsic to the data/corpus, and (d) is minimally supervised i.e., uses pre-existing summarization labels only on a limited validation set for hyperparameter tuning.

On this last point, extractive summarization labeling tasks are much harder than standard machine learning labeling and/or annotation tasks — the reason is that a training "set" must be of the form $\mathcal{D} = \{(V_i, A_i)\}_{i=1}^{l}$ where $\forall i$, V_i is the i^{th} dataset and $A_i \subseteq V_i$ is a summary of that dataset. For a human annotator, creating this is extremely difficult; imagine, for example, the task of selecting a size-1000 representative subset from 100,000 images, i.e., where $|V_i| = 100,000$ and $|A_i| = 1000$. Hence, minimal supervision (if any at all) is not only desirable but necessary for the general task of training or tuning big data summarization processes.

In this paper, we develop a new minimally supervised conditional summarization framework based on a method that we call the *submodular span problem*. This method produces a *conditional summary*, i.e., a summary that is relevant to a given query set $Q \subseteq V$. We formulate this as an optimization problem over a submodular function $f : 2^V \to \mathbb{R}$ where the ground set V involves both the query set Q and the data items being summarized $V \setminus Q$. The utility function f is expected to capture the same fundamental properties that a utility function would capture for a generic summary (i.e., diversity, representativeness, etc.). Also the utility function does not need to be reformulated as the query set Q changes.

Our conditional summarization framework, called S3, is formulated as a two-stage submodular optimization problem where the first stage aims to select a large subset that is relevant to the query set. Specifically, we minimize a monotone, non-decreasing conditional submodular function $f(A|Q)$ (representing the conditional redundancy) subject to a cardinality lower-bound constraint. Here $f(A|Q) = f(A \cup Q) - f(Q)$. This first stage retrieves all data points relevant to the query, but that might be redundant, as follows:

$$\text{Stage 1:} \min_{A \subseteq V \setminus Q, |A| \geq k_1} f(A|Q). \tag{1}$$

The second stage is a standard cardinality constrained submodular maximization problem starting from the solution of stage one as follows:

$$\text{Stage 2:} \max_{A \subseteq A_Q^*, |A| \leq k_2} f(A) \tag{2}$$

where A_Q^* is the solution of stage one. This second stage summarizes the redundant output of stage one, and therefore produces a diverse and succinct summary of the data that is still relevant to Q (i.e., stage two filters out the redundancy in A_Q^*). To solve stage two, we use the standard greedy algorithm (Nemhauser, Wolsey, and Fisher 1978).

Our main contributions are summarized as follows:

- We propose a new task called *submodular span problem* which involves finding a large set of data items that is redundant with respect to a query set at hand. We analyze its theoretical properties in context of both matroids and polymatroids.
- Based on the above task, we develop a novel minimally supervised two-stage conditional summarization framework called *submodular span summarization* i.e., S3 framework which produces a query-focused summary. It utilizes a single submodular function for both stages and the utility function does not need to be reformulated as the query set changes.
- We demonstrate that the S3 framework leads to either competitive or state-of-the-art results when applied to three conditional data summarization problems: conditional multi-document, video, and image corpus summarization.

The remainder of the paper is organized as follows. In Section 2, we first discuss related work about existing unsupervised and supervised methods for conditional data summarization of different modalities. In Section 3, we discuss the *submodular span problem*, where less is known. In particular, we offer a scalable approximation algorithm for stage one based on various modular approximations that we motivate via an analysis of a version of the problem applied to matroids, where the problem (as we show) is equivalent to computing the matroid span. In Section 3.2, we generalize this analysis to the submodular case where we show a constant factor approximation for the submodular span problem based on the curvature of f. Finally, we set forth to demonstrate the application of our proposed S3 framework on different conditional summarization tasks in Section 4. We leverage unsupervised representation learning methods such as a pre-trained BERT model (Devlin et al. 2018) for encoding textual data in Section 4.1, AlexNet trained on SentiBank (Chen et al. 2014) dataset for encoding video snippets in Section 4.2, and BiGAN (Donahue, Krähenbühl, and Darrell 2016) for encoding the ImageNet data in Section 4.3.

2 Related Work

Conditional Document Summarization: The majority of existing extractive MDS methods are based on two tasks: query based relevance ranking and sentence saliency score based selection. One of the earlier standard methods is maximum marginal relevance (MMR) (Carbonell and Goldstein 1998) which uses a greedy approach to select the most relevant sentences while considering the trade-off between relevance and redundancy. (McDonald 2007; Gillick and Favre 2009) propose an optimal reformulation to the MMR framework in the form of an integer linear programming problem.

The methods based on data reconstruction, for example DSDR (He et al. 2012) reconstructs each sentence by a non-negative linear combination of summary sentences and then uses sparse coding to select summary sentences that minimize the document reconstruction error. SpOpt (Yao, Wan, and Xiao 2015) adds a sentence dissimilarity term to the objective to maximize diversity. DocRebuild (Ma, Deng, and Yang 2016) further builds upon the DSDR framework using a neural document model. CTSUM (Wan and Zhang 2014) utilizes several hand-crafted features to predict sentence uncertainty scores and then uses them in a graph-based ranking scheme. More recently, deep learning based techniques such as DocEmb (Kobayashi, Noguchi, and Yatsuka 2015) and the vector space model (Kågebäck et al. 2014) utilize the sum of trained word embeddings to represent sentences or documents and formalize the task as maximizing a submodular function defined on the similarity of embeddings. The state-of-the-art unsupervised method called Dual-CES (Roitman et al. 2020) proposes a two-step dual-cascade optimization framework, where both steps utilize the cross-entropy method to handle trade-offs between sentence saliency and focus. Among the supervised methods, the state-of-the-art method SRSum (Ren et al. 2018) uses a deep neural network based model which comprises five sub-models, PriorSum, CSRSum, TSRSum, QSRSum, and SFSum. The individual models encode surface features and latent semantic sentence meaning, and use attention to simulate the context aware reading of a human.

Conditional Video Summarization: Existing methods for this task are supervised in terms of using the summarization labels. (Sharghi, Gong, and Shah 2016; Sharghi, Laurel, and Gong 2017) propose a sequential and hierarchical DPP to model a shot's relevance to the given query and representativeness in the video. In (Jiang and Han 2019), the authors have proposed a Hierarchical Variational Network (HVN) consisting of a query-focused attention module and a multi-level self-attention variational block that captures the multilevel visual content of the scenes and adds to the user-oriented diversity as well. (Xiao et al. 2020) trains a Query-biased Self-Attentive Network (QSAN) which learns the mapping between the visual content and textual captions. It is then augmented with a query-aware scoring MLP to generate a query-focused summary.

Conditional Image Corpus Summarization: This domain is relatively new and the existing work does not meet all requirements of a conditional image summarization system. For example, (Tschiatschek et al. 2014) proposes learning a mixture of submodular functions for generic image collection summarization. (Arandjelovic and Zisserman 2012) focuses on image retrieval given multiple queries of the same object, resulting in improved recall of the system when compared to a single query.

Although the existing methods perform well in their respective domains, there is no simple, effective, and unifying framework for conditional data summarization that requires minimal learning and that can be used irrespective of the data modality. We believe the submodular span approach we present in this work fits this bill.

3 Submodular Span

A given set function $f : 2^V \to \mathbb{R}_+$ is non-negative, monotone, non-decreasing, and submodular if $f(j|A) \geq f(j|B) \geq 0$ for all $A \subseteq B \subseteq V$ and $j \in V$. Such a function is often called a *polymatroid* function (Cunningham 1983). Given a polymatroid function $f : 2^V \to \mathbb{R}_+$, and a query set $Q \subseteq V$ and defining $V_Q = V \setminus Q$, we define the *submodular span problem* as

$$\text{maximize } \{|A| \text{ s.t. } A \subseteq V_Q, f(A|Q) \leq \epsilon\}, \tag{3}$$

where $\epsilon \geq 0$ is small. W.l.o.g., we assume all polymatroid functions are normalized so that not only $f(\emptyset) = 0$ but $f(V) = 1$. Dual to the submodular span problem is Eq. 1. We see that these problems are related, in that they generally ask for large sets A that have low f-valuation when conditioned on the query set Q. We also see that the dual form is cardinality constrained submodular minimization, a problem that is known to have no constant factor approximation algorithm in general (Svitkina and Fleischer 2008), although in the limited curvature case, it is constant-factor approximable (see Theorem 4 analogous to Theorem 5.4 in (Iyer, Jegelka, and Bilmes 2013)).

Submodular span is used as the first step in our conditional summarization strategy, i.e., given a domain V over which a submodular function f is defined, and given a query set $Q \subseteq V$, the objective is to produce a Q-related summary of the remainder V_Q. Submodular span produces a large set A that is related to Q, but to be a good summary, it should also be non-redundant. Hence, given a solution A^*_Q to either Eq. (1) or (3), one can apply standard submodular maximization (via the greedy algorithm), approximately solving Eq. 2. The resulting solution is both related to Q and non-redundant. Conditional summarization uses only one submodular function f defined both on Q and everything else V_Q.

3.1 Matroids, Span, and Redundancy

The reason we call the above the *submodular span problem* is that for a matroid rank function, it is identical to the matroid span. A matroid $\mathcal{M} = (V, \mathcal{I})$ is an algebraic system consisting of a pair $(V, \mathcal{I})$, where V is a ground set, and $\mathcal{I} = \{I_1, I_2, \dots\}$ is a *non-empty* set of *independent* subsets $I_i \subseteq V$ satisfying the two properties: (1) down-closed, if $I \in \mathcal{I}$ then $A \in \mathcal{I}$ for any $A \subseteq I$, and (2) exchangeable, for all $I_1, I_2 \in \mathcal{I}$ with $|I_1| < |I_2|$, then $\exists j \in I_2 \setminus I_1$ such that $I_1 \cup \{j\} \in \mathcal{I}$. The rank function $r_\mathcal{M} : 2^V \to \mathbb{R}$ of a matroid is defined as $r_\mathcal{M}(A) = \max_{I \in \mathcal{I}} |A \cap I|$, i.e., the maximum independent subset of A which is an integer valued unit-increment polymatroid function. The rank function also defines the matroid so we can refer to the matroid simply as $r_\mathcal{M}$. Given $r_\mathcal{M}$ and a query set Q, the span function (Oxley 2011) is defined as:

$$\mathrm{span}_{r_\mathcal{M},0}(Q) = \{v \in V : r_\mathcal{M}(Q \cup \{v\}) = r_\mathcal{M}(Q)\} \tag{4}$$

The subscript "$r_\mathcal{M}, 0$" notation will become apparent below. The span is also called the "closure" of Q, and the span of Q produces a "flat" (or a "subspace") that contains Q. We also define a Q-specific "redundancy" function redn for a matroid as follows:

$$\mathrm{redn}_{r_\mathcal{M},0}(Q) \in \mathrm{argmax}\{|A| : A \subseteq V_Q, r_\mathcal{M}(A|Q) = 0\} \tag{5}$$

We see that Eq. (3) with $f = r_\mathcal{M}$ being a matroid rank function and $\epsilon = 0$ computes $\mathrm{redn}_{r_\mathcal{M},0}(Q)$. By simple inspection, we see that computing $\mathrm{span}_{r_\mathcal{M},0}(Q)$ is much more straightforward via a simple $O(n)$ process than computing $\mathrm{redn}_{r_\mathcal{M},0}(Q)$ which appears to be a form of constrained submodular minimization. Therefore, in the next several sections, we study $\mathrm{span}_{r_\mathcal{M},0}(Q)$ as a surrogate for $\mathrm{redn}_{r_\mathcal{M},0}(Q)$, starting with the case of pure matroids, where there is good news. Specifically:

Lemma 1. *$\mathrm{redn}_{r_\mathcal{M},0}(Q)$ is unique when $r_\mathcal{M}$ is a matroid rank function.*

Theorem 1. *$\mathrm{span}_{r_\mathcal{M},0}(Q) = \mathrm{redn}_{r_\mathcal{M},0}(Q)$ when $r_\mathcal{M}$ is a matroid rank function.*

3.2 Polymatroids, Span, and Redundancy

We can easily generalize span and redn to polymatroids. Given a polymatroid function f, a set Q such that $Q \subseteq V$, and $\epsilon \geq 0$, the ϵ-span function $\operatorname{span}_{f,\epsilon}(Q)$ is defined as:

$$\operatorname{span}_{f,\epsilon}(Q) = \{v \in V_Q : f(v|Q) \leq \epsilon\}. \quad (6)$$

We also define a Q-specific ϵ-redundancy function $\operatorname{redn}_{f,\epsilon}$ for a polymatroid function f as:

$$\operatorname{redn}_{f,\epsilon}(Q) \in \operatorname{argmax}\{|A| : A \subseteq V_Q, f(A|Q) \leq \epsilon\}. \quad (7)$$

We see that $\operatorname{redn}_{f,\epsilon}(Q)$ computes the submodular span defined in Eq. (3), and hence involves constrained submodular minimization. The question we wish to address is the extent to which $\operatorname{span}_{f,\epsilon}(Q)$ can be used as a surrogate function for $\operatorname{redn}_{f,\epsilon}(Q)$. Analysis comparing the above for the cases when $\epsilon = 0$ and $\epsilon > 0$ follows.

Theorem 2. *For a polymatroid* $f : 2^V \rightarrow \mathbb{R}_+$, $\operatorname{span}_{f,0}(Q) = \operatorname{redn}_{f,0}(Q)$.

For $\epsilon = 0$, we observe that computing submodular span and redundancy lead to the same result, analogous to the matroid case. When $\epsilon > 0$, however, this is not the case.

Lemma 2. $\operatorname{redn}_{f,\epsilon}(Q)$ *is not always unique for* $\epsilon > 0$.

Theorem 3. *For a polymatroid* $f : 2^V \rightarrow \mathbb{R}_+$ *such that* $n = |V|$, $\operatorname{redn}_{f,\epsilon}(Q) \subseteq \operatorname{span}_{f,\epsilon}(Q) \subseteq \operatorname{redn}_{f,n\epsilon}(Q)$ *when* $\epsilon \geq 0$.

For a given ϵ, since $\operatorname{span}_{f,\epsilon}(Q)$ covers all the elements of $\operatorname{redn}_{f,\epsilon}(Q)$, we can compute the $\operatorname{span}_{f,\epsilon}(Q)$ as a surrogate function for $\operatorname{redn}_{f,\epsilon}(Q)$ and then summarize it. But first we ask if for some value of $\epsilon' \leq \epsilon$, their f-valuations are equal. Unfortunately, this is also not the case.

Lemma 3. *There does not, in general, exist an* $\epsilon' \leq \epsilon$ *such that* $f(\operatorname{span}_{f,\epsilon'}(Q)) = f(\operatorname{redn}_{f,\epsilon}(Q))$ *for all* Q *and* $\epsilon > 0$.

Since there does not exist an $\epsilon' \leq \epsilon$ for which $f(\operatorname{span}_{f,\epsilon'}(Q)) = f(\operatorname{redn}_{f,\epsilon}(Q))$ when $\epsilon > 0$, we can form an upper bound on $f(\operatorname{span}_{f,\epsilon}(Q))$ as follows.

Lemma 4. $f(\operatorname{span}_{f,\epsilon}(Q)|Q) \leq (k_s - k_r + 1)\epsilon$ *where* $k_s = |\operatorname{span}_{f,\epsilon}(Q)|$ *and* $k_r = |\operatorname{redn}_{f,\epsilon}(Q)|$

Lemma 4 shows that for our surrogate function $\operatorname{span}_{f,\epsilon}(Q)$, the worst case bound on its f-valuation with respect to Q could be $n\epsilon$ where $n = |V_Q|$. This is most likely when the ground set V contains many elements that are redundant to Q but that are mostly mutually non-redundant.

Lemma 5. *With the conditional submodular curvature with respect to* Q *defined as*

$$\kappa_{f_Q}(A) \triangleq 1 - \min_{a \in A} \frac{f((a|(A \backslash a)), Q)}{f(a|Q)}, \quad (8)$$

$f(\operatorname{span}_{f,\frac{\epsilon}{n}}(Q)|Q) \leq \epsilon - \frac{\epsilon}{n}(k_r - k_s)(1 - \kappa_{f_Q}(\operatorname{redn}_{f,\epsilon}(Q)))$ *where* $k_s = |\operatorname{span}_{f,\frac{\epsilon}{n}}(Q)|$ *and* $k_r = |\operatorname{redn}_{f,\epsilon}(Q)|$.

To solve Eq. (1) or (3), a modular approximation of $f(A|Q)$ i.e., $m_Q(A) = \sum_{a \in A} f(a|Q)$ can be optimized. The Majorization-Minimization algorithm based on submodular semi-differentials, as proposed in (Iyer, Jegelka, and Bilmes 2013) can also be used for the constrained submodular minimization. The approximation factor for these algorithms is expressed in terms of conditional submodular curvature with respect to Q as proved in Theorem 4 and has a worst-case upper bound of $\mathcal{O}(n)$ where $n = |V \setminus Q|$

Theorem 4. *Let* A^* *be the optimal solution to Eq.* (1)*, then* A *returned by the modular approximation of* $f(A|Q)$ *such that* $A = \text{argmin}_{A \subseteq V_Q, |A| \geq k} m_Q(A)$ *satisfies:*

$$f(A|Q) \leq \frac{|A^*|}{1 + (|A^*| - 1)(1 - \kappa_{f_Q}(A^*))} f(A^*|Q)$$

Proof. (Iyer, Jegelka, and Bilmes 2013) show a bound for submodular function minimization in terms of generic submodular curvature and our proof follows theirs.

4 Experiments

In this section, using optimization procedures based on the analysis given in Section 3.2, we evaluate the S3 framework on three conditional summarization tasks: (1) conditional multi-document summarization (2) conditional video summarization (3) conditional image corpus summarization.

4.1 Conditional Multi-Document Summarization

Dataset: We use DUC 2005-2007 datasets which are the benchmark datasets for query-focused MDS, made available by the Document Understanding Conference [1]. DUC 2005-2006 and DUC 2007 contain 50 and 45 document clusters respectively, with each cluster containing 25 news articles (32 in case of DUC 2005) related to the same topic, and the task is to generate a query-focused summary of at most 250 words for each document cluster. As a pre-processing step, we remove special characters from the sentences and we augment the query set for each document cluster with its topic as well as concatenate each query sentence with the cluster topic.

Feature Representation: In order to obtain sentence representations, we use the English uncased variant of the BERT-base model (Devlin et al. 2018) and fine-tune it for the Rouge-2 recall score prediction task using two years of DUC 2005-2007 as the training set. For example, we fine-tune the network on the DUC 2005-2006 datasets in order to extract fixed-size sentence representations for DUC 2007 (which is the test set in this example). We do not use any oracle summarization labels for the test set. In addition to using fine-tuned BERT models, we also try a minimally supervised approach where we use the pre-trained BERT model for computing sentence representations.

Since BERT's encoder has 12 transformer layers, each of which outputs contextualized WordPiece representations, the most transferable layer l (Ethayarajh 2019) for the MDS task is a hyperparameter which is tuned on the development set. Given l, we take a smoothed inverse frequency (SIF) based weighted average of hidden activations of each wordpiece (Peters, Ruder, and Smith 2019; Arora, Liang, and Ma 2017) from layer l to construct 768-dimensional sentence embeddings v_{s_i} for the sentences s_i in the test set i.e., $v_{s_i} = \frac{1}{|s_i|} \sum_{w \in s_i} \frac{a}{a + p(w)} h_l(w)$. Here, $h_l(w)$ is the hidden layer representation of wordpiece w corresponding to layer l, $p(w)$ is probability of wordpiece w estimated from the entire DUC corpus, and a is a weighting parameter fixed at 10^{-3}.

Summary Generation: We use facility location (Mirchandani and Francis 1990) as the objective function for stage one

[1] https://duc.nist.gov

and two of the S3 framework. The facility location function is defined as $f(X) = \sum_{s_i \in V} \max_{s_j \in X} \mathrm{sim}(s_i, s_j)$ where $\mathrm{sim}(s_i, s_j)$ is the similarity between sentence embeddings v_{s_i} and v_{s_j} of sentences s_i and s_j. We compute the similarity matrix using a Gaussian kernel of width σ which is tuned on the development set in each case.

Stage one of the S3 framework caters to finding relevant sentences A_Q from a document set which answer given queries. In order to filter irrelevant noisy sentences which are either too small or too long, we prune the candidate set by considering sentences whose length ranges between 11 and 80 and are a subset of the top 30% nearest neighbors set of the query sentences. Once we obtain the relevant answers (A_Q) using the majorization-minimization (MMin) (Iyer, Jegelka, and Bilmes 2013; Iyer and Bilmes 2013) algorithm for solving Eq. (3), stage two removes the redundant answers and produces a succinct relevant summary for that document set via constrained submodular maximization using the greedy algorithm (Nemhauser, Wolsey, and Fisher 1978). The algorithm at iteration i selects the sentence s_i such that $s_i = \mathrm{argmax}_{s \in A_Q} \frac{f(A_{i-1} \cup s) - f(A_{i-1})}{(c(s))^r}$ if $c(A_{i-1} \cup s_i) \leq \mathcal{B}$. $c(s)$ denotes the sentence length, $r > 0$ is the scaling factor, and $\mathcal{B}$ represents the overall budget which is 250 words for DUC 2005-2007. For DUC-2005, we use DUC-2006 to tune the hyperparameters which include $\{l, \sigma, \epsilon, r\}$. Similarly, for DUC-2006 and DUC-2007, we use DUC-2005 as the development set.

Evaluation: We use the ROUGE toolkit (Lin 2004)[2] which assesses the summary quality by counting the overlapping units such as n-grams, word sequences, and word-pairs between the candidate summary and the reference summaries. We report recall and F-measure corresponding to Rouge-1, Rouge-2 , and Rouge-SU4.

Since our approach requires minimal learning of hyperparameters, we compare against other state-of-the-art unsupervised and supervised approaches. In addition to existing supervised methods, we also design another strong supervised baseline method called *MixModSub* which utilizes a submodular function $f' : 2^{V'} \rightarrow \mathbb{R}_+$, i.e., the query set Q is extrinsic to the submodular function f' i.e., $V' \cap Q = \emptyset$ and it only considers V' which contains sentences s_i where $i \in \{1, 2, \ldots |V'|\}$ that are to be summarized. Here, f' is a facility location function defined using the fine-tuned BERT-based feature vectors v_{s_i} for each $s_i \in V'$. We define a relevance based modular function $m_Q : 2^{V'} \rightarrow \mathbb{R}_+$ where for any $A \subseteq V', m_Q(A) = \sum_{s_i \in A} m_Q(s_i)$. Since m captures the relevance of each sentence s_i to the query set Q, $m_Q(s_i) = \frac{1}{|Q|} \sum_{s_q \in Q} \mathrm{sim}(s_q, s_i)$. Finally, we define a submodular function $g : 2^{V'} \rightarrow \mathbb{R}_+$ as $g(A) = \lambda f'(A) + (1 - \lambda) m_Q(A)$ which is a convex mixture of submodular f' and modular m_Q. We use a Gaussian kernel of width σ to define the similarity matrix and perform budget constrained submodular maximization given budget $\mathcal{B}$. Similar to the previous experiments, we tune the hyperparameters $\{\sigma, r, \lambda\}$ on another year of DUC as the development set.

Table 1 shows the average recall and F-measure with respect to Rouge-1 (R1), Rouge-2 (R2) and Rouge-SU4 (RSU4) scores on DUC 2005-2007 datasets against different methods. The performance of S3 framework is competitive with the current unsupervised state-of-the-art method, Dual-CES (Roitman et al. 2020) on each of the Rouge-1 , Rouge-2, and Rouge-SU4 F-measure.

4.2 Conditional Video Summarization

Dataset: We use the query-focused video summarization dataset from (Sharghi, Laurel, and Gong 2017) which is compiled using the UT Egocentric dataset (Lee, Ghosh, and Grauman 2012). The UTE dataset consists of four daily life egocentric videos of 3-5 hours duration. Based on the overlap of the video-shot captions (Yeung, Fathi, and Fei-Fei 2014) with SentiBank (Borth et al. 2013), a lexicon of 48 concepts such as street, tree, phone etc. is constructed that denotes the basis for encoding the semantic information in each video shot. For each video, there are 46 different sets of queries, with each query set covering two or three concepts. We use the oracle summaries released by (Sharghi, Laurel, and Gong 2017) and follow their video summarization evaluation strategy based on the user-annotated semantic vectors of the video shots.

Feature Representation: We uniformly partition each video into five seconds long shots. For each frame belonging to a shot, we construct a 2089-dimensional feature vector using an off-the-shelf deep model called DeepSentiBank (Chen et al. 2014) for fair comparison to existing methods. The network has an architecture similar to the AlexNet (Krizhevsky, Sutskever, and Hinton 2012) which was pre-trained on the ImageNet classification task and for the SentiBank classification task, the last fully connected layer is replaced to produce a softmax distribution across 2089 class labels. Then for each shot, we average its frame-level feature representations to obtain a shot-level feature representation. The SentiBank classes consist of ANP (Adjective Noun Pairs), for instance *beautiful sky, clear sky, sunny sky* are different ANPs corresponding to the concept *sky*. We max-pool their shot-level detection scores to get one detection score for each concept belonging to the lexicon consisting of 48 concepts. This results into a 48-dimensional feature representation for each video-shot with detection scores ranging between 0 and 1.

Summary Generation: Similar to section 4.1, we use facility location as the objective function, but here $\mathrm{sim}(s_i, s_j)$ is the similarity between the DeepSentiBank-based shot-level features for shots s_i and s_j. We compute the similarity matrix using cosine similarity after carefully validating different similarity measures on a development set in terms of F1 score performance. For Video-1, we use Video-3 to tune the hyperparameters which include $\{k_1, k_2\}$; k_1 and k_2 are the cardinality constraints for optimizing stage one and stage two respectively. For Video 2-4, we use Video-1 as the development set.

Evaluation: Similar to (Sharghi, Laurel, and Gong 2017), we use the user-annotated semantic vectors of video shots to quantify the semantic similarity between the oracle summary's shots and our system generated summary's shots. A maximum weight based bipartite graph matching between them enables us to compute precision, recall, and F1 score between the matched pairs. Similar to document

[2]ROUGE version 1.5.5 used with option -n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -d -l 250

	System	R1-R	R1-F	R2-R	R2-F	RSU4-R	RSU4-F
DUC 2005	MixModSub*	38.64	38.17	7.74	7.65	13.65	13.49
	SRSum (Ren et al. 2018)*	39.83	-	**8.57**	-	-	-
	Dual-CES (Roitman et al. 2020)	**40.82**	38.08	8.07	7.54	**14.13**	13.17
	S3 (Ours)*	39.11	**38.66**	7.87	**7.79**	13.80	**13.65**
	S3 (Ours)	38.64	38.20	7.60	7.52	13.52	13.37
DUC 2006	MixModSub*	39.80	39.57	8.62	8.58	14.40	14.32
	DSDR (He et al. 2012)	-	33.17	-	6.05	-	-
	SpOpt (Yao, Wan, and Xiao 2015)	39.96	-	8.68	-	14.23	-
	DocRebuild (Ma, Deng, and Yang 2016)	-	40.86	-	8.48	-	14.45
	SRSum (Ren et al. 2018)*	42.82	-	**10.46**	-	-	-
	Dual-CES (Roitman et al. 2020)	**43.94**	41.23	10.09	**9.47**	**15.96**	14.97
	S3 (Ours)*	41.62	**41.42**	9.48	9.43	15.10	**15.02**
	S3 (Ours)	41.13	40.95	9.24	9.20	14.85	14.79
DUC 2007	MixModSub*	40.87	40.42	10.26	10.15	15.66	15.49
	DSDR (He et al. 2012)	-	39.57	-	7.44	-	-
	CTSUM (Wan and Zhang 2014)	43.10	42.66	10.93	10.82	16.32	16.16
	SpOpt (Yao, Wan, and Xiao 2015)	42.36	-	11.11	-	16.47	-
	DocRebuild (Ma, Deng, and Yang 2016)	-	42.73	-	10.31	-	15.81
	SRSum (Ren et al. 2018)*	45.01	-	**12.80**	-	-	-
	Dual-CES (Roitman et al. 2020)	**46.02**	**43.24**	12.53	**11.78**	**17.91**	**16.83**
	S3 (Ours)*	43.42	42.95	11.24	11.12	16.70	16.52
	S3 (Ours)	42.50	42.32	11.12	11.07	16.35	16.28

Table 1: ROUGE results on DUC 2005, 2006, and 2007 in terms of Recall and F-measure. Methods marked with $(^*)$ are supervised in terms of using oracle summarization labels for training or model fine-tuning.

summarization, we also compare against the designed minimally supervised baseline *MixModSub* (not fully supervised as we are not fine-tuning any model using oracle summary labels). We tune relevant hyperparameters $\{k, \lambda\}$ on another video as the development set.

Table 2 shows the performance of the S3 framework against other supervised state-of-the-art methods and our designed baseline *MixModSub*, in terms of precision, recall, and F1 score. In terms of recall and F1 score on Video 1 and 3, our S3 framework (requiring minimal learning) outperforms the current supervised state-of-the-art methods which use the oracle summary labels for learning different deep neural networks; on average, it outperforms the previous supervised methods in terms of F1 score and achieves comparable results in terms of precision and recall.

4.3 Conditional Image Corpus Summarization

Dataset: ImageNet-1k (Deng et al. 2009) is a large scale image database which contains nearly 1.28 million training images and 50,000 validation images. The dataset is organized according to the WordNet (Miller 1995) hierarchy, with each node depicting hundreds and thousands of images. In our experiments, we randomly sample 1,000 images, one from each class, from the dataset to form the development query set and the remaining training images are used as the gallery set that needs to be summarized. We use this set for hyperparameter tuning, and show qualitative results corresponding to test query images having no overlap with the development query set.

Feature Representation: We use a Bidirectional Generative Adversarial Network (BiGAN) (Donahue, Krähenbühl, and Darrell 2016) for unsupervised learning of feature representations for the ImageNet database. We use the pre-trained encoder weights from (Donahue, Krähenbühl, and Darrell 2016) learned in an unsupervised fashion to encode the images into a 1024-dimensional latent feature representation. In order to reduce problems arising from the curse of dimensionality, after examining the histogram of pairwise similarities, we use PCA to reduce the features dimensions to 512.

Summary Generation: We use a sparse facility location as the objective function for our S3 framework. In order to deal with this large-scale dataset, we use *faiss*[3], which is an efficient similarity search library from Facebook, to build a k-NN similarity graph using cosine similarity. For stage one of the S3 framework, we prune the candidate set by taking top-k nearest neighbors of each $q \in Q$ and optimize Eq. (1) using modular approximation. In all experiments, k is set to 1000. In stage two, we condense A_Q to generate a conditional summary of 25 images.

Evaluation: To assess the quality of the query-focused summary, we propose a function, $\mathcal{R}(A|Q)$ that captures the relevance of the summary images to the query set.

$$\mathcal{R}(A|Q) = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{|A|} \sum_{a \in A} \mathbb{1}_{y_q = y_a} \tag{9}$$

Here, y_a denotes the label/class of the image a. In case of single query image based summarization, $\mathcal{R}(A|Q)$ is simply the precision of retrieving the query class in the summary. We also assess the diversity of the summary using $f(A)$ which is the sparse-facility location function valuation of summary A with respect to the ground set V.

In Table 3, for different query sets, we show the conditional summaries after fitting the two-dimensional tSNE

[3]https://ai.facebook.com/tools/faiss/

	System	Precision	Recall	F1 Score
Video-1	Mem-SeqDPP (2017)	49.86	53.38	48.68
	HVN (2019)	52.55	52.91	51.45
	DSAN (2020)	48.41	52.34	48.52
	MixModSub	**53.26**	51.77	51.20
	S3 (Ours)	51.86	**55.24**	**52.18**
Video-2	Mem-SeqDPP (2017)	33.71	62.09	41.66
	HVN (2019)	38.66	62.70	**47.49**
	DSAN (2020)	**46.51**	51.36	46.64
	MixModSub	36.29	62.37	45.56
	S3 (Ours)	37.24	**63.88**	46.71
Video-3	Mem-SeqDPP (2017)	55.16	62.40	56.47
	HVN (2019)	60.28	62.58	61.08
	DSAN (2020)	56.78	61.14	56.93
	MixModSub	58.35	63.24	60.36
	S3 (Ours)	**60.51**	**65.72**	**62.66**
Video-4	Mem-SeqDPP (2017)	21.39	**63.12**	29.96
	HVN (2019)	26.79	54.21	**35.47**
	DSAN (2020)	**30.54**	46.90	34.25
	MixModSub	17.78	53.51	26.44
	S3 (Ours)	26.54	52.94	34.97
Avg	Mem-SeqDPP (2017)	40.03	**60.25**	44.19
	HVN (2019)	44.57	58.10	48.87
	DSAN (2020)	**45.56**	52.94	46.59
	MixModSub	41.42	57.72	45.89
	S3 (Ours)	44.04	59.44	**49.13**

Table 2: Results on the UTE dataset for conditional video summarization. The cited methods given in the first row corresponding to each video are supervised, in terms of using the oracle summarization labels for model training.

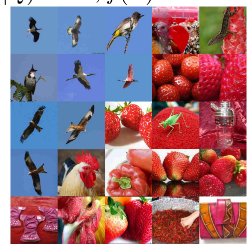
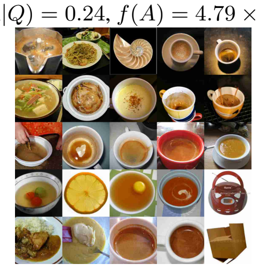
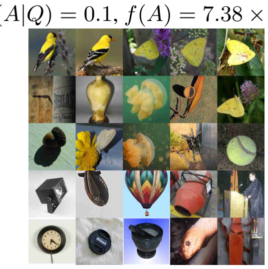
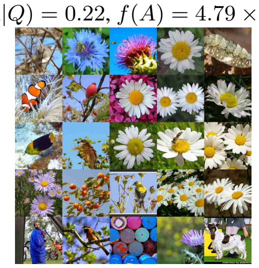
Query	**Submodular Span Summary**
	$\mathcal{R}(A\vert Q) = 0.2$, $f(A) = 4.07 \times 10^{-3}$ 
	$\mathcal{R}(A\vert Q) = 0.24$, $f(A) = 4.79 \times 10^{-3}$ 
	$\mathcal{R}(A\vert Q) = 0.1$, $f(A) = 7.38 \times 10^{-3}$ 
	$\mathcal{R}(A\vert Q) = 0.22$, $f(A) = 4.79 \times 10^{-3}$ 

Table 3: Qualitative Results on the ImageNet Dataset corresponding to $|Q| > 1$

vectors of their image embeddings onto a square grid using the Jonker-Volgenant algorithm (Jonker and Volgenant 1987). As it can be seen, apart from consisting of diverse images belonging to the query classes, the conditional summary also consists of images which share properties of both query classes. For example, given a query set comprising $\{$*strawberry, kite*$\}$ images, the conditional summary also consists of bird images having red hues and a lizard (sharing *kite* hues) on a red flower. In the last example where the queries comprise $\{$*bee, daisy*$\}$, the summary consists of two images of a *bee* on *daisy* as well as birds resting on twigs.

5 Conclusion

In this paper, we have proposed and studied the *submodular span problem* as an extension to the matroid span problem in terms of retrieving elements relevant to a query set. We have designed a minimally supervised two-stage query-focused summarization framework called S3 and showed its applications for conditional data summarization of different data modalities. Our analysis and results also shed light on the feasibility and scalability of the modular approximation algorithm for the polymatroid submodular span problem. Our results on three real-world datasets, DUC 2005-2007, UT-Egocentric video dataset, and ImageNet verify the significance of the two stages (retrieval followed by summarization) of the S3 framework.

6 Acknowledgments

This work was supported in part by the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. We would also like to thank the Melodi lab students: Tianyi Zhou, Shengjie Wang, and Chandrashekhar Lavania for their useful discussions.

References

Arandjelovic, R.; and Zisserman, A. 2012. Multiple queries for large scale specific object retrieval. In *BMVC*, 1–11.

Arora, S.; Liang, Y.; and Ma, T. 2017. A simple but tough-to-beat baseline for sentence embeddings. *5th International*

Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings .

Borth, D.; Chen, T.; Ji, R.; and Chang, S.-F. 2013. Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia*, 459–460.

Carbonell, J.; and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 335–336.

Chen, T.; Borth, D.; Darrell, T.; and Chang, S.-F. 2014. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586* .

Cunningham, W. H. 1983. Decomposition of submodular functions. *Combinatorica* 3(1): 53–68.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning. *arXiv preprint arXiv:1605.09782* .

Ethayarajh, K. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *arXiv preprint arXiv:1909.00512* .

Feigenblat, G.; Roitman, H.; Boni, O.; and Konopnicki, D. 2017. Unsupervised query-focused multi-document summarization using the cross entropy method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 961–964.

Fujishige, S. 2005. *Submodular functions and optimization*. Elsevier.

Gillick, D.; and Favre, B. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, 10–18.

He, Z.; Chen, C.; Bu, J.; Wang, C.; Zhang, L.; Cai, D.; and He, X. 2012. Document summarization based on data reconstruction. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Iyer, R.; and Bilmes, J. 2013. Submodular Optimization with Submodular Cover and Submodular Knapsack Constraints. In *Neural Information Processing Society (NeurIPS, formerly NIPS)*. Lake Tahoe, CA.

Iyer, R.; Jegelka, S.; and Bilmes, J. 2013. Fast semidifferential-based submodular function optimization: Extended version. In *ICML*.

Jiang, P.; and Han, Y. 2019. Hierarchical Variational Network for User-Diversified & Query-Focused Video Summarization. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, 202–206.

Jonker, R.; and Volgenant, A. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* 38(4): 325–340.

Kågebäck, M.; Mogren, O.; Tahmasebi, N.; and Dubhashi, D. 2014. Extractive summarization using continuous vector space models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, 31–39.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980.

Kobayashi, H.; Noguchi, M.; and Yatsuka, T. 2015. Summarization based on embedding distributions. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1984–1989.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Lee, Y. J.; Ghosh, J.; and Grauman, K. 2012. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, 1346–1353. IEEE.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Lin, H.; and Bilmes, J. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 510–520. Association for Computational Linguistics.

Lin, H.; and Bilmes, J. A. 2012. Learning mixtures of submodular shells with application to document summarization. *arXiv preprint arXiv:1210.4871* .

Ma, S.; Deng, Z.-H.; and Yang, Y. 2016. An unsupervised multi-document summarization framework based on neural document model. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1514–1523.

McDonald, R. 2007. A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, 557–564. Springer.

Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11): 39–41.

Mirchandani, P. B.; and Francis, R. L. 1990. *Discrete location theory*. Wiley.

Nemhauser, G. L.; Wolsey, L. A.; and Fisher, M. L. 1978. An analysis of approximations for maximizing submodular set functions—I. *Mathematical programming* 14(1): 265–294.

Oxley, J. 2011. *Matroid Theory: Second Edition*. Oxford University Press.

Peters, M.; Ruder, S.; and Smith, N. A. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987* .

Ren, P.; Chen, Z.; Ren, Z.; Wei, F.; Nie, L.; Ma, J.; and De Rijke, M. 2018. Sentence relations for extractive summarization with deep neural networks. *ACM Transactions on Information Systems (TOIS)* 36(4): 1–32.

Roitman, H.; Feigenblat, G.; Cohen, D.; Boni, O.; and Konopnicki, D. 2020. Unsupervised Dual-Cascade Learning with Pseudo-Feedback Distillation for Query-Focused Extractive Summarization. In *Proceedings of The Web Conference 2020*, 2577–2584.

Sharghi, A.; Gong, B.; and Shah, M. 2016. Query-focused extractive video summarization. In *European Conference on Computer Vision*, 3–19. Springer.

Sharghi, A.; Laurel, J. S.; and Gong, B. 2017. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4788–4797.

Svitkina, Z.; and Fleischer, L. 2008. Submodular approximation: Sampling-based algorithms and lower bounds. In *FOCS*, 697–706.

Tschiatschek, S.; Iyer, R. K.; Wei, H.; and Bilmes, J. A. 2014. Learning mixtures of submodular functions for image collection summarization. In *Advances in neural information processing systems*, 1413–1421.

Wan, X.; and Zhang, J. 2014. CTSUM: extracting more certain summaries for news articles. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 787–796.

Xiao, S.; Zhao, Z.; Zhang, Z.; Guan, Z.; and Cai, D. 2020. Query-Biased Self-Attentive Network for Query-Focused Video Summarization. *IEEE Transactions on Image Processing* 29: 5889–5899.

Yao, J.-g.; Wan, X.; and Xiao, J. 2015. Compressive document summarization via sparse optimization. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yeung, S.; Fathi, A.; and Fei-Fei, L. 2014. Videoset: Video summary evaluation through text. *arXiv preprint arXiv:1406.5824* .

Zhu, H.; Dong, L.; Wei, F.; Qin, B.; and Liu, T. 2019. Transforming Wikipedia into Augmented Data for Query-Focused Summarization. *arXiv preprint arXiv:1911.03324* .

A Proof of Results in Section 3.1

Lemma 1. $\mathrm{redn}_{r_{\mathcal{M}},0}(Q)$ *is unique when* $r_{\mathcal{M}}$ *is a matroid rank function.*

Proof. Let $A_1, A_2 \in \{A \subseteq V \setminus Q : r_{\mathcal{M}}(A|Q) = 0\}$. We prove that $A_1 \cap A_2 \in \{A \subseteq V \setminus Q : r_{\mathcal{M}}(A|Q) = 0\}$ and $A_1 \cup A_2 \in \{A \subseteq V \setminus Q : r_{\mathcal{M}}(A|Q) = 0\}$.

Take any subset $B \subseteq A$, then $0 = r_{\mathcal{M}}(A|Q) \geq r_{\mathcal{M}}(B|Q) \geq 0$. Hence $r_{\mathcal{M}}(B|Q) = 0$. Since $A_1 \cap A_2 \subseteq A$, then $r_{\mathcal{M}}(A_1 \cap A_2|Q) = 0$.

Next, by submodularity we have:

$$0 = r_{\mathcal{M}}(A_1|Q) + r_{\mathcal{M}}(A_2|Q) \geq \tag{10}$$

$$r_{\mathcal{M}}(A_1 \cup A_2|Q) + r_{\mathcal{M}}(A_1 \cap A_2|Q) \geq 0 \tag{11}$$

Hence $r_{\mathcal{M}}(A_1 \cup A_2|Q) = 0$. The uniqueness of redn follows by considering the following equivalent definition:

$$\mathrm{redn}_{r_{\mathcal{M}},0}(Q) = \bigcup_{A' \in \{A \subseteq V \setminus Q | r_{\mathcal{M}}(A|Q)=0\}} A' \tag{12}$$

Theorem 1. $\mathrm{span}_{r_{\mathcal{M}},0}(Q) = \mathrm{redn}_{r_{\mathcal{M}},0}(Q)$ *when* $r_{\mathcal{M}}$ *is a matroid rank function.*

Proof. We first show that $\mathrm{span}_{r_{\mathcal{M}},0}(Q) \subseteq \mathrm{redn}_{r_{\mathcal{M}},0}(Q)$.

We prove this by induction on $i \in [k]$. Order elements $\{v_1, v_2, \ldots, v_k\} = \mathrm{span}_{r_{\mathcal{M}},0}(Q) \setminus Q$. Then $r_{\mathcal{M}}(v_i|Q) = 0$ for all $i \in [k]$. The base case is as follows: from submodularity, we have $r_{\mathcal{M}}(Q \cup \{v_1\}) + r_{\mathcal{M}}(Q \cup \{v_2\}) \geq r_{\mathcal{M}}(Q) + r_{\mathcal{M}}(Q \cup \{v_1, v_2\})$ and hence $r_{\mathcal{M}}(Q \cup \{v_1\}) \geq r_{\mathcal{M}}(Q \cup \{v_1, v_2\})$ but by monotonicity, $r_{\mathcal{M}}(Q \cup \{v_1\}) \leq r_{\mathcal{M}}(Q \cup \{v_1, v_2\})$. Hence, $r_{\mathcal{M}}(Q) = r_{\mathcal{M}}(Q \cup \{v_1\}) = r_{\mathcal{M}}(Q \cup \{v_2\}) = r_{\mathcal{M}}(Q \cup \{v_1, v_2\})$. Inductively, assume that $r_{\mathcal{M}}(Q) = r_{\mathcal{M}}(Q \cup \{v_1\}) = r_{\mathcal{M}}(Q \cup \{v_1, v_2, \ldots, v_{i-1}\})$.

Then by submodularity,

$$r_{\mathcal{M}}(Q \cup \{v_1, v_2, \ldots, v_{i-1}\}) + r_{\mathcal{M}}(Q \cup \{v_i\}) \geq \tag{13}$$

$$r_{\mathcal{M}}(Q) + r_{\mathcal{M}}(Q \cup \{v_1, v_2, \ldots, v_i\}) \tag{14}$$

which implies $r_{\mathcal{M}}(Q \cup \{v_i\}) \geq r_{\mathcal{M}}(Q \cup \{v_1, v_2, \ldots, v_i\})$, and along with monotonicity we get $r_{\mathcal{M}}(Q) = r_{\mathcal{M}}(Q \cup \{v_i\}) = r_{\mathcal{M}}(Q \cup \{v_1, v_2, \ldots, v_i\})$. Taking $i = k$, this implies that $r_{\mathcal{M}}(A|Q) = 0$ where $A = \mathrm{span}_{r_{\mathcal{M}},0}(Q) \setminus Q$. Since $\mathrm{redn}_{r_{\mathcal{M}},0}(Q)$ is the largest set having this property, and is unique, the result follows.

Next we show that $\mathrm{redn}_{r_{\mathcal{M}},0}(Q) \subseteq \mathrm{span}_{r_{\mathcal{M}},0}(Q)$. Let $A = \mathrm{redn}_{r_{\mathcal{M}},0}(Q)$ and let $A' \subseteq A$. Then,

$$0 = r_{\mathcal{M}}(A|Q) \geq r_{\mathcal{M}}(A'|Q) \geq 0 \tag{15}$$

This implies $r_{\mathcal{M}}(A'|Q) = 0$. Take, in turn, $A' = \{a\}$ for all $a \in A$ giving $r_{\mathcal{M}}(a|Q) = 0$.

Thus, from these results, we conclude that $\mathrm{span}_{r_{\mathcal{M}},0}(Q) \subseteq \mathrm{redn}_{r_{\mathcal{M}},0}(Q)$.

B Proof of Results in Section 3.2

Theorem 2. *For a polymatroid* $f : 2^V \rightarrow \mathbb{R}_+$, $\mathrm{span}_{f,0}(Q) = \mathrm{redn}_{f,0}(Q)$.

Proof. The proof is identical to the proof of Theorem 1 since the uniqueness of $\mathrm{redn}_{f,0}$ still holds by the proof of Lemma 1, and nothing in that proof assumed anything more than polymatroidality (i.e., monotone non-decreasing, normalized, and submodular) and uniqueness of $\mathrm{redn}_{f,0}$.

Lemma 2. $\mathrm{redn}_{f,\epsilon}(Q)$ *is not always unique for* $\epsilon > 0$.

Proof. For the uniqueness of $\mathrm{redn}_{f,\epsilon}(Q)$, we still have that $\epsilon \geq f(A|Q) \geq f(A'|Q) \geq 0$, so that if $A, A' \in \mathrm{redn}_{f,\epsilon}(Q)$, then $A \cap A' \in \mathrm{redn}_\epsilon(Q)$. However, the union case has

$$2\epsilon \geq f(A|Q) + f(A'|Q) \geq \tag{16}$$

$$f(A \cup A'|Q) + f(A \cap A'|Q) \geq \epsilon \tag{17}$$

We also, by monotonicity, have that $f(A \cap A'|Q) \leq \min(f(A|Q), f(A'|Q)) \leq \max(f(A|Q), f(A'|Q)) \leq f(A \cup A'|Q)$. From this, if it was the case that $\max(f(A|Q), f(A'|Q)) \leq f(A \cap A'|Q)$ then we could get that $f(A \cup A'|Q) \leq \epsilon$. But since this is not necessarily the case, the only bound we get is $f(A \cup A'|Q) \leq 2\epsilon$. Hence, $\mathrm{redn}_\epsilon(Q)$ might not be unique.

As an example, take $f(A) = \sqrt{|A|}$. Then take any two non-intersecting same sized sets B, B' that don't overlap with Q. Then $f(B|Q) = f(B'|Q) = \sqrt{|B \cup Q|} - \sqrt{|Q|}$, and if we take $\epsilon = \sqrt{|B \cup Q|} - \sqrt{|Q|}$ we have two non-unique maximal sets that could serve as $\mathrm{redn}_{f,\epsilon}(Q)$.

Theorem 3. *For a polymatroid* $f : 2^V \rightarrow \mathbb{R}_+$ *such that* $n = |V|$, $\mathrm{redn}_{f,\epsilon}(Q) \subseteq \mathrm{span}_{f,\epsilon}(Q) \subseteq \mathrm{redn}_{f,n\epsilon}(Q)$ *when* $\epsilon \geq 0$.

Proof. If $f(A|Q) \leq \epsilon$ then, for any ordering $\{a_1, a_2, \ldots, a_k\} = A$, we have

$$f(A|Q) = \sum_{i=1}^{k} \underbrace{f(a_i|a_1, \ldots, a_{i-1}, Q)}_{\geq 0 \text{ and thus } \leq \epsilon} \leq \epsilon \tag{18}$$

Since each term is non-negative, and they sum to ϵ, they all must be at most ϵ. Hence $f(a_1|Q) \leq \epsilon$. Since the order is arbitrary, each a_i takes its turn in the first position, giving $f(a|Q) \leq \epsilon$ for $a \in A$.

A simpler proof is to just note that if $f(A|Q) \leq \epsilon$, then by monotonicity, $f(a|Q) \leq f(A|Q) \leq \epsilon$ for any $a \in A$.

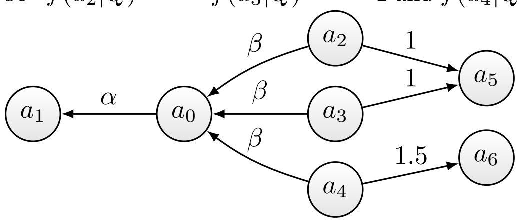
To get the upper set bound, i.e., that $\mathrm{redn}_{n\epsilon}(Q) \supseteq \mathrm{span}_\epsilon(Q)$ when $\epsilon \geq 0$, Assume for a given A that $f(a|Q) \leq \epsilon$. Then $f(A|Q) \leq \sum_{a \in A} f(a|Q) \leq |A|\epsilon \leq n\epsilon$.

We can find a strict example where $\mathrm{redn}_{f,\epsilon}(Q) \subset \mathrm{span}_{f,\epsilon}(Q)$ when $\epsilon > 0$. For $f(A) = \sqrt{A}$, we only need to find a value $m = |Q|$ and ϵ such that $\sqrt{m+2} - \sqrt{m} > \epsilon$ and $\sqrt{m+1} - \sqrt{m} < \epsilon$. Take $m = 1$ and $\epsilon = 0.5$. Then $\sqrt{3} - \sqrt{1} \approx 0.7321 > \epsilon = 0.5$ and $\sqrt{2} - \sqrt{1} \approx 0.4142 < \epsilon = 0.5$.

Lemma 3. *There does not, in general, exist an* $\epsilon' \leq \epsilon$ *such that* $f(\mathrm{span}_{f,\epsilon'}(Q)) = f(\mathrm{redn}_{f,\epsilon}(Q))$ *for all* Q *and* $\epsilon > 0$.

Proof. Assume that $\exists\ \epsilon' \leq \epsilon$ such that $f(\text{span}_{f,\epsilon'}(Q)) = f(A)$ where $A \in \text{redn}_{f,\epsilon}(Q)$. Since $\text{redn}_{f,\epsilon}(Q)$ is not always unique (from Lemma 2), there could exist a set $B \in \text{redn}_{f,\epsilon}(Q)$ such that $f(B) \neq f(\text{span}_{f,\epsilon'}(Q))$.

As an strict example, consider the directed graph below where $Q = \{a_0\}, \epsilon' = 1$ and $\epsilon = 2.5$. Let $f(A) = |\delta^+(A)|$ be the directed cut function, so $f(a_2|Q) = f(a_3|Q) = 1$ and $f(a_4|Q) = 1.5$.



Thus, $A = \text{span}_{f,\epsilon'}(Q) = \{a_2, a_3\}$. For the given ϵ, one can see that all three sets $A = \{a_2, a_3\}, B = \{a_3, a_4\}, C = \{a_2, a_4\}$ can serve as $\text{redn}_{f,\epsilon}(Q)$, but $f(B) = f(C) \neq f(A)$.

Similarly, for any value of $\epsilon' \leq \epsilon$, it can be shown that for the given example, $f(\text{span}_{f,\epsilon'}(Q)) \neq f(\text{redn}_{f,\epsilon}(Q))$ for some set in $\text{redn}_{f,\epsilon}(Q)$.

C Upper Bounds on Functional Evaluations of Span

In this section, we provide two additional bounds that are elucidating regarding bounds resulting functional evaluations of span in various different cases as shown below. While these theorems are not used elsewhere in the paper, since they offer upper bounds on the evaluation of span, they give an indication of worst case results in terms of violations of ϵ the evaluation of span might be. For the two theorems below, and for a given set A, we define $f_Q(A)$ such that $f_{\text{Q}}(A) = f(A \cup Q) - f(Q) = f(A|Q)$.

Lemma 4. *$f(\text{span}_{f,\epsilon}(Q)|Q) \leq (k_s - k_r + 1)\epsilon$ where $k_s = |\text{span}_{f,\epsilon}(Q)|$ and $k_r = |\text{redn}_{f,\epsilon}(Q)|$*

Proof. Let $A_{\text{s}} = \text{span}_{f,\epsilon}(Q)$ and $|A_{\text{s}}| = k_{\text{s}}$. Similarly $A_{\text{r}} = \text{redn}_{f,\epsilon}(Q)$ and $|A_{\text{r}}| = k_{\text{r}}$.

From Theorem 3, we know that $\text{redn}_{f,\epsilon}(Q) \subseteq \text{span}_{f,\epsilon}(Q)$, so $A_r \subseteq A_s$. Let $A_s = A_r \cup A$ such that $A \cap A_r = \emptyset$.

For a submodular function f, we know that:

$$f(Y) \leq f(X) + \sum_{j \in Y \setminus X} f(j|X) - \sum_{j \in X \setminus Y} f(j|(X \cup Y) \setminus j) \tag{19}$$

$\forall\ X, Y \subseteq V$. Substituting X with A_r and Y with A_s in Eq. 19 and using the modified submodular function f_Q, we get:

$$f_{\text{Q}}(A_{\text{s}}) \leq f_{\text{Q}}(A_{\text{r}}) + \sum_{j \in A} f_{\text{Q}}(j|A_{\text{r}}) \tag{20}$$

Since, $f_{\text{Q}}(j|A_{\text{s}}) \leq f_{\text{Q}}(j) \leq \epsilon \quad j \in A$, we modify the upper bound on $f_Q(A_s)$ as follows:

$$f_{\text{Q}}(A_{\text{s}}) \leq \epsilon + (k_{\text{s}} - k_{\text{r}})\epsilon = (k_{\text{s}} - k_{\text{r}} + 1)\epsilon \tag{21}$$

In a situation where k_s is much larger than k_r which could happen when there are many elements in the ground set which are redundant to Q but are mostly non-redundant with respect to each other. So, the worst case bound in Eq. 21 could be $n\epsilon$ where $n = |V|$

A strict example where $f_Q(A_s) \approx n\epsilon$ for $\epsilon \geq 0$ is as follows: Let $f(A) = \min(|A|, n)$ where $n = |V|$ and $\epsilon = 1$. For all $j \in V \setminus Q$, $f_Q(j) = (|Q| + 1) - |Q| = 1 \leq \epsilon$. So, all such data points will belong to the spanning set of Q such that $f_Q(A_s) \approx n\epsilon$ for $|Q| \ll n$.

Lemma 5. *With the conditional submodular curvature with respect to Q defined as*

$$\kappa_{f_Q}(A) \triangleq 1 - \min_{a \in A} \frac{f((a|(A \setminus a)), Q)}{f(a|Q)}, \tag{8}$$

$f(\text{span}_{f,\frac{\epsilon}{n}}(Q)|Q) \leq \epsilon - \frac{\epsilon}{n}(k_r - k_s)(1 - \kappa_{f_Q}(\text{redn}_{f,\epsilon}(Q)))$ *where $k_s = |\text{span}_{f,\frac{\epsilon}{n}}(Q)|$ and $k_r = |\text{redn}_{f,\epsilon}(Q)|$.*

Proof. Let $A_s = \text{span}_{f,\frac{\epsilon}{n}}(Q)$ and $|A_s| = k_s$. Similarly, let $A_r = \text{redn}_{f,\epsilon}(Q)$ and $|A_r| = k_r$.

From Theorem 3, we know that $\text{span}_{f,\frac{\epsilon}{n}}(Q) \subseteq \text{redn}_{\epsilon}(Q)$, so $A_s \subseteq A_r$. Let $A_r = A_s \cup A$ such that $A \cap A_s = \emptyset$.

Let $j_{\text{min}} = \text{argmin}_{j \in A} f_Q(j)$. Since $A \notin A_s$, we have $\frac{\epsilon}{n} \leq f_Q(j_{\text{min}}) \leq \epsilon$.

Substituting X with A_r and Y with A_s in Eq. 19, and using submodular function f_Q, we get:

$$f_Q(A_s) \leq f_Q(A_r) - \sum_{j \in A} f_Q(j|(A_r \setminus j)) \tag{22}$$

$$\overset{\text{a}}{\leq} \epsilon - (k_{\text{r}} - k_{\text{r}})(1 - \kappa_{f_{\text{Q}}}(A_{\text{r}}))f_{\text{Q}}(j_{\text{min}}) \tag{23}$$

$$\overset{\text{b}}{\leq} \epsilon - (k_r - k_s)(1 - \kappa_{f_Q}(\text{redn}_{f,\epsilon}(Q)))\frac{\epsilon}{n} \tag{24}$$

where a holds because $f_Q(A_r) \leq \epsilon$ and from the definition of conditional submodular curvature defined as:

$$\kappa_{f_Q}(A) \triangleq 1 - \min_{a \in A} \frac{f_Q(a|(A \setminus a))}{f_Q(a)} \tag{25}$$

and b holds because of the lower bound on $f_Q(j_{\text{min}})$.

D Proof of Theorem 4

Theorem 4. *Let A^* be the optimal solution to Eq.* (1)*, then A returned by the modular approximation of $f(A|Q)$ such that $A = \text{argmin}_{A \subseteq V_Q, |A| \geq k} m_Q(A)$ satisfies:*

$$f(A|Q) \leq \frac{|A^*|}{1 + (|A^*| - 1)(1 - \kappa_{f_Q}(A^*))} f(A^*|Q)$$

Proof. Here also, lets define $f_Q(A)$ such that $f_{\text{Q}}(A) = f(A \cup Q) - f(Q) = f(A|Q)$.

Let A^* denotes the optimal solution such that $A^* = \text{argmin}_{A \subseteq V', |A| \geq k} f_Q(A)$, and A denotes the approximate solution such that $A = \text{argmin}_{A \subseteq V', |A| \geq k} m_Q(A)$ where

$m_Q(A) = \sum_{a \in A} f_Q(a)$ and $V^{'} = V \setminus Q$. Let $a_{\max} = \text{argmax}_{a \in A^*} f_Q(a)$.

$$f_Q(A) \leq \sum_{a \in A} f_Q(a) = m_Q(A) \leq m_Q(A^*) \quad (26)$$

$$= \sum_{a \in A^*} f_Q(a) \leq |A^*| f_Q(a_{\max}) \quad (27)$$

Substituting Y with $a_{\max}$, X with A^* and using the modified submodular function f_Q, we get:

$$f_Q(A^*) \geq f_Q(a_{\max}) + \sum_{a \in (A^* \setminus a_{\max})} f_Q(a|A^* \setminus a) \quad (28)$$

$$\geq f_Q(a_{\max}) + (|A^*| - 1) \min{}_{a \in A^*} f_Q(a|A^* \setminus a) \quad (29)$$

Combining Eq. 26 and 28, we get:

$$\frac{f_Q(A)}{f_Q(A^*)} \leq \frac{|A^*| f_Q(a_{\max})}{f_Q(a_{\max}) + (|A^*| - 1) \min{}_{a \in A^*} f_Q(a|A^* \setminus a)} \quad (30)$$

$$\leq \frac{|A^*|}{1 + (|A^*| - 1) \frac{\min_{a \in A^*} f_Q(a|A^* \setminus a)}{f_Q(a_{\max})}} \quad (31)$$

$$\leq \frac{|A^*|}{1 + (|A^*| - 1)(1 - \kappa_{f_Q}(A^*))} \quad (32)$$

E Experiments: Additional Details

E.1 Conditional Multi-Document Summarization

Fine-Tuning BERT We use the uncased variant of BERT-base model and fine-tune it on a pair of DUC 2005-2007 datasets, in order to generate feature representations for the remaining dataset which is the test set. We flatten the query tokens and the sentences tokens and then insert [Q#] token before the #-th query sentence as well as insert a [CLS] token at the beginning and [SEP] token at the end of a sentence to mark the sentence boundary. We then fine-tune the BERT-based conditional summarization model as proposed in (Zhu et al. 2019) for a sentence regression task. This is achieved by adding a linear layer on top of the BERT-encoder which predicts score for sentence s_i using h_i where $h_i = \text{BERT}(s_i)_{[\text{CLS}]}$. For network architecture diagram, we ask the readers to please refer to Fig. 2 in (Zhu et al. 2019).

For each sentence in the document, the model predicts a regression score $r(s_i|\mathcal{Q}, \mathcal{D})$ based on its relevance to the query $\mathcal{Q}$ and its salience to the document $\mathcal{D}$. The training objective is to minimize the mean square error between the predicted scores and the oracle scores that we evaluate using the gold summary $\mathcal{S}^*$.

$$\mathcal{L} = \frac{1}{n} \sum_{i=n}^{n} (t(s_i|\mathcal{S}^*) - r(s_i|\mathcal{Q}, \mathcal{D}))^2 \quad (33)$$

Since the maximum sequence token for BERT-base is set to 512, we split long documents into multiple training examples such that the overlap of a new training example with the previously generated example is four sentences. Using this setup, we are able to generate 12704 training examples for DUC 2005, 9605 training examples for DUC 2006 and 6727n training examples for DUC 2007.

We fine-tune the BERT-encoder model three times, on each pair of query-focused DUC datasets. For training, similar to (Zhu et al. 2019), we use Adam optimizer (Kingma and Ba 2015) with learning rate of $3e-5$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, linear decay of learning rate, and L2 weight decay of 0.01. We set the batch size as 32 and train the above-mentioned architecture for three epochs. For each pair of training set, for example DUC 2005-2006, we instantiate the network using the original BERT-base weights and do not make use of any fine-tuned weights obtained via training on other pairs of DUC 2005-2007. This way, we ensure that the computed sentence representations for the test set (DUC 2007 in this example) do not capture any information about the oracle summarization scores.

After fine-tuning, we use the network fine-tuned on DUC 2005-2006 sets to extract sentence representations for the left-out DUC 2007. Similarily we use the network fine-tuned on DUC 2005 and 2007 to extract sentence representations for DUC 2006 and so on.

S3: Hyperparameter Tuning Since BERT-base has 12 transformer layers, we first set to investigate which of these layers is most transferable for the conditional summarization task. We find out that the layers close to the input are most transferable for this task. This is in assent with our feature extraction strategy which does not allow the representations to encode any information about the oracle summary. Specifically, we found that the second layer (starting from the input layer which is layer 0) is the most transferable after comparing the Rouge-2 F-measure scores across all 12 layers. We also tried using the transformer's output corresponding to the [CLS] token but it led to rather poor performance as it is suited mostly for sentence classification tasks. Also in case of the minimally supervised case where we used the pre-trained BERT weights to generate sentence representations for instantiating the submodular function, we found that the first and second layers close to the input were most transferable for the summarization task. So, when using the pre-trained BERT encoder weights, average pooled sentence representations across the first and second layers lead to the best minimally supervised results on the DUC 2005-2007 datasets.

For computing ϵ such that $f(A|Q) \leq \epsilon$, we modify the inequality $f(A|Q) \leq f(V|Q)$ to $f(A|Q) \leq \epsilon' f(V|Q)$ where $0 < \epsilon' \leq 1$. Among similarity measures for instantiating the facility location function, we compare three similarity measures: (1) modified cosine similarity which truncates all negative similarities to zero (2) d - *euclidean distance* where d is the maximum euclidean distance between any pair of feature representations (3) Gaussian kernel with width σ. The hyperparameters set includes: (1) $\epsilon' \in \{0.05, 0.1, 0.125, 0.15, 0.175, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45\}$ (2) $\sigma \in \{0.75, 1.0, 1.25, 1.5, 2, 5, 10, 20, 50, 100, 1000, 5000, 10000\}$ (3) $r \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. We use DUC 2005 as development set for DUC 2006 and DUC 2007. For DUC 2005, we use DUC 2006 as the development set for hyperparameter tuning. We found that similarity based on Gaussian kernel with width $\sigma = 1.0$, $\epsilon' = 0.25$, and $r = 0.3$ lead to the best reported performance on all three DUC datasets.

Additional Results In Table 4, 5 and 6, we provide different examples of conditional summaries generated using the S3 framework and compare them with their generic summaries (sentence ordering in place). It can be seen that the summary conditioned on the title of the paper as query is better in terms of capturing the most salient and relevant sentences from the abstract of the paper. In all examples, we set the size constraint corresponding to the stage one as six and the final summary size as three sentences.

Generic Summary: We then propose a two-stage Submodular Span Summarization (S3) framework to achieve a form of conditional or query-focused data summarization. The first stage encourages the summary to be relevant to a given query set, and the second stage encourages the final summary to be diverse, thus achieving two important necessities for a good query-focused summary. *We find that a scalable approximation algorithm to the polymatroid submodular span problem has good both theoretical and empirical properties.*
Conditional Summary: We then propose a two-stage Submodular Span Summarization (S3) framework to achieve a form of conditional or query-focused data summarization. The first stage encourages the summary to be relevant to a given query set, and the second stage encourages the final summary to be diverse, thus achieving two important necessities for a good query-focused summary. *Unlike previous methods, our framework uses only a single submodular function defined over both data and query.*

Table 4: Generic summary and conditional summary of the abstract of this paper using its title as the query. Dissimilar sentences among the summaries are marked in *italics*.

E.2 Conditional Video Summarization

Dataset: Detailed Description The query-focused video summarization dataset from (Sharghi, Laurel, and Gong 2017) consists of four egocentric videos (compiled from the UTE dataset), each of 3-5 hours, captured at a frame rate of 15 fps. The videos are diverse with respect to the set of events that they capture. Each video shot is annotated with captions in English. Using the nouns in the video shot captions and SentiBank (Borth et al. 2013), the authors composed a lexicon of 48 concepts which includes {*Baby, Beach, Bed, Blonde, Boat, Book, Building, Car, Chair, Chocolate, Computer, Cup-Glass, Dance, Desk, Drink, Exercise, Face, Flower, Food, Garden, Grass, Hall, Hands, Hat, Kids, Lady, Lake, Market, Men, MusicalInstrument, Ocean, Office, Park, Party, Pets-Animal, Phone, Room, School, Shoes, Sign, Sky, Sports, Street, Student, Sun, Toy, Tree, Window*}.

Each video shot is annotated for the presence and absence of these concepts leading to a binary semantic video shot representation. But these binary representations are used only for evaluation purposes. For each video, there are 46 queries, each of which contain a pair of the lexicon concepts. Thus, for representing a query, we use a binary vector with 1's representing the concepts present in the query.

For evaluation purposes, (Sharghi, Laurel, and Gong 2017) have provided user-annotated semantic vectors of each video shot and have defined a similarity function between two video

Title: WinoGrande: An Adversarial Winograd Schema Challenge at Scale
Abstract: The Winograd Schema Challenge (WSC) (Levesque, Davis, and Morgenstern 2011), a benchmark for commonsense reasoning, is a set of 273 expert-crafted pronoun resolution problems originally designed to be unsolvable for statistical models that rely on selectional preferences or word associations. However, recent advances in neural language models have already reached around 90% accuracy on variants of WSC. This raises an important question whether these models have truly acquired robust commonsense capabilities or whether they rely on spurious biases in the datasets that lead to an overestimation of the true capabilities of machine commonsense. To investigate this question, we introduce WinoGrande, a large-scale dataset of 44k problems, inspired by the original WSC design, but adjusted to improve both the scale and the hardness of the dataset. The key steps of the dataset construction consist of (1) a carefully designed crowdsourcing procedure, followed by (2) systematic bias reduction using a novel AfLite algorithm that generalizes human-detectable word associations to machine-detectable embedding associations. The best state-of-the-art methods on WinoGrande achieve 59.4-79.1%, which are 15-35% below human performance of 94.0%, depending on the amount of the training data allowed. Furthermore, we establish new state-of-the-art results on five related benchmarks - WSC (90.1%), DPR (93.1%), COPA (90.6%), KnowRef (85.6%), and Winogender (97.1%). These results have dual implications: on one hand, they demonstrate the effectiveness of WinoGrande when used as a resource for transfer learning. On the other hand, they raise a concern that we are likely to be overestimating the true capabilities of machine commonsense across all these benchmarks. We emphasize the importance of algorithmic bias reduction in existing and future benchmarks to mitigate such overestimation.
Generic Summary: *This raises an important question whether these models have truly acquired robust commonsense capabilities or whether they rely on spurious biases in the datasets that lead to an overestimation of the true capabilities of machine commonsense.* To investigate this question, we introduce WinoGrande, a large-scale dataset of 44k problems, inspired by the original WSC design, but adjusted to improve both the scale and the hardness of the dataset. Furthermore, we establish new state-of-the-art results on five related benchmarks - WSC (90.1%), DPR (93.1%), COPA (90.6%), KnowRef (85.6%), and Winogender (97.1%).
Conditional Summary: *The Winograd Schema Challenge (WSC), a benchmark for commonsense reasoning, is a set of 273 expert-crafted pronoun resolution problems originally designed to be unsolvable for statistical models that rely on selectional preferences or word associations.* To investigate this question, we introduce WinoGrande, a large-scale dataset of 44k problems, inspired by the original WSC design, but adjusted to improve both the scale and the hardness of the dataset. Furthermore, we establish new state-of-the-art results on five related benchmarks - WSC (90.1%), DPR (93.1%), COPA (90.6%), KnowRef (85.6%), and Winogender (97.1%).

Table 5: Generic summary and conditional summary of the abstract of the AAAI-20 outstanding paper using its title as the query. Dissimilar sentences among the summaries are marked in *italics*.

shots as Intersection Over Union (IOU) of their corresponding concepts. For instance, if one shot is tagged by {*car, street*} and another by {*street, tree, sign*}, then the IOU similarity between them is $1/4 = 0.25$. Next, a maximum weight based bipartite graph matching between the oracle summary shots and the system generated summary shots is computed using the IOU scores, which enables us to evaluate precision, recall and F1 score between the matched pairs.

S3: Hyperparameter Tuning For the facility location function, we again investigate three similarity measures: (1)

Title: How to Combine Tree-Search Methods in Reinforcement Learning
Abstract: Finite-horizon lookahead policies are abundantly used in Reinforcement Learning and demonstrate impressive empirical success. Usually, the lookahead policies are implemented with specific planning methods such as Monte Carlo Tree Search (e.g. in AlphaZero). Referring to the planning problem as tree search, a reasonable practice in these implementations is to back up the value only at the leaves while the information obtained at the root is not leveraged other than for updating the policy. Here, we question the potency of this approach. Namely, the latter procedure is non-contractive in general, and its convergence is not guaranteed. Our proposed enhancement is straightforward and simple: use the return from the optimal tree path to back up the values at the descendants of the root. This leads to a γ^h- contracting procedure, where γ is the discount factor and h is the tree depth. To establish our results, we first introduce a notion called multiple-step greedy consistency. We then provide convergence rates for two algorithmic instantiations of the above enhancement in the presence of noise injected to both the tree search stage and value estimation stage.
Generic Summary: *Referring to the planning problem as tree search, a reasonable practice in these implementations is to back up the value only at the leaves while the information obtained at the root is not leveraged other than for updating the policy. Here, we question the potency of this approach.* We then provide convergence rates for two algorithmic instantiations of the above enhancement in the presence of noise injected to both the tree search stage and value estimation stage.
Conditional Summary: *Finite-horizon lookahead policies are abundantly used in Reinforcement Learning and demonstrate impressive empirical success. This leads to a γ^h- contracting procedure, where γ is the discount factor and h is the tree depth.* We then provide convergence rates for two algorithmic instantiations of the above enhancement in the presence of noise injected to both the tree search stage and value estimation stage.

Table 6: Generic summary and conditional summary of the abstract of the AAAI-19 outstanding paper using its title as the query. Dissimilar sentences among the summaries are marked in *italics*.

cosine similarity (2) d - *euclidean distance* where d is the maximum euclidean distance between any pair of video shots (3) Gaussian kernel with width σ. In this case, we do not need to modify the cosine similarity to be non-negative as the video shot features are simply probability scores for each lexicon concept. For determining the cardinality constraints, k_1 for stage one and k_2 for stage two of the S3 framework, we use $k_1^{'}$ such that $k_1 = \left\lfloor k_1^{'}|V_i| \right\rfloor$ where V_i denotes the set of video shots present in the video i that has to be summarized. Similarly, we introduce $k_2^{'}$ corresponding to k_2. The hyperparameters set includes: (1) $\sigma \in \{$0.05, 0.1, 0.5, 1, 2, 10, 100, 200, 1000, 10000$\}$ (2) $k_1^{'} \in \{$0.2, 0.25, 0.275, 0.3, 0.325, 0.35, 0.375, 0.4, 0.45$\}$ (3) $k_2^{'} \in \{$0.01, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05, 0.07, 0.09, 0.1$\}$. We tried both three-fold cross validation i.e., use three videos for hyperparameter tuning and test them on the held-out test video, and one-fold cross validation which led to best set of results. For Video 1, we use Video 3 as the development set and for all remaining videos, we use Video 1 as the development set. Among different similarity measure, cosine similarity led to the best results in terms of all three metrics, precision, recall, and F1 score. For Video 1, corresponding to $k_1^{'} = 0.275$ and $k_2^{'} = 0.03$, we get the results reported in Table 2. For remaining videos where we use Video 1 as the development set, we get the reported results corresponding to $k_1^{'} = 0.375$ and $k_2^{'} = 0.03$.

E.3 Conditional Image Corpus Summarization

BiGAN Features BiGAN is an extension of GAN with a third component as an encoder which maps the data space x to a latent space z. In addition, the discriminator is trained to discriminate between the joint samples (x, z) coming from the encoder distribution and the decoder (or generator) distribution. The encoder architecture of BiGAN in (Donahue, Krähenbühl, and Darrell 2016) follows AlexNet (Krizhevsky, Sutskever, and Hinton 2012). We use its last convolutional layer i.e., *conv5* to encode the ImageNet images as we do not want to use any trained model that uses the class information. After building a sparse k-NN similarity graph ($k = 1000$) using the 1024-dimensional features, and analyzing the pairwise cosine similarities, we found that most of the non-negative similarities are concentrated around 0.95. So, we reduce the feature dimension to 512 using PCA (retaining 95% of the total variance). We also used the averaged top-k accuracy of the nearest neighbors of the validation query images to further validate the feature dimension chosen. Here, we set $k \in \{10, 50, 100\}$.

S3: Hyperparameter Tuning For the facility location function, we experimented with two similarity measures: (1) cosine similarity (2) Gaussian kernel with width σ. Our hyperparameters set includes: (1) $\sigma \in \{$1, 10, 50, 100, 200, 500, 1000, 10000$\}$ (2) $k_1 \in \{$50, 75, 100, 125, 150, 175, 200, 225, 250$\}$. We find that cosine similarity with $k_1 = 225$ leads to best relevance and diversity scores on the validation query set.

Additional Results In Table 7, 8, 9, and 10, we show a comparison between the submodular span summary and the k-nearest neighbors for different query images when $|Q| = 1$. We generate a submodular span summary of 25 images and compare it side-by-side with the top-25 nearest neighbors of the query. We utilize tSNE plots to better visualize the similarity/dissimilarity among the submodular span summary images. It can be seen that the tSNE plot corresponding to k-nearest neighbors consists of clusters of similar images that are redundant among themselves, while the submodular span summary consists of diverse views of images belonging to the query class or images which share visually similar features with the query image. For example, in the first example in Table 8 corresponding to the query *great white shark*, the submodular span summary consists of images with diverse views of *shark*, images of different aquatic creatures as well as images of *scuba diver* inside water. Similarly, corresponding to the query *zebra* in Table 9, the k-nearest neighbors consists of redundant images belonging to the query class while the submodular span summary consists of diverse images of *zebra* in different backgrounds as well as images from other classes having features similar to stripes. This shows that the submodular span summary can capture different properties associated with the query set, thus leading to

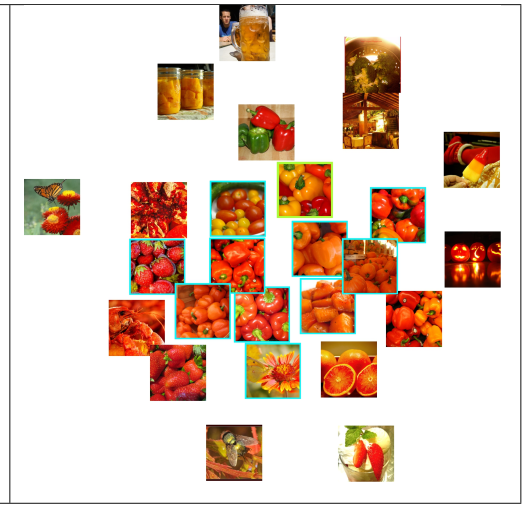
k-Nearest Neighbors	Submodular Span Summary
	

Table 7: Qualitative Results on the ImageNet Dataset when $|Q| = 1$. The first column shows the top-25 nearest neighbors given a query and the second column shows the query-focused summary. Images marked with a surrounding green box and light blue box denote the query image and common images among the two columns respectively.

a diverse query-focused summary that is highly related (or redundant) to the query set.

In Table 11 and 12, for different query sets where $|Q| > 1$, we show their conditional summaries after fitting the two-dimensional tSNE vectors of the image embeddings onto a square grid using the Jonker-Volgenant algorithm (Jonker and Volgenant 1987). For query examples where the query images consist of very diverse views of the same class, the resulting summary can capture both views as in the last example in Table 12 for query class *strawberry*. For example, in Table 12 corresponding to the query set consisting of images belonging to $\{$*ambulance, cab*$\}$ classes, the resulting conditional summary consists of *ambulance* images matching *cab* hues as well as *racing cars* matching *ambulance* hues. These examples demonstrate that given a query set compassing multiple objects, the resultant S3 summary is not only able to capture the relevant object properties but the hybrid properties among them as well.

k-Nearest Neighbors	Submodular Span Summary

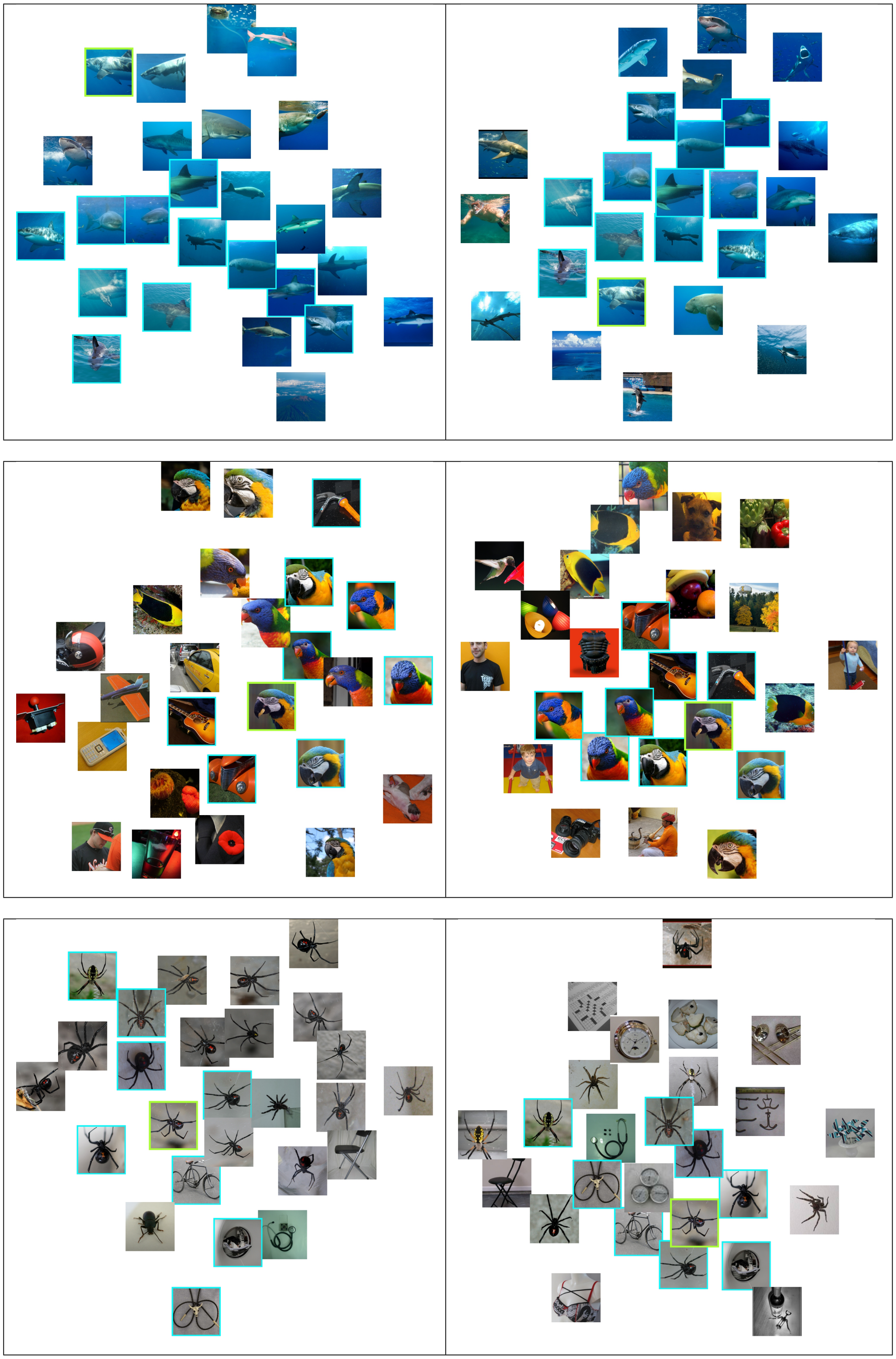


Table 8: Qualitative Results on the ImageNet Dataset when $|Q| = 1$. The first column shows the top-25 nearest neighbors given a query and the second column shows the query-focused summary. Images marked with a surrounding green box and light blue box denote the query image and common images among the two columns respectively.

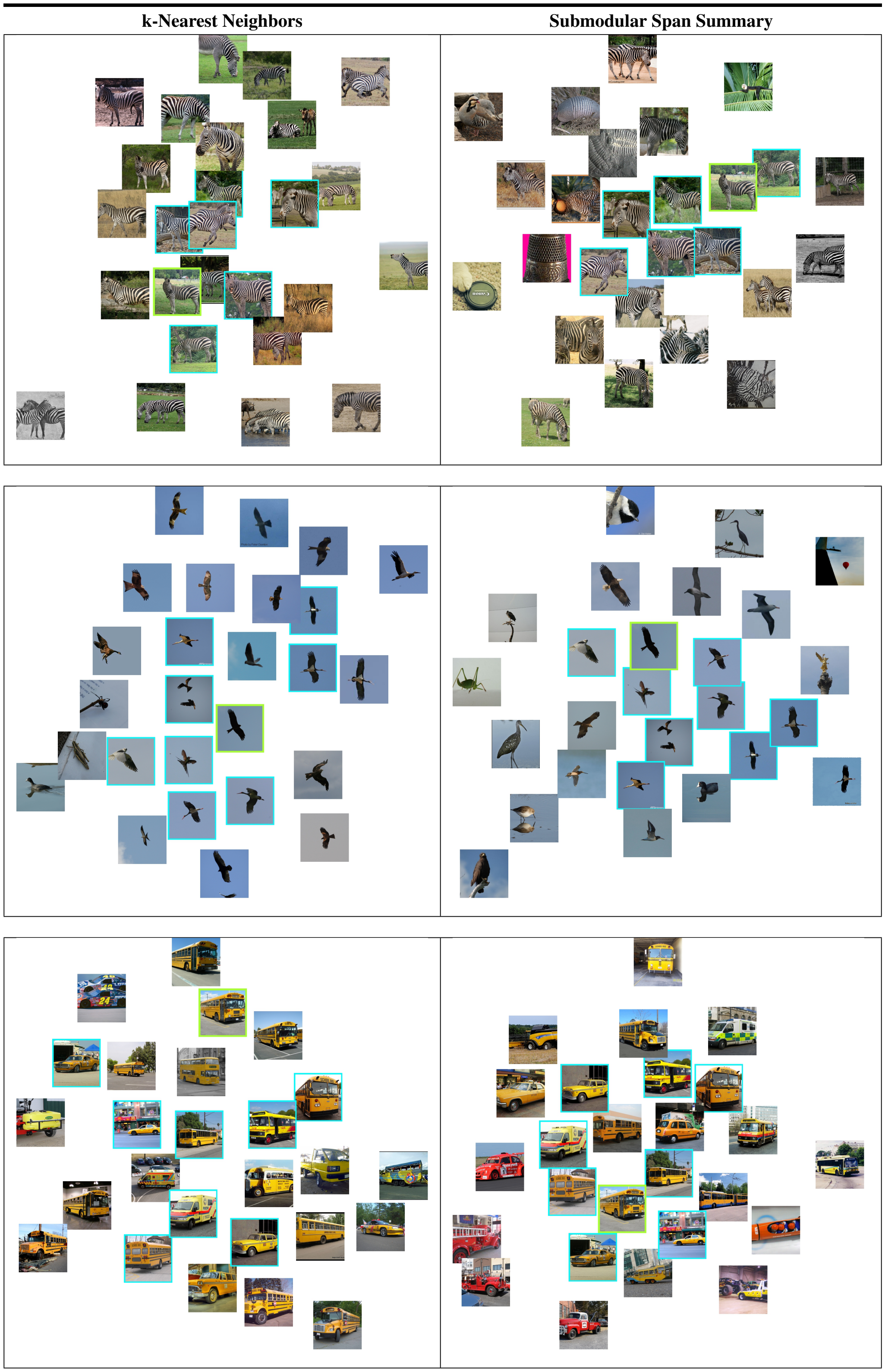


Table 9: Qualitative Results on the ImageNet Dataset when $|Q| = 1$. The first column shows the top-25 nearest neighbors given a query and the second column shows the query-focused summary. Images marked with a surrounding green box and light blue box denote the query image and common images among the two columns respectively.

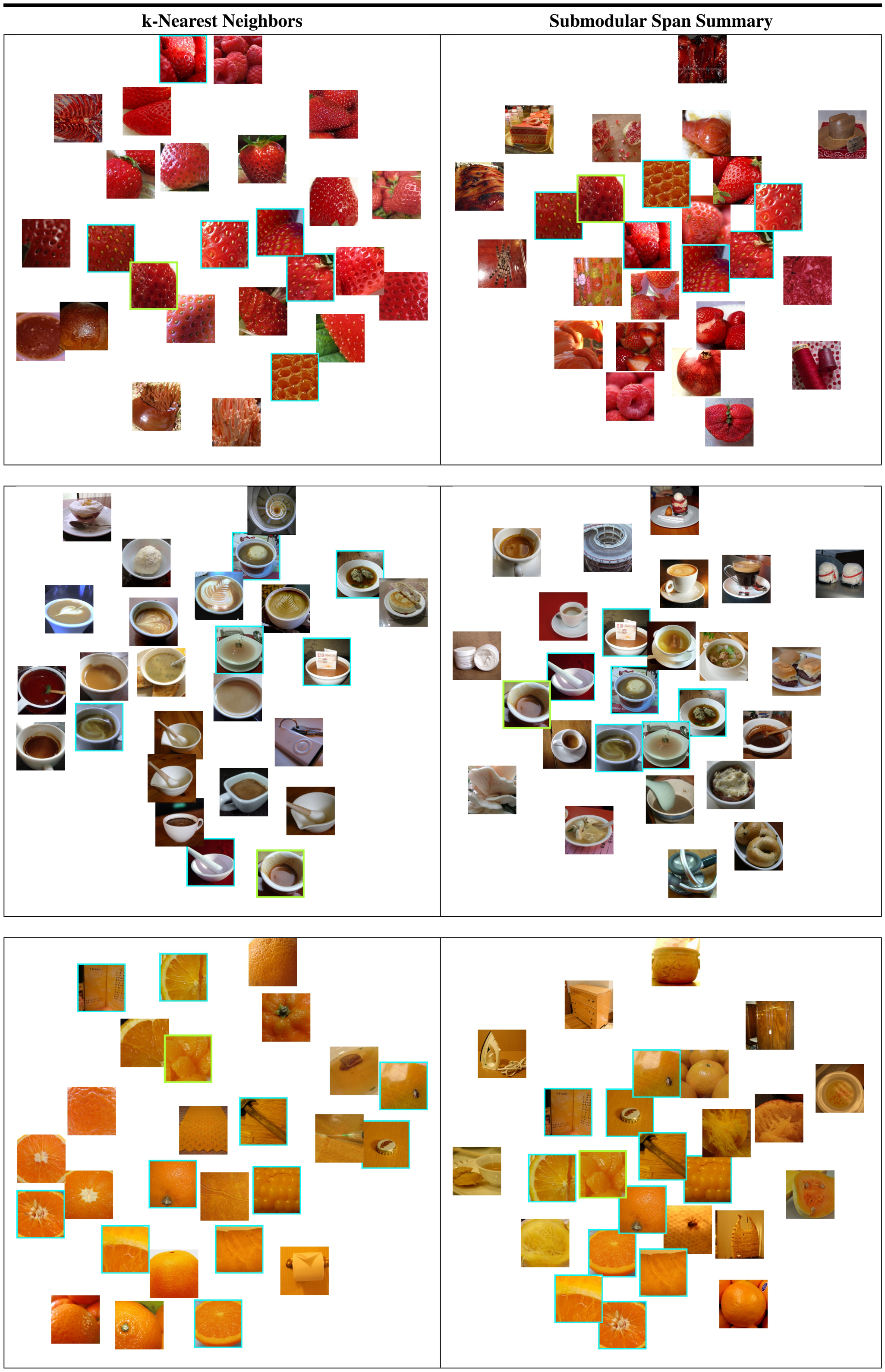


Table 10: Qualitative Results on the ImageNet Dataset when $|Q| = 1$. The first column shows the top-25 nearest neighbors given a query and the second column shows the query-focused summary. Images marked with a surrounding green box and light blue box denote the query image and common images among the two columns respectively.

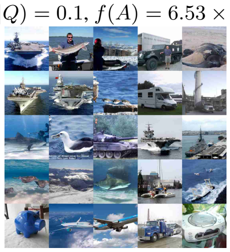
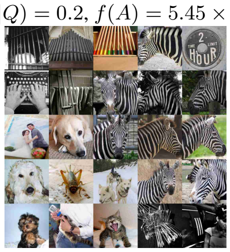
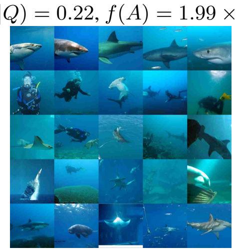
Query	Submodular Span Summary
	$\mathcal{R}(A\|Q) = 0.16, f(A) = 3.63 \times 10^{-3}$ 
	$\mathcal{R}(A\|Q) = 0.1, f(A) = 6.53 \times 10^{-3}$ 
	$\mathcal{R}(A\|Q) = 0.2, f(A) = 5.45 \times 10^{-3}$ 
	$\mathcal{R}(A\|Q) = 0.22, f(A) = 1.99 \times 10^{-3}$ 
	$\mathcal{R}(A\|Q) = 0.24, f(A) = 5.09 \times 10^{-3}$ 

Table 11: Qualitative Results on the ImageNet Dataset corresponding to $|Q| > 1$

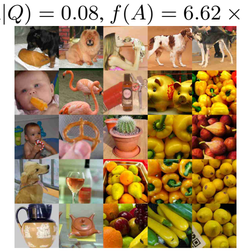
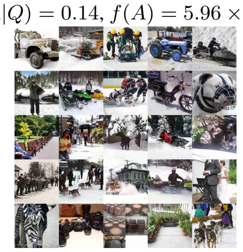
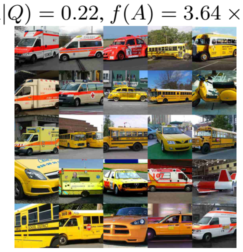
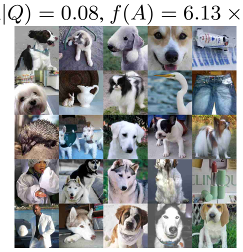
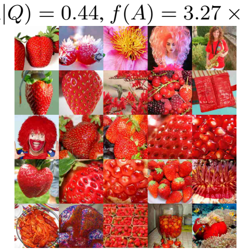
Query	Submodular Span Summary
	$\mathcal{R}(A\|Q) = 0.08, f(A) = 6.62 \times 10^{-3}$ 
	$\mathcal{R}(A\|Q) = 0.14, f(A) = 5.96 \times 10^{-3}$ 
	$\mathcal{R}(A\|Q) = 0.22, f(A) = 3.64 \times 10^{-3}$ 
	$\mathcal{R}(A\|Q) = 0.08, f(A) = 6.13 \times 10^{-3}$ 
	$\mathcal{R}(A\|Q) = 0.44, f(A) = 3.27 \times 10^{-3}$ 

Table 12: Qualitative Results on the ImageNet Dataset corresponding to $|Q| > 1$