Trainable Decoding of Sets of Sequences for Neural Sequence Models

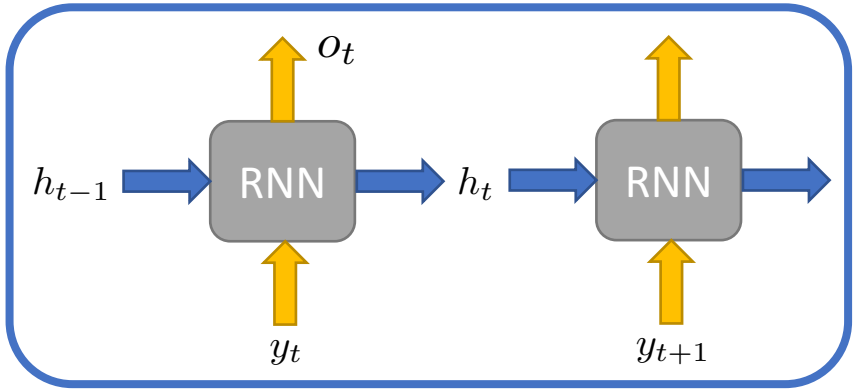



Ashwin Kalyan Peter Anderson Stefan Lee Dhruv Batra

Georgia Tech

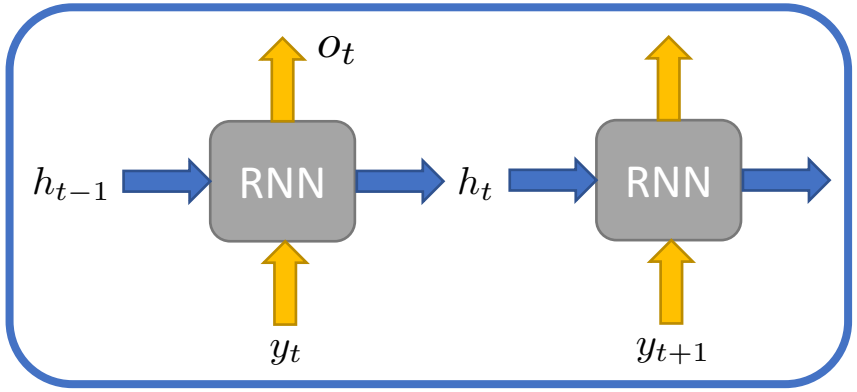
facebook Artificial Intelligence Research

Standard Sequence Prediction Pipeline

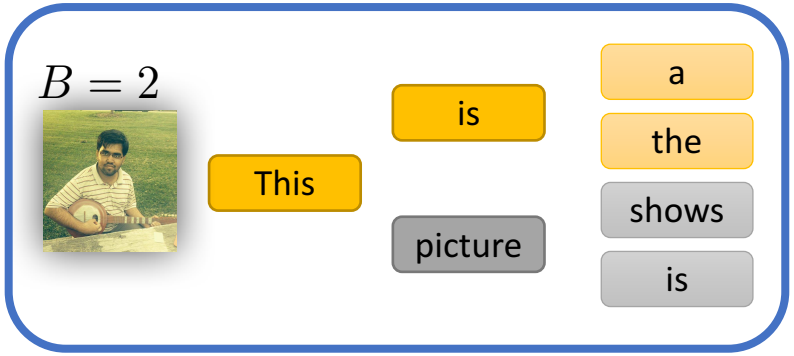


1. Train RNNs to maximize Log Likelihood

Standard Sequence Prediction Pipeline

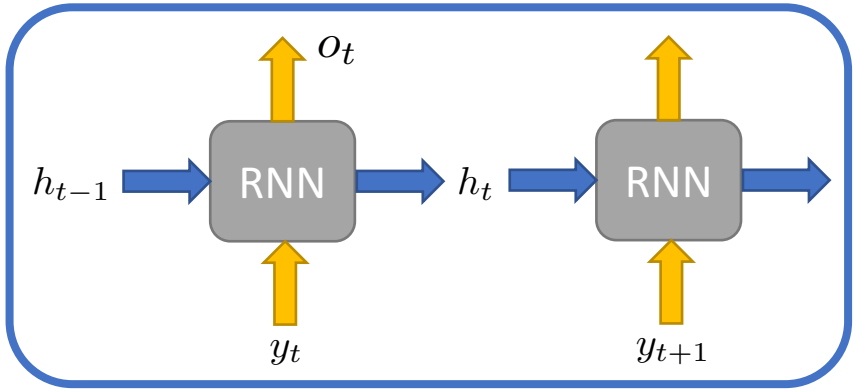


1. Train RNNs to maximize Log Likelihood

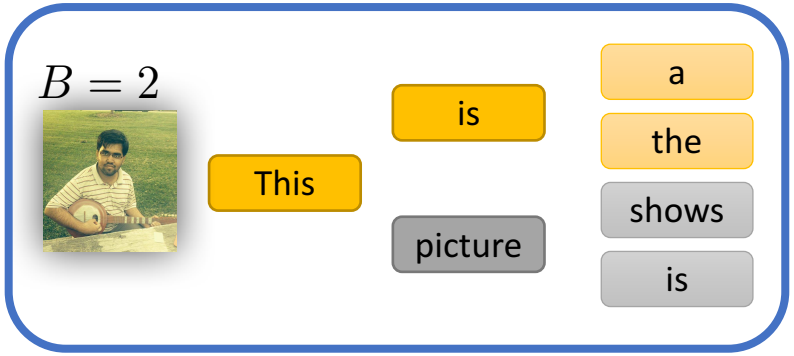


2. Perform Beam Search to decode top K

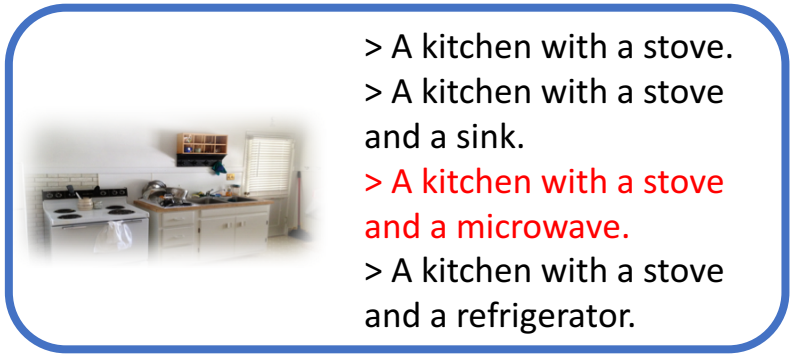
Standard Sequence Prediction Pipeline



1. Train RNNs to maximize Log Likelihood



2. Perform Beam Search to decode top K



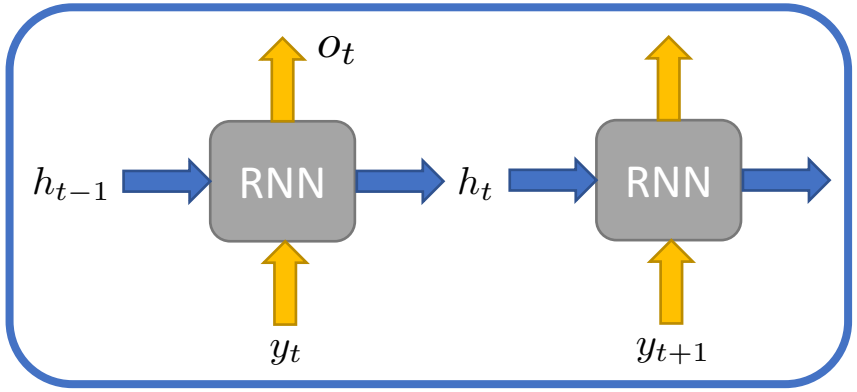
3. Return the best sequence in the top K

But… many real world tasks are multi-modal!

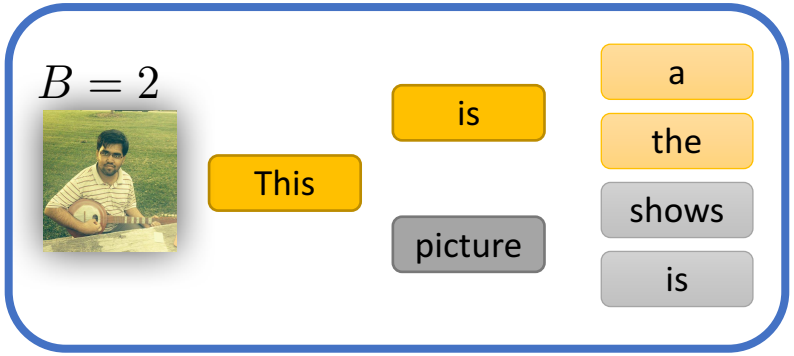
- ✓ A group of people riding horses.
- ✓ Kids riding horses with adults help.
- ✓ A girl poses on her horse in equestrian dress by a small crowd.
- ✓ Some people stand near some horses in a field.
- ✓ People are standing around children riding horses in a grassy area.
- ✓ A small girl is riding a large light brown horse.
- ✓ A young girl in riding gear mounts a pony in front of a group.
- ✓ A group of people with a jockey and her horse
- ✓ Several people playing with ponies in a park.

How to model more than one correct output?

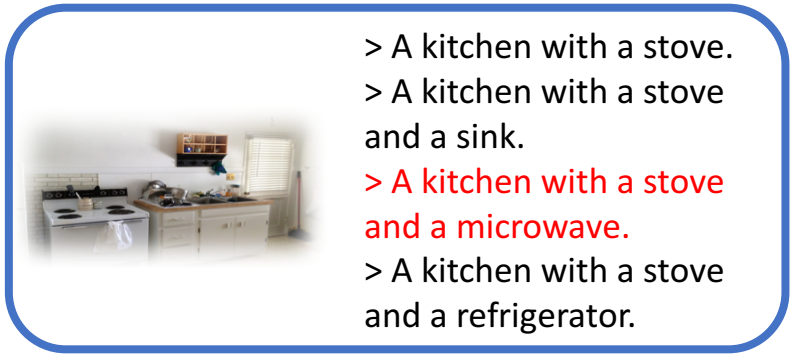
Retool the Standard Sequence Prediction Pipeline



1. Train RNNs to maximize Log Likelihood

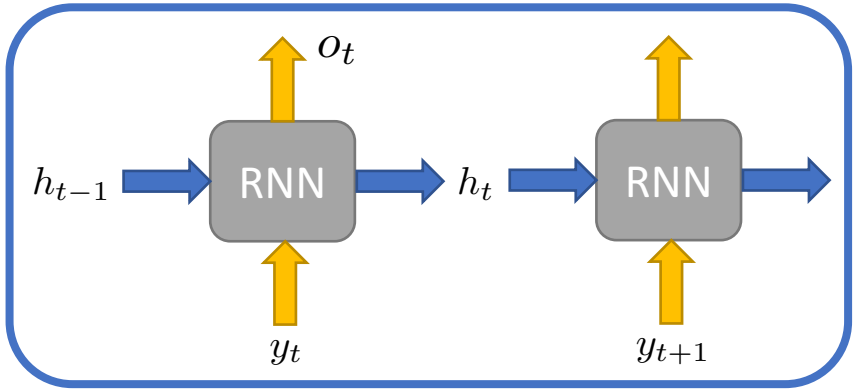


2. Perform Beam Search to decode top K

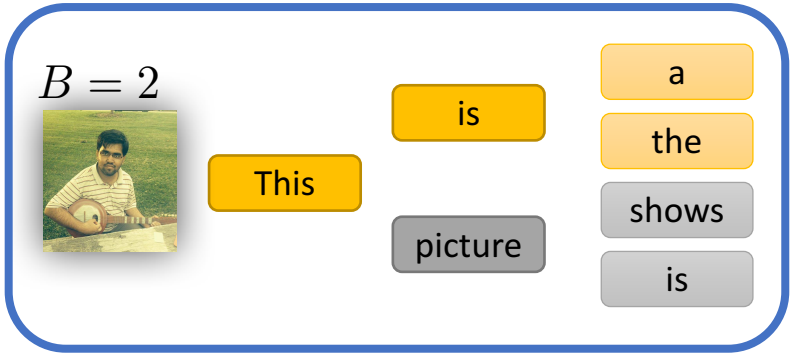


3. Return the best sequence in the top K

Retool the Standard Sequence Prediction Pipeline



1. Train RNNs to maximize Log Likelihood



2. Perform Beam Search to decode top K

3. Re best sequ e top K

Beam Search outputs are nearly identical!

- A group of people riding horses on a field.
- A group of people riding horses in a field.
- A group of people riding horses down a dirt road.
- A group of people riding horses through a field.
- A group of people riding on the back of horses.
- A group of people riding on the back of a horse.
- A group of people riding on a horse.
- A couple of people riding on the back of horses.
- A couple of people riding on the back of a horse.
- A couple of people riding horses on a field.

Doesn't model intra-set interactions!

Beam Search outputs are nearly identical!

- A group of people riding horses on a field.
- A group of people riding horses in a field.
- A group of people riding horses down a dirt road.
- A group of people riding horses through a field.
- A group of people riding on the back of horses.
- A group of people riding on the back of a horse.
- A group of people riding on a horse.
- A couple of people riding on the back of horses.
- A couple of people riding on the back of a horse.
- A couple of people riding horses on a field.

Doesn't model intra-set interactions!

Fails to COVER the variation in the output space!

Learning to Decode Sets of Sequences

Select top-B words at each time step

$B = 2$

This

is

picture

t

Learning to Decode Sets of Sequences

Select top-B words at each time step

$B = 2$

This

is

picture

a

the

shows

is

t

Learning to Decode Sets of Sequences

Select top-B words at each time step

$B = 2$

This

is

a

picture

shows

t

Learning to Decode Sets of Sequences

Select top-B words at each time step

$B = 2$

This

is

a

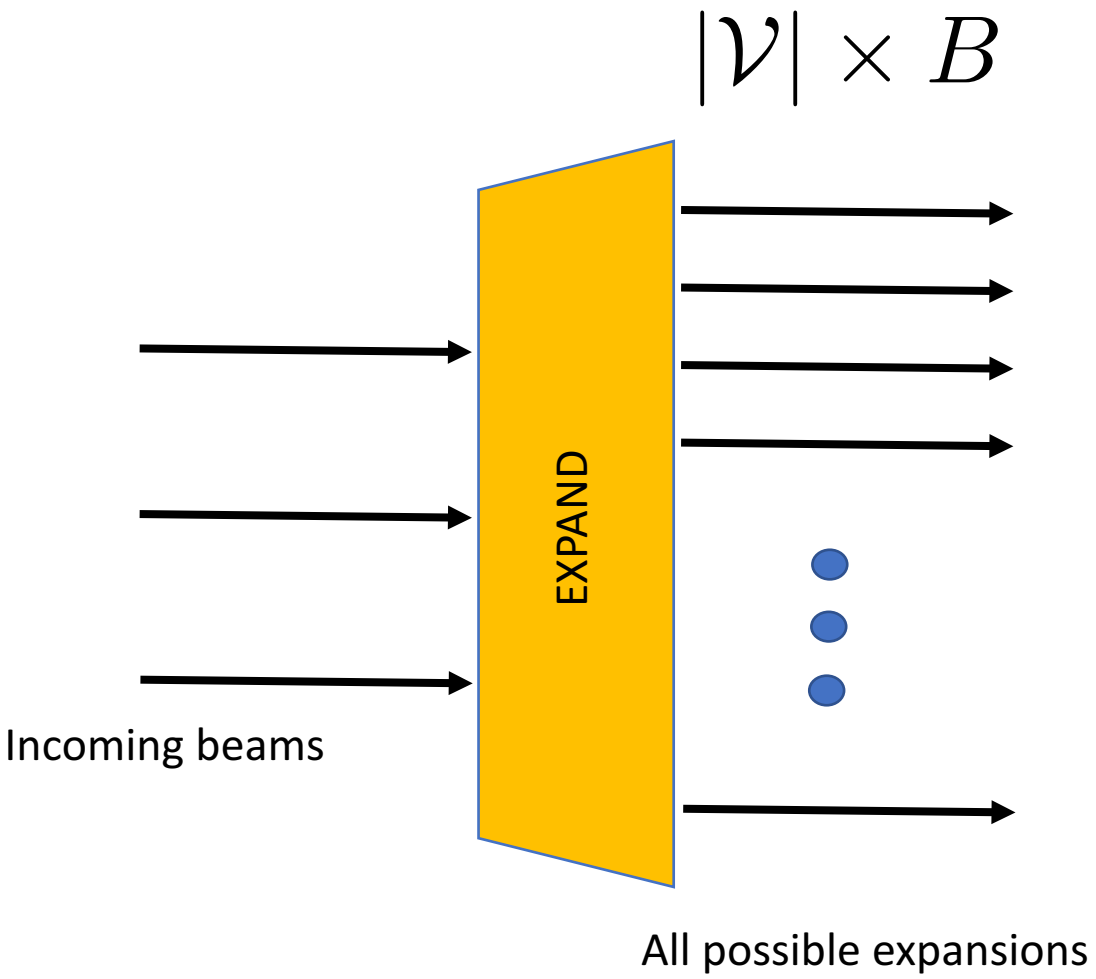
picture

shows

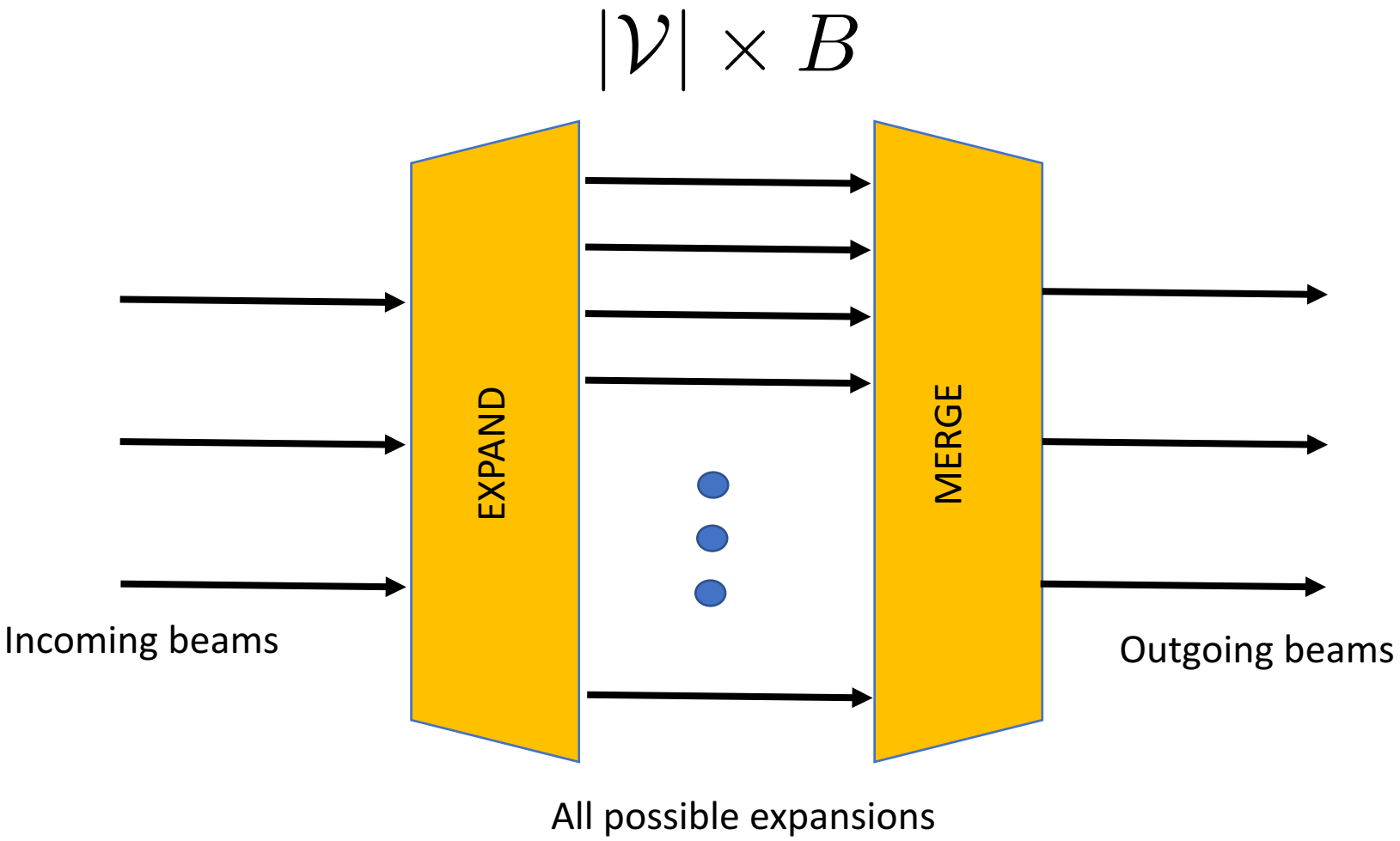
Till end token is generated or max time

t

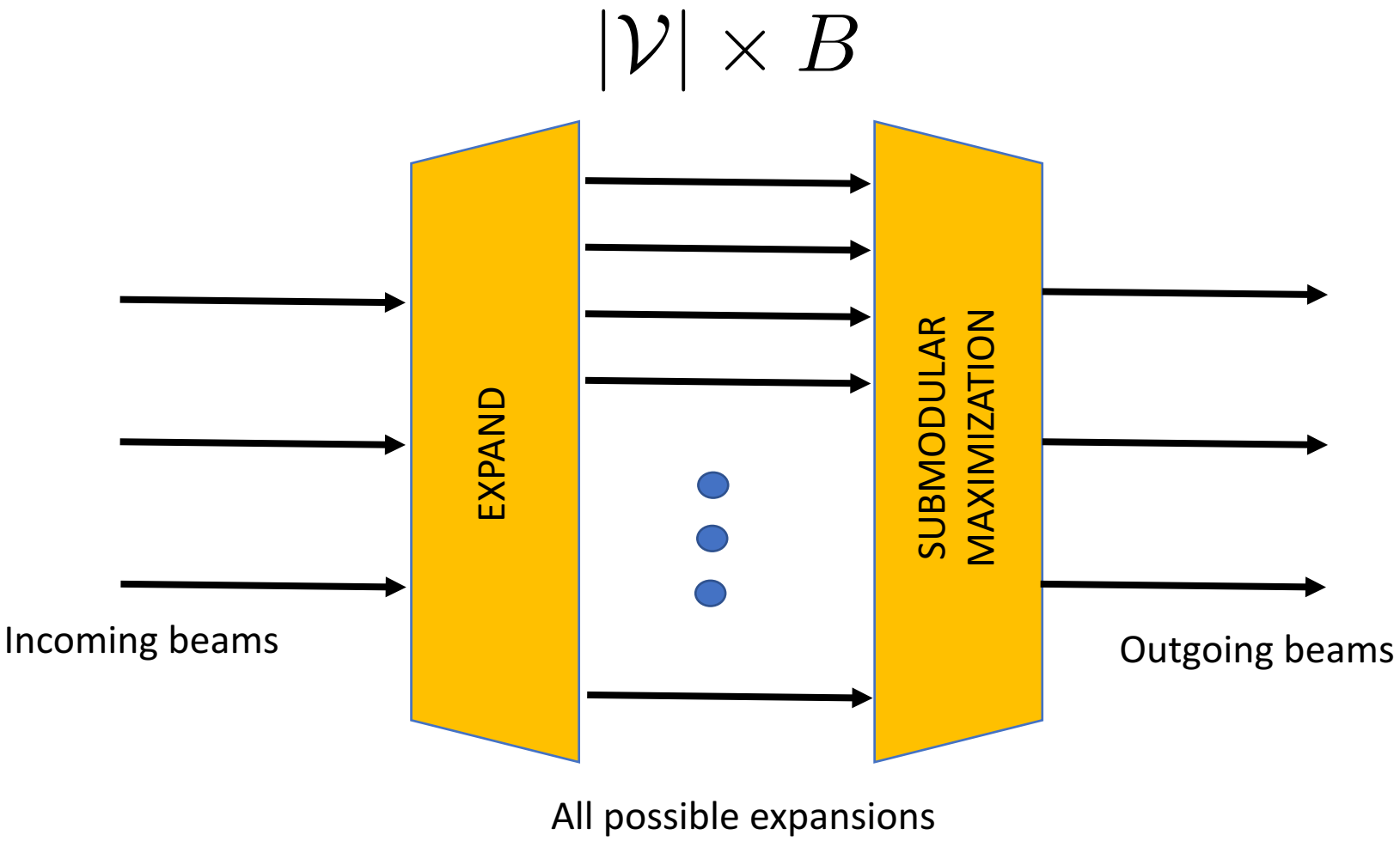
Beam Search as Subset Selection



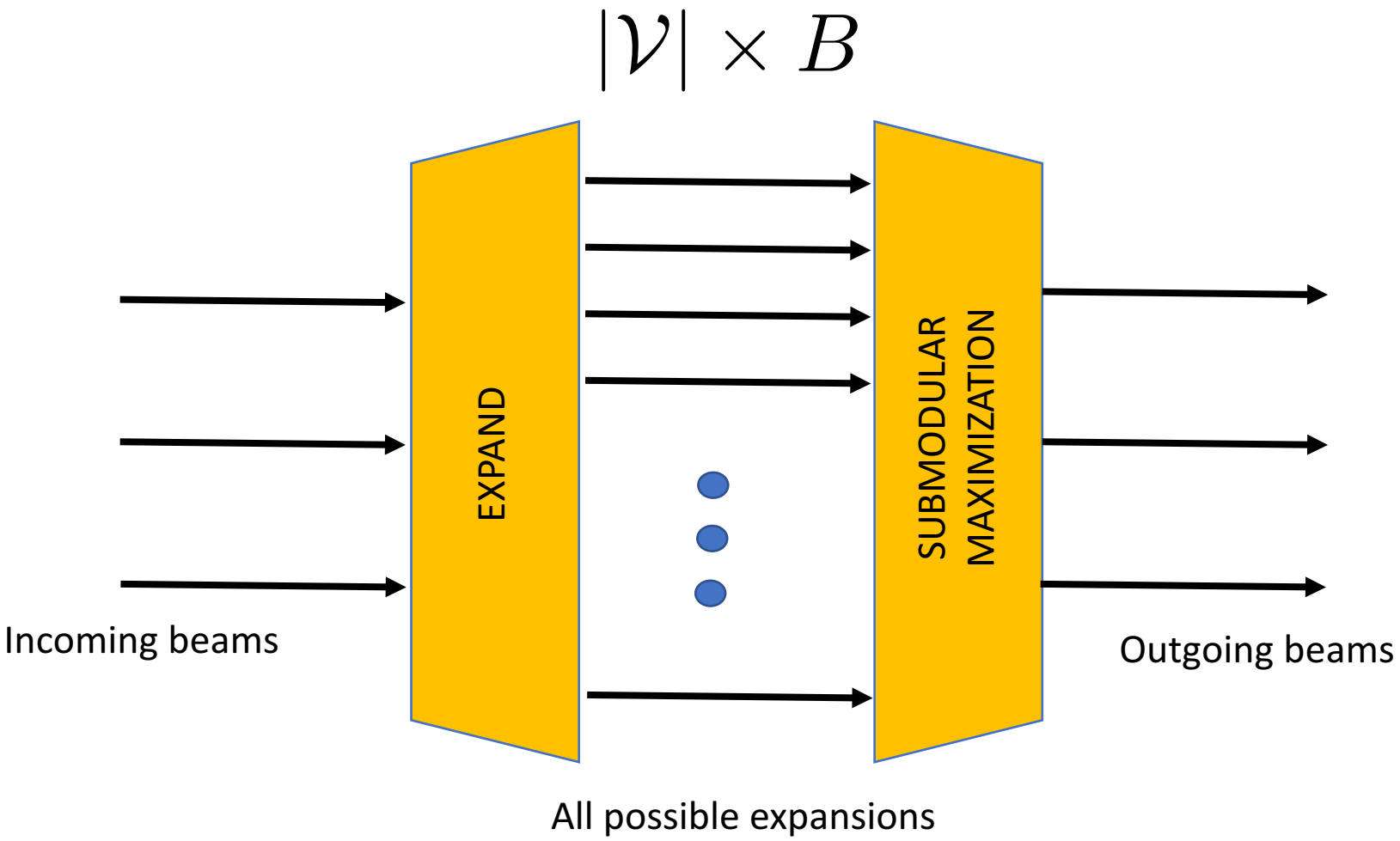
Beam Search as Subset Selection



Beam Search as Subset Selection

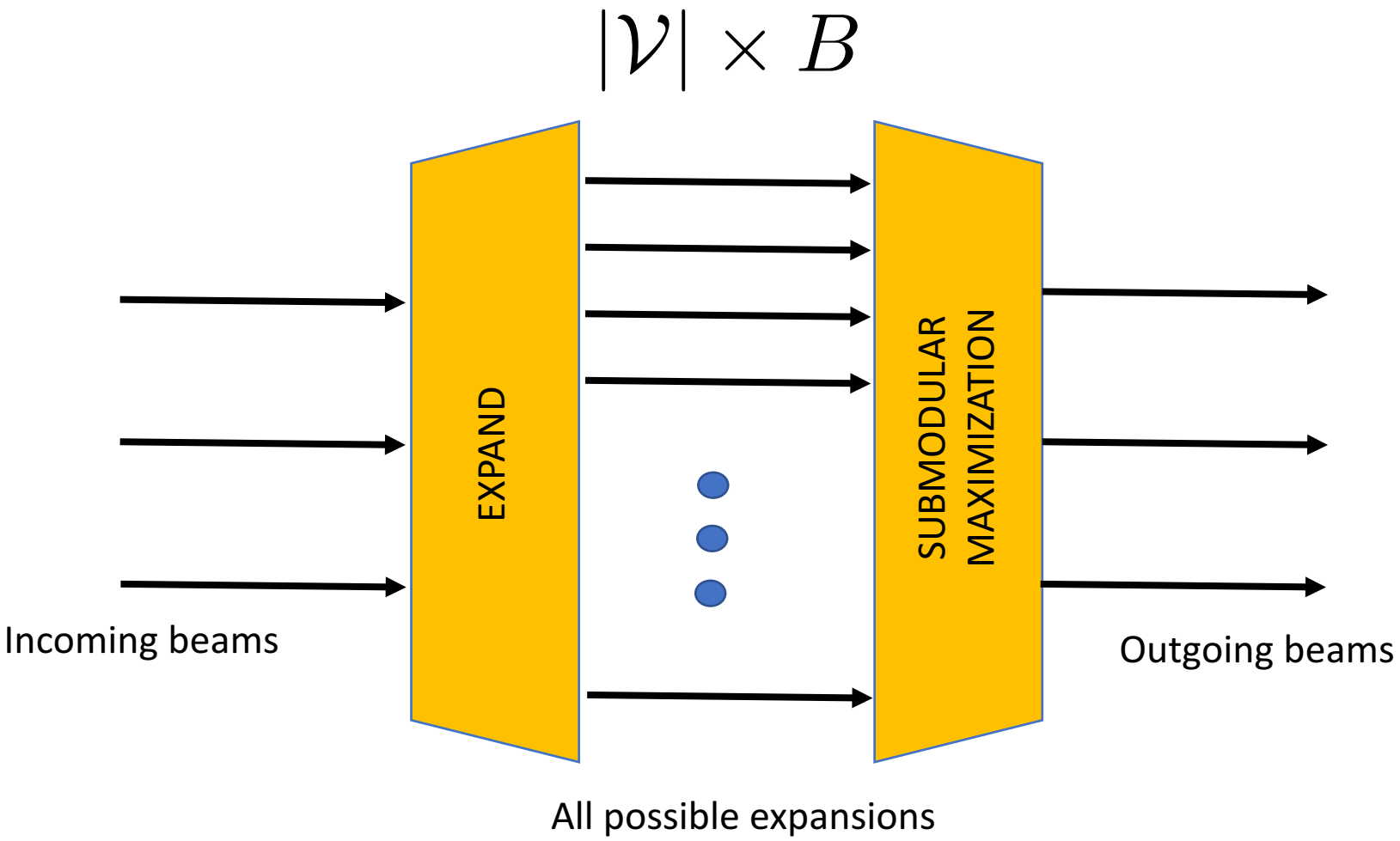


Submodular Maximization for Subset Selection



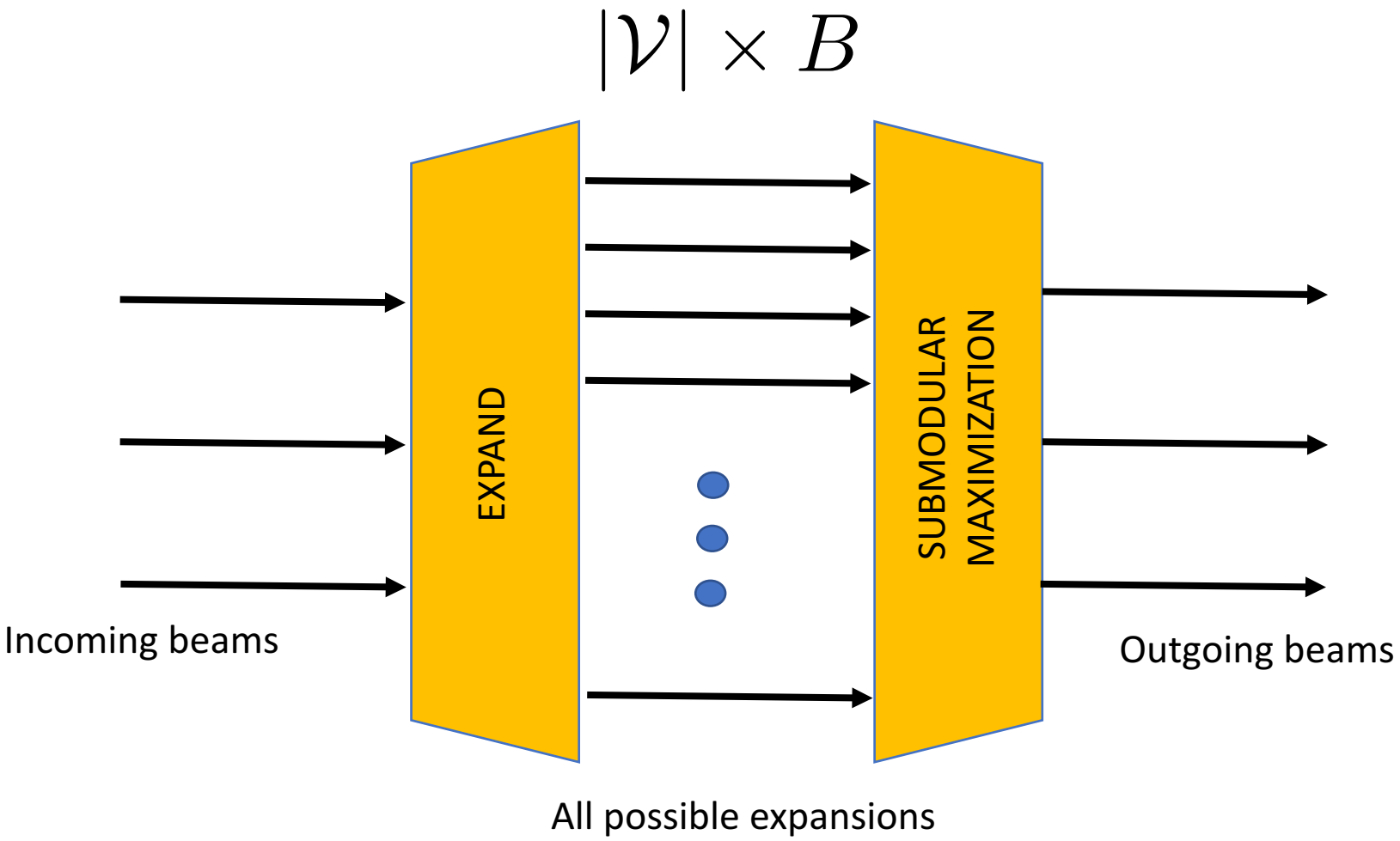
- Naturally models coverage, promoting diversity

Submodular Maximization for Subset Selection



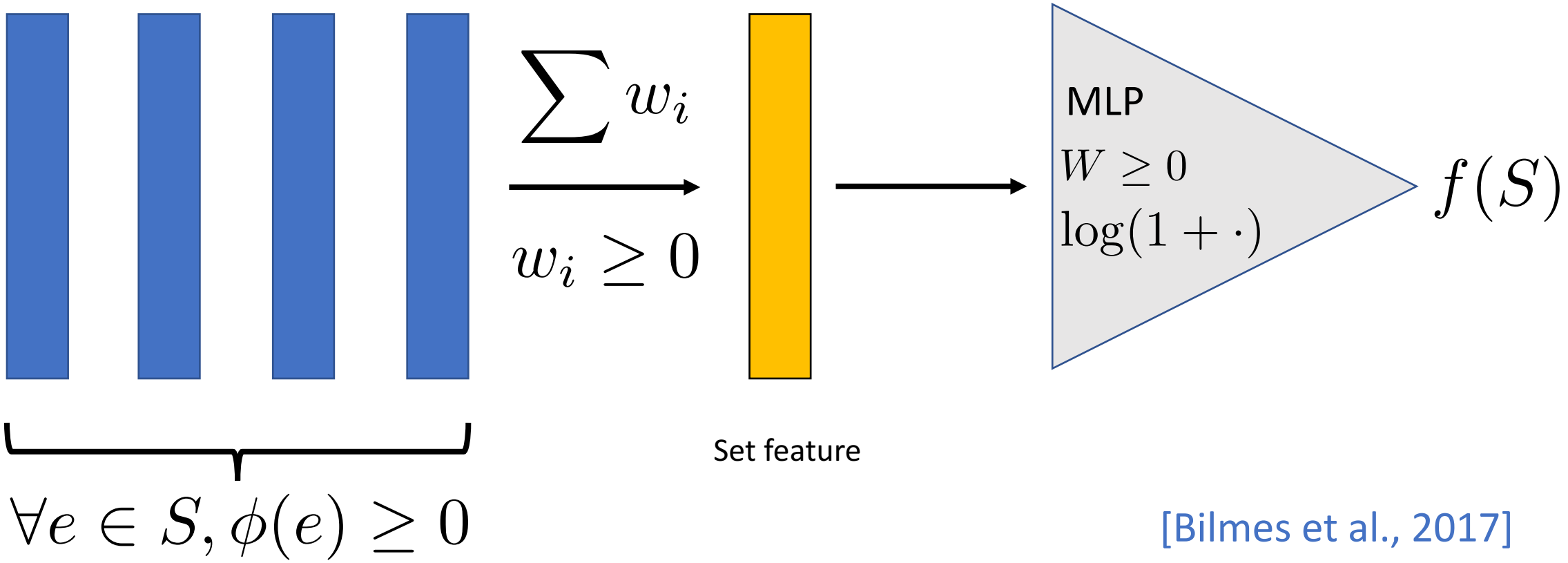
- Naturally models coverage, promoting diversity
- NP Hard!

Submodular Maximization for Subset Selection



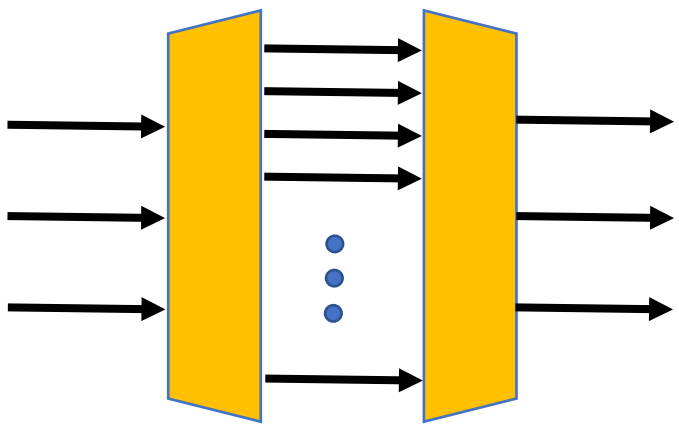
- Naturally models coverage, promoting diversity
- NP Hard!
- Greedy algorithms with approximation guarantees exist!

Learning Submodular Functions



[Bilmes et al., 2017]

∇BS (diff-BS)



FOR t = 1 to T:

1. Construct set of all possible extensions $\mathcal{Y}_{t-1} \times |\mathcal{V}|$

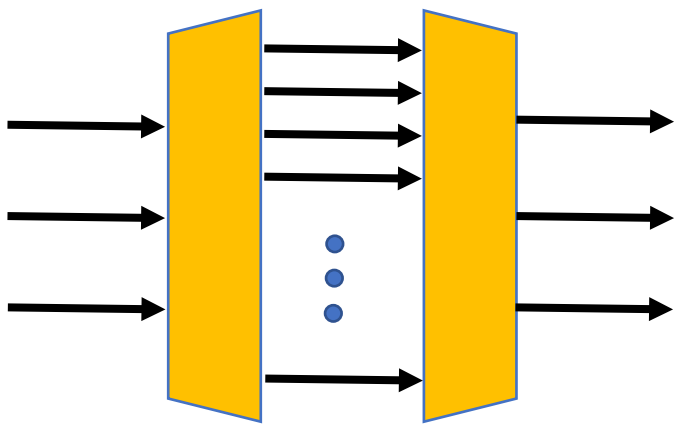
FOR k = 1 to K:

2. Compute marginal gain of each extension

3. Sample an extension proportional to marginal gain

RETURN Set of K Sequences of length T

∇BS (diff-BS)



FOR t = 1 to T:

1. Construct set of all possible extensions

$$\mathcal{Y}_{t-1} \times |\mathcal{V}|$$

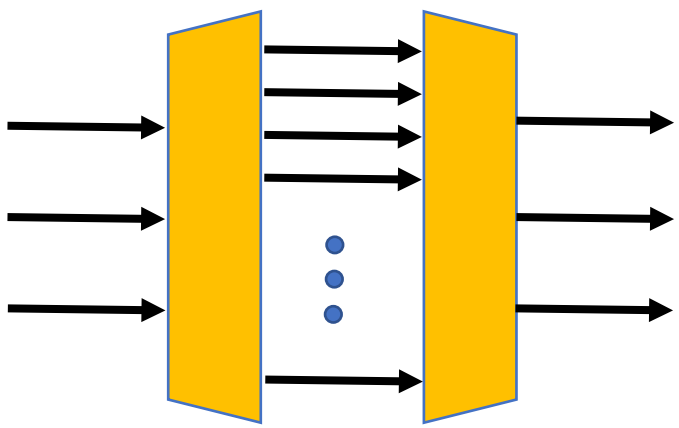
FOR k = 1 to K:

2. Compute marginal gain of each extension

3. Sample an extension proportional to marginal gain

RETURN Set of K Sequences of length T

∇BS (diff-BS)



FOR t = 1 to T:

1. Construct set of all possible extensions $\mathcal{Y}_{t-1} \times |\mathcal{V}|$

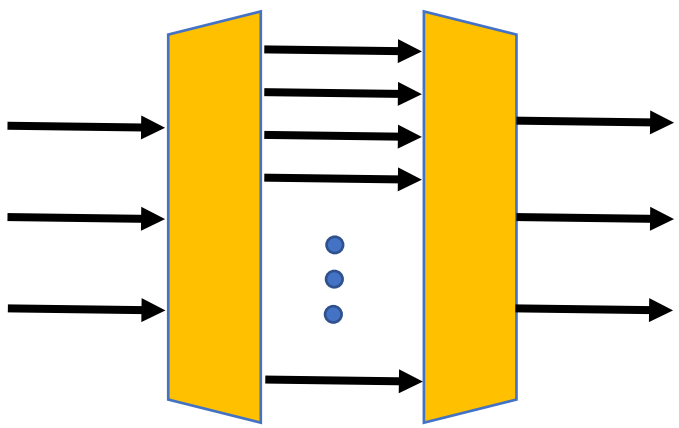
FOR k = 1 to K:

2. Compute marginal gain of each extension

3. Sample an extension proportional to marginal gain

RETURN Set of K Sequences of length T

∇BS (diff-BS)



FOR t = 1 to T:

1. Construct set of all possible extensions $\mathcal{Y}_{t-1} \times |\mathcal{V}|$

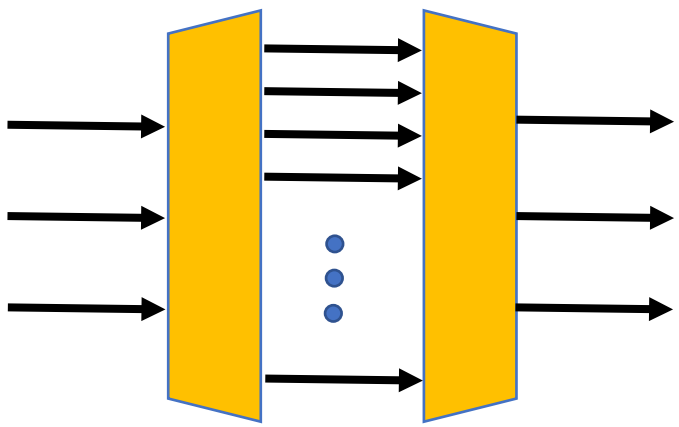
FOR k = 1 to K:

2. Compute marginal gain of each extension

3. Sample an extension proportional to marginal gain

RETURN Set of K Sequences of length T

∇BS (diff-BS)



FOR t = 1 to T:

1. Construct set of all possible extensions $\mathcal{Y}_{t-1} \times |\mathcal{V}|$

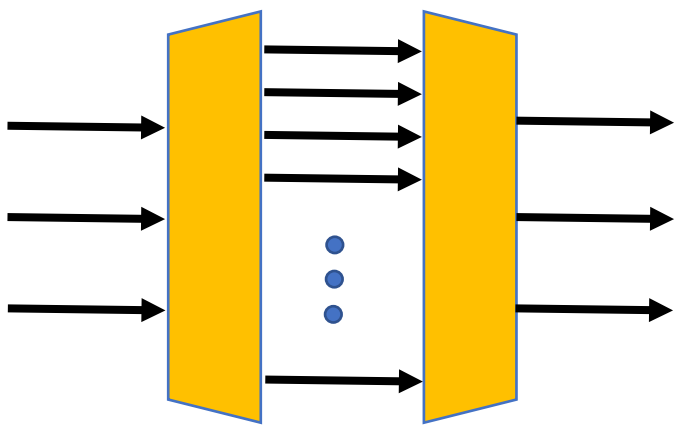
FOR k = 1 to K:

2. Compute marginal gain of each extension

3. Sample an extension proportional to marginal gain

RETURN Set of K Sequences of length T

∇BS (diff-BS)



FOR t = 1 to T:

1. Construct set of all possible extensions $\mathcal{Y}_{t-1} \times |\mathcal{V}|$

FOR k = 1 to K:

2. Compute marginal gain of each extension

3. Sample an extension proportional to marginal gain

RETURN Set of K Sequences of length T

"Set of Sequences" Level Training

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{(Y_1,\ldots,Y_T)\sim\pi(\cdot|\mathbf{x})} SET\text{-}METRIC(\mathbf{Y}|\mathbf{x})$$

“Set of Sequences” Level Training

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{(Y_1,\ldots,Y_T)\sim\pi(\cdot|\mathbf{x})} SET\text{-}METRIC(\mathbf{Y}|\mathbf{x})$$

- Set-metric?
  - Oracle, average accuracy

"Set of Sequences" Level Training

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{(Y_1,\ldots,Y_T)\sim\pi(\cdot|\mathbf{x})} SET\text{-}METRIC(\mathbf{Y}|\mathbf{x})$$

- Set-metric?
  - Oracle, average accuracy
  - Facility Location Accuracy [NEW]

"Set of Sequences" Level Training

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{(Y_1,\ldots,Y_T)\sim\pi(\cdot|\mathbf{x})} SET\text{-}METRIC(\mathbf{Y}|\mathbf{x})$$

- Set-metric?
  - Oracle, average accuracy
  - Facility Location Accuracy [NEW]
- Training?
  - Teacher Forcing if multiple annotations are available.

"Set of Sequences" Level Training

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{(Y_1,\ldots,Y_T)\sim\pi(\cdot|\mathbf{x})} SET\text{-}METRIC(\mathbf{Y}|\mathbf{x})$$

- Set-metric?
  - Oracle, average accuracy
  - Facility Location Accuracy [NEW]
- Training?
  - Teacher Forcing if multiple annotations are available.
  - Imitation Learning if expert is available

"Set of Sequences" Level Training

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{(Y_1,\ldots,Y_T)\sim\pi(\cdot|\mathbf{x})} SET\text{-}METRIC(\mathbf{Y}|\mathbf{x})$$

- Set-metric?
  - Oracle, average accuracy
  - Facility Location Accuracy [NEW]
- Training?
  - Teacher Forcing if multiple annotations are available
  - Imitation Learning if expert is available
  - REINFORCE to directly optimize for the set-metric

In Summary

- Novel perspective. Beam Search as Subset Selection
- Models intra-set dependencies
- Can be used with arbitrary set constraints
- No train-test or loss-evaluation mismatch
- Outperforms Beam Search and other baselines on captioning

Doesn’t scale very well with beam size (some tricks in the paper)

In Summary

- Novel perspective. Beam Search as Subset Selection
- Models intra-set dependencies
- Can be used with arbitrary set constraints
- No train-test or loss-evaluation mismatch
- Outperforms Beam Search and other baselines on captioning

Doesn’t scale very well with beam size (some tricks in the paper)

In Summary

- Novel perspective. Beam Search as Subset Selection
- Models intra-set dependencies
- Can be used with arbitrary set constraints
- No train-test or loss-evaluation mismatch
- Outperforms Beam Search and other baselines on captioning

Doesn’t scale very well with beam size (some tricks in the paper)

In Summary

- Novel perspective. Beam Search as Subset Selection
- Models intra-set dependencies
- Can be used with arbitrary set constraints
- No train-test or loss-evaluation mismatch
- Outperforms Beam Search and other baselines on captioning

Doesn’t scale very well with beam size (some tricks in the paper)

In Summary

- Novel perspective. Beam Search as Subset Selection
- Models intra-set dependencies
- Can be used with arbitrary set constraints
- No train-test or loss-evaluation mismatch
- Outperforms Beam Search and other baselines on captioning

Doesn’t scale very well with beam size (some tricks in the paper)

In Summary

- Novel perspective. Beam Search as Subset Selection
- Models intra-set dependencies
- Can be used with arbitrary set constraints
- No train-test or loss-evaluation mismatch
- Outperforms Beam Search and other baselines on captioning

Doesn’t scale very well with beam size (some tricks in the paper)

Poster: Pacific Ballroom #48
June 13th 6:30 pm

Paper: http://proceedings.mlr.press/v97/kalyan19a.html
Code: https://github.com/ashwinkalyan/diff-bs