Enhancing Contrastive Learning using Optimal Transport

Improving Negative Sampling for Multi-modal Contrastive Learning

Aishwarya Kumaran, Akash Mittal, Stephen Owesney

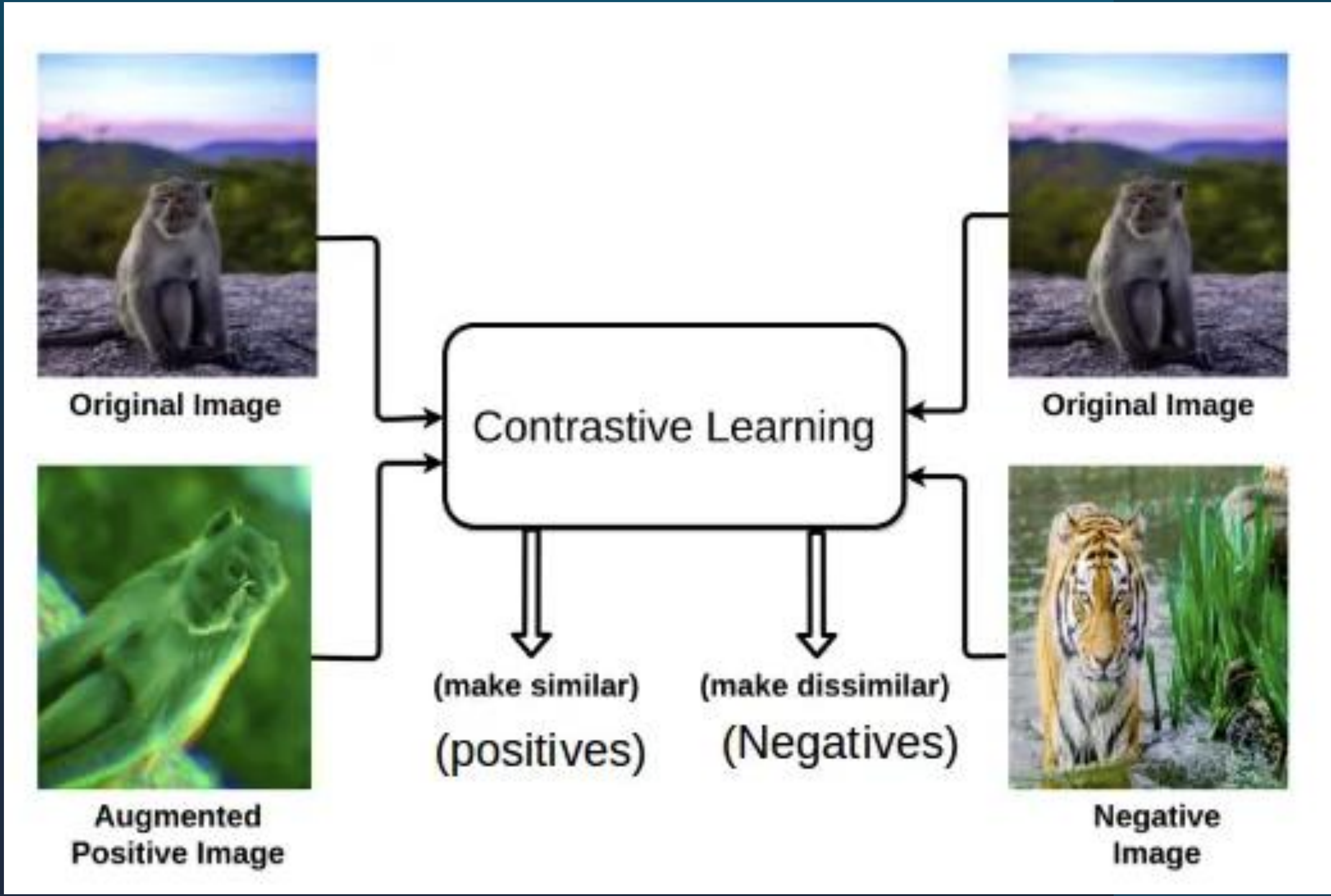
Introduction to Contrastive Learning

Contrastive Learning:

- Self-supervised learning technique
- Aims to learn rich representations by..
  - Aligning similar (positive)
  - Distinguishing dissimilar (negative)

Common Applications:

- Aligning multi-modal data (especially image-text pairs)
- Models like CLIP and ALIGN have achieved significant success


Original Image
Augmented Positive Image
Contrastive Learning
(make similar)
(positives)
(make dissimilar)
(Negatives)
Original Image
Negative Image

Why Enhance Contrastive Learning?

Key Challenges:

- Generation of diverse negative samples is crucial but challenging
- Current methods often lead to trivial / redundant negative examples

Problem with Negative Sampling:

- Poorly sampled negatives result in suboptimal representations
- Limits model generalization
- Reduces effectiveness of learned features

Our Goal:

- Improve the diversity and quality of negative samples using OT

Problem Statement

Objective:

- To enhance the contrastive learning process by computing an entropy-regularized OT plan
- Generate diverse and challenging negative examples while aligning modalities

Approach:

- Utilize a combinatorial OT algorithm to dynamically adjust the OT plan
- Evolving OT plan ensures stable alignment across batches
- Helps the model learn better representations

Why Optimal Transport?

Optimal Transport

A framework for distribution alignment

Enables precise alignment of positive and negative pairs

Benefits of OT:

Generates diverse negative pairs, addressing redundancy

Prevents sudden shifts in the representation space

Drawbacks of OT:

EXPENSIVE!

Existing Contrastive Learning Approaches

CLIP and ALIGN:

- Align image-text pairs by maximizing agreement between corresponding pairs
- Minimize similarity between non-matching pairs

Limitations:

- Negative sampling is often random or insufficiently diverse
- Poor model generalization
- Reduced feature discrimination

Negative Sampling in Existing Models

SimCLR:

- Uses random sampling of negatives from current batches
- Limitation: Can still lead to redundancy

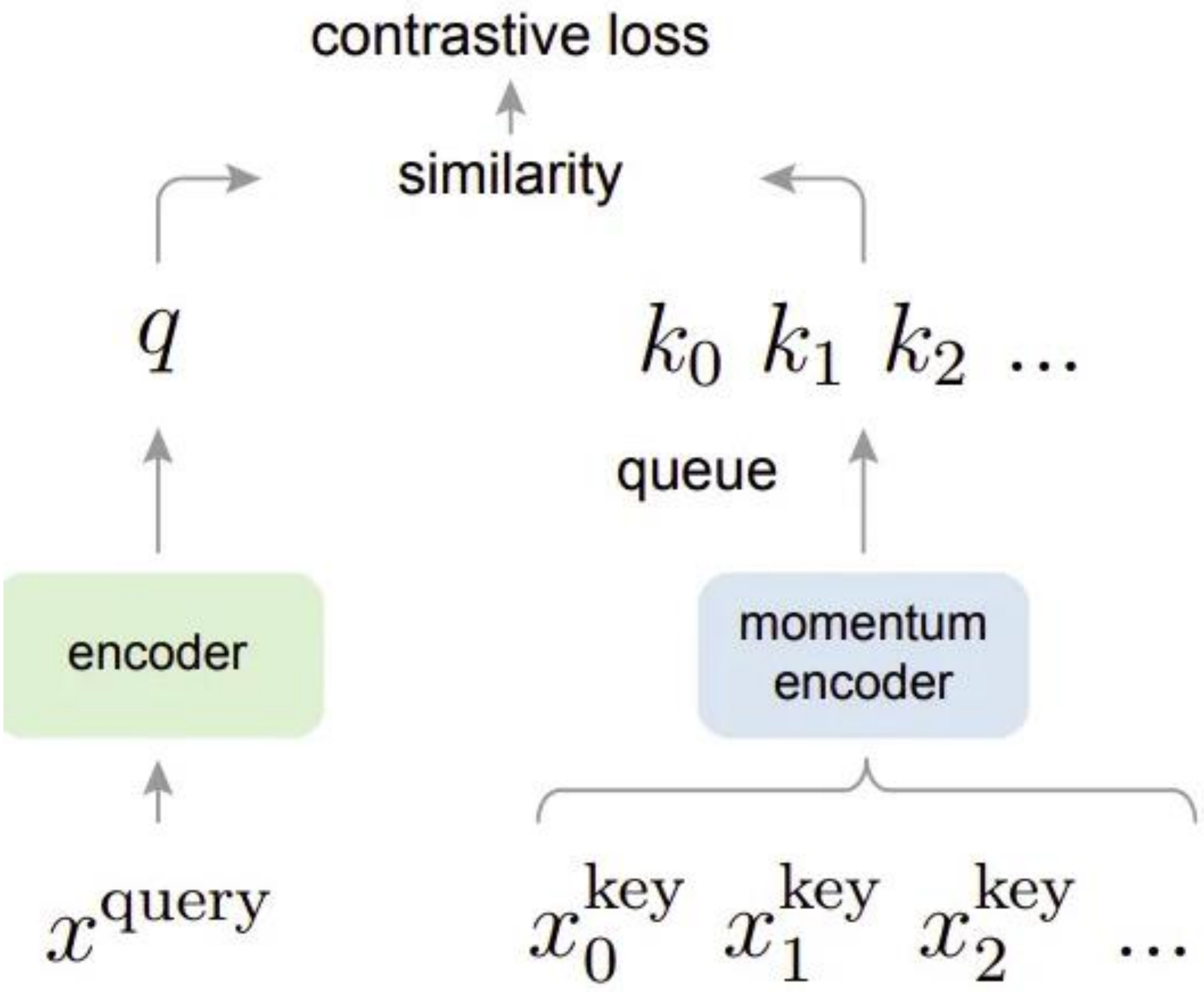
MoCo:

- Introduces a momentum encoder and dynamic queue to store negatives across batches
- Provides more diverse negatives but relies on stability of queue

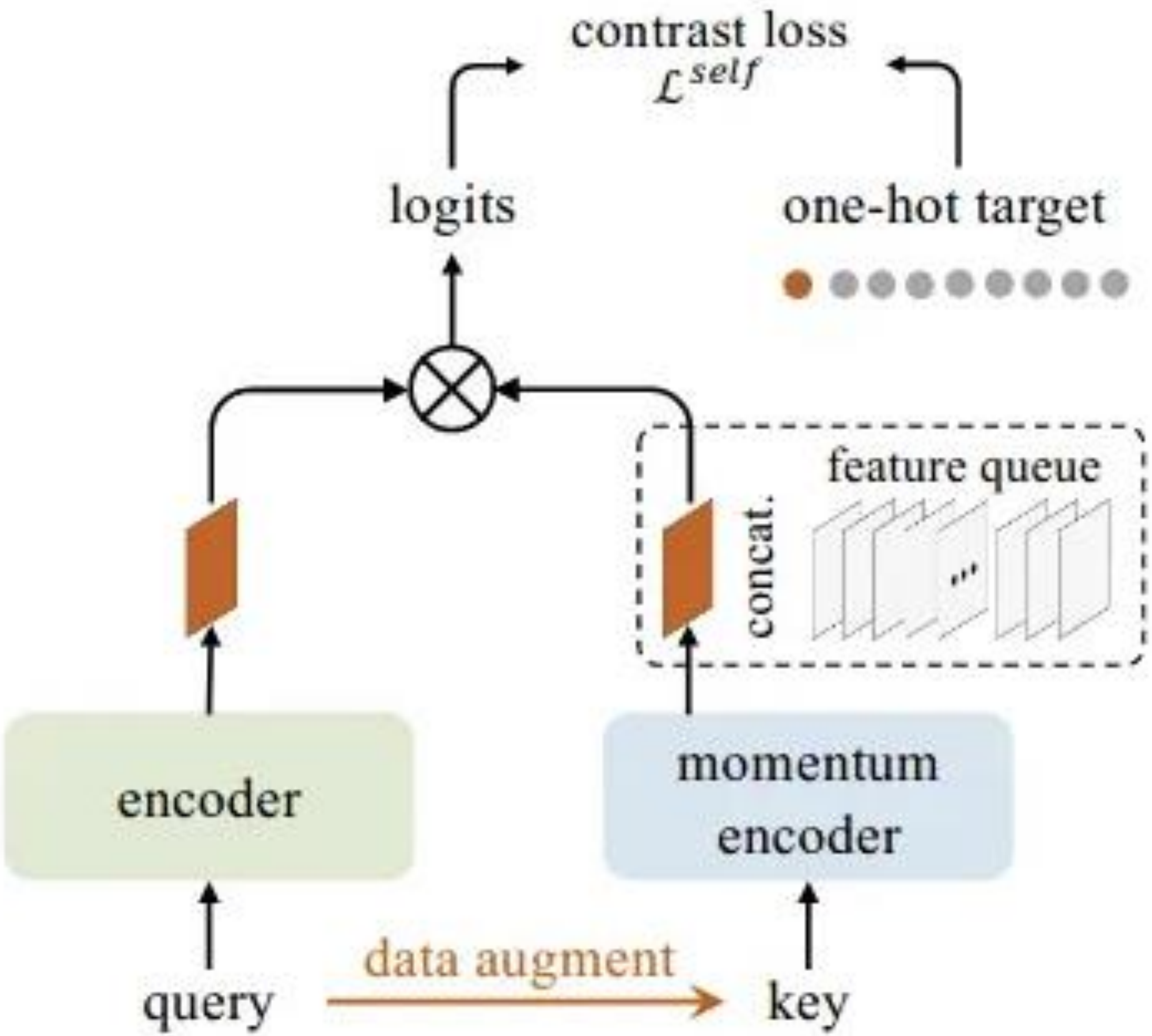
Our Approach:

- OT Plan offers more structured and evolving negative sampling
- Addresses MoCo’s limitations

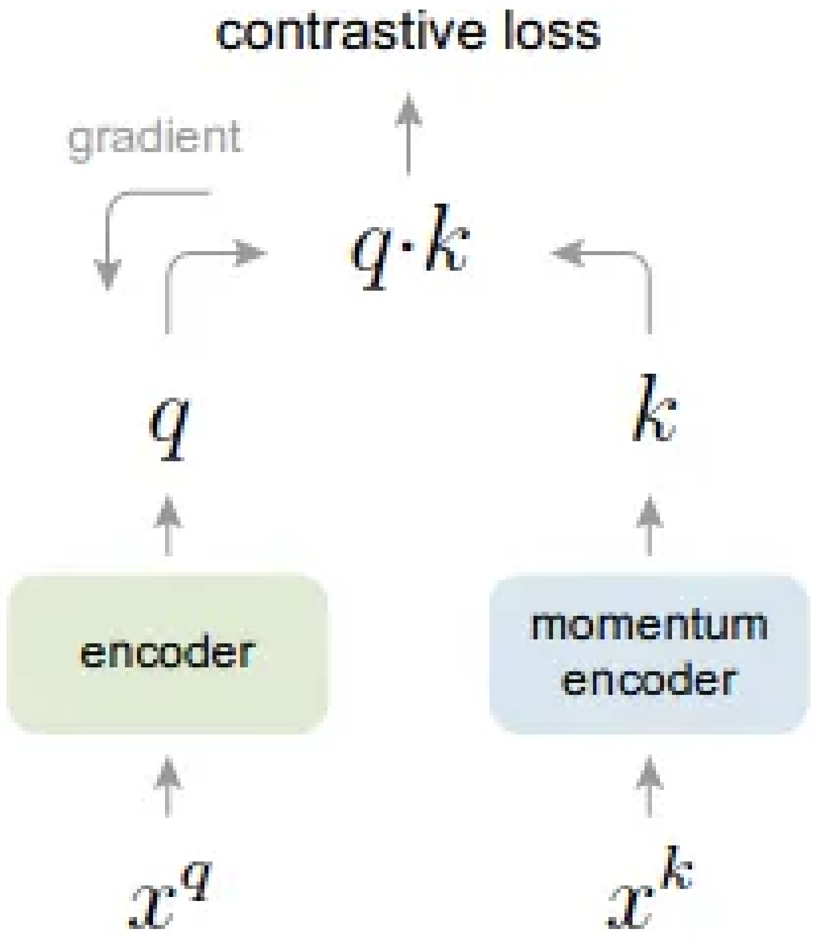
MoCo’s Architecture



MoCo's Architecture



Momentum Encoder Architecture with a gradient on the encoder branch only



Our Methodology: Overview

We introduce a contrastive learning framework leveraging OT

Challenges in Existing Methods:

- Overfitting on easy negatives
- Limited diversity in negative samples

Our Solution:

- Use OT to generate high-quality, diverse negative samples
- Preserve representation diversity through caching OT plans
- Regularize using Sinkhorn Divergence for stability

Feature Extraction: Image & Text Encoders

Input:

- A batch of images:
- A batch of corresponding text descriptions:

Encoding:

- Visual Features: Extracted using ResNet
- Textual Features: Extracted using BERT

Cross-Modal OT Plan Computation

Accumulate visual and text features over Δ batches

Compute the Optimal Transport (OT) plan:

Input: V (visual) and T (text) features.

Algorithm: Based on combinatorial approach to generate initial OT plan σ

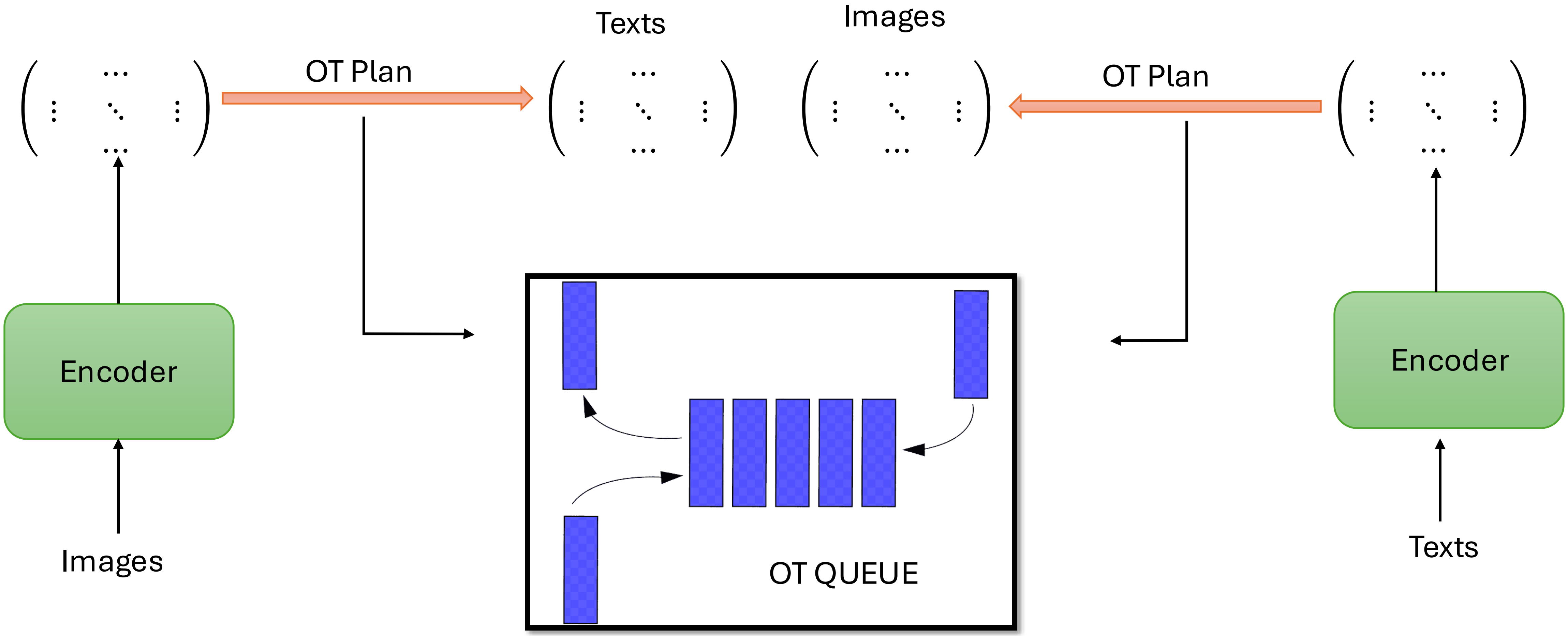
Key Insight:

OT plan is cached

Allows for revisiting previous samples

Retains rich structural details across time

OTCO Architecture for Generating Negative Pairs



Sinkhorn Divergence for Entropy Regularization

Objective:

- Stabilize and regularize the Optimal Transport (OT) plan across batches for smooth feature alignment.

Key Steps

$$\sigma' = argmin_{\sigma} \sum_{i,j} \sigma_{ij} C(v_i, t_j) + \epsilon \sum_{i,j} \sigma_{ij} \log(\sigma_{ij})$$

- $C(v_i, t_j)$ = Cost between visual and text
- ϵ = Parameter that controls the level of smoothness

Contrastive Learning with InfoNCE Loss

- Use Common Contrastive Loss InfoNCE Loss:

$$\mathcal{L}_q = -\log \frac{\exp\left(q.k_r/\tau\right)}{\sum_{i=0}^{K} \exp\left(q.k_i/\tau\right)}$$

- For each visual image v_i and text t_i
- OT transformed image $v_i{}'$ and text $t_i{}'$

Positive Pairs:

- Each pair of aligned image and text features and their respective transformations
- $(v_i, t_i), (v'_i, t_i)$ *and* (v_i, t'_i) are the positive pairs

Negative Pairs:

- Each misaligned pair of images and text $(v_i, t_j), (v'_i, t_j)$ *and* (v_i, t'_j)

Periodic Updates to the OT Plan

After processing each set of Δ batches,

- Rerun the combinatorial algorithm and apply Sinkhorn regularization to update the OT plan σ.
- Store the OT Plan: Save the updated OT plan in a dictionary for future reference
- Storing of OT plans allows us to have a dictionary of generator functions which allows us to "infinite" negative examples.

Results

- We are still training OTCO…

- Some Challenges we had:
  - During Thanksgiving break, we faced significant delays in convergence while processing a few batches of FLICKR 30K images and captions. We think it’s because of our OT solver.

  - We have changed our OT solver to an additive approximation OT solver a week back and running our experiments again.

Future Work

- Results required to move forwards
- Our original idea had a unimodal OT plan that would create a way to rollback between batches.
- Further exploration of setup


Thank You

Questions