Dual Encoding^{++}: Optimization of Text-Video Retrieval via Fine-tuning and Pruning

Dongyoun Lee, Dong-hun Lee, Vani Priyanka Gali, Sang-hyo Park

Kyungpook National University, School of Computer Science and Engineering

Overview

1. Introduction
2. Methods
   - Learning strategies
   - Model size reduction
3. Experiments and Result
4. Conclusion

Introduction

- Text-video retrieval is the task of searching relevant videos when candidate videos and text are given, or vice versa.
- Video encoder – creates embedding for videos
- Text encoder – creates embedding for texts
- Common space learning
  - Each embedding is projected into the common space.
  - calculates the similarity between embeddings

Introduction

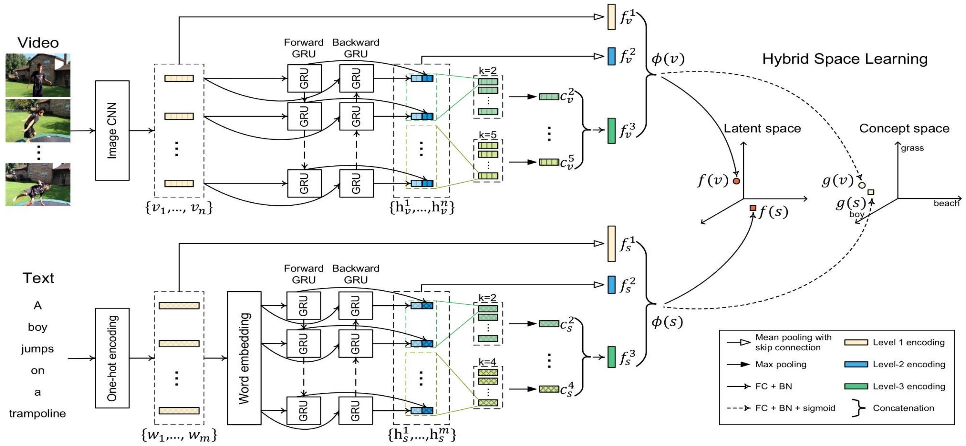
- Challenges
    - High complexity - Upcoming models are gradually becoming heavier.
    - High computational cost
    - Slow learning speed
    - A large amount of data

Introduction

- Challenges
  - High complexity - Upcoming models are gradually becoming heavier.
  - High computational cost
  - Slow learning speed
  - A large amount of data

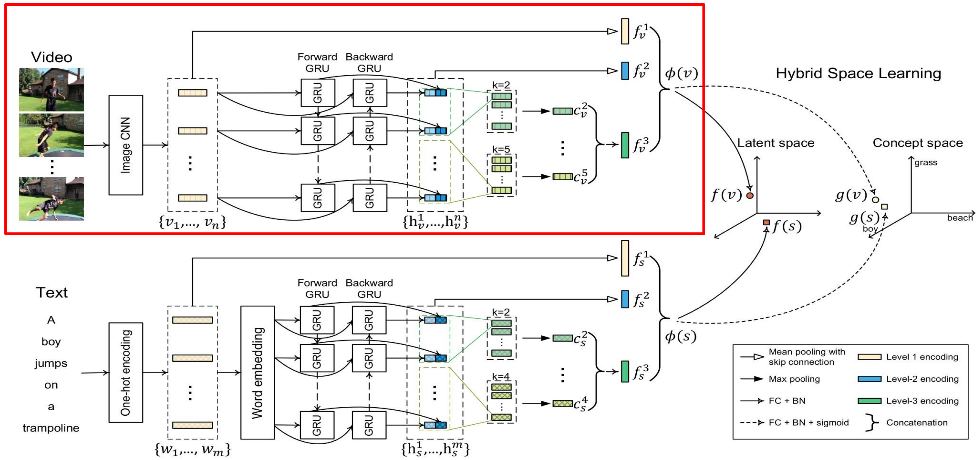
- Effects of solving the above challenges
  - Reduce computation and inference time
  - Alleviate memory storage and execution burdens
  - Make real-time text-video retrieval more feasible

Introduction



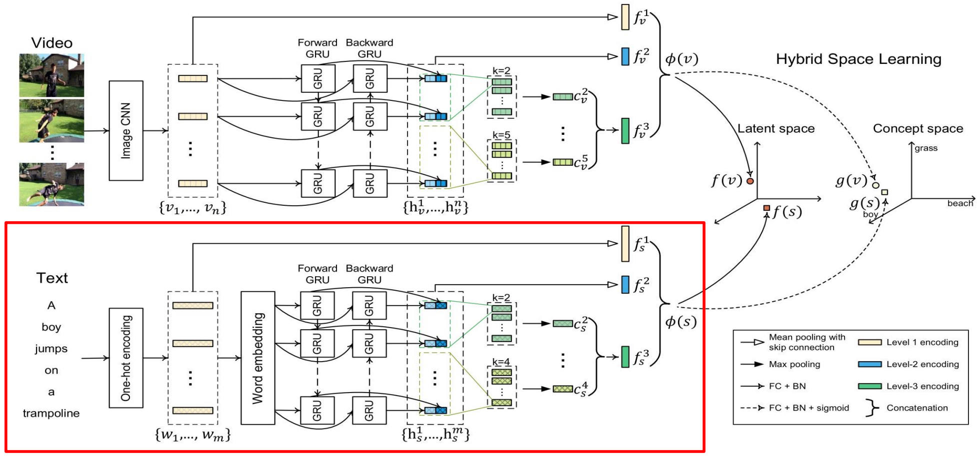
- Optimized existing text-video retrieval model, called Dual Encoding

Introduction



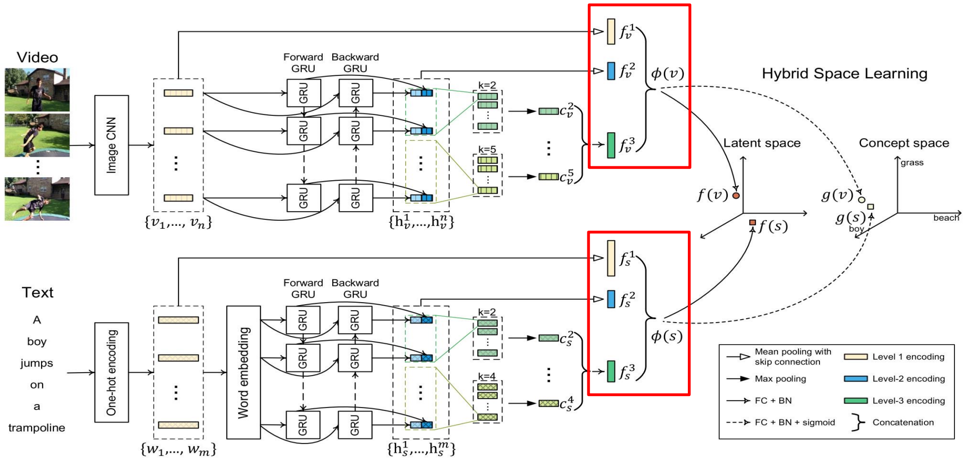
- Optimized existing text-video retrieval model, called Dual Encoding
- Dual Encoding
    - Video encoder – pre-trained ImageNet embedding, Bi-GRU, 1D Conv (4 diff. kernel sizes)

Introduction



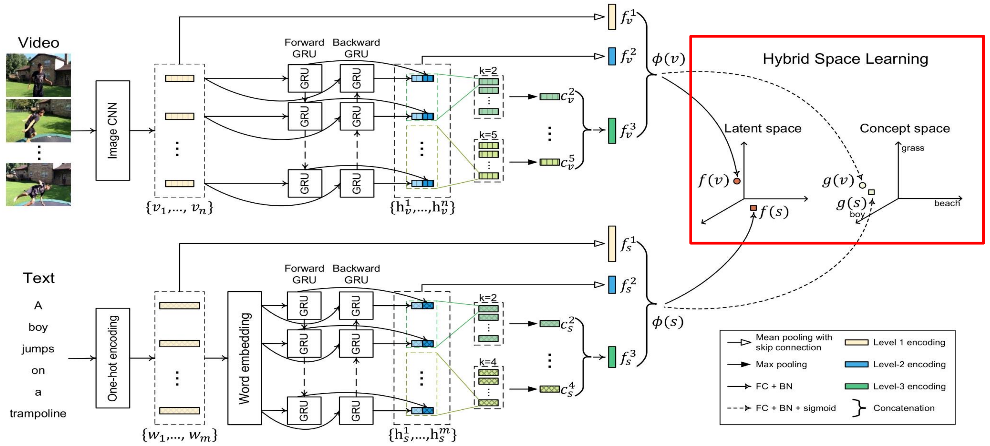
- Optimized existing text-video retrieval model, called Dual Encoding
- Dual Encoding
  - Video encoder – pre-trained ImageNet embedding, Bi-GRU, 1D Conv (4 diff. kernel sizes)
  - Text encoder – Bag of Words, Bi-GRU, 1D Conv (3 diff. kernel sizes)

Introduction



- Optimized existing text-video retrieval model, called Dual Encoding
- Dual Encoding
  - Video encoder – pre-trained ImageNet embedding, Bi-GRU, 1D Conv (4 diff. kernel sizes)
  - Text encoder – Bag of Words, Bi-GRU, 1D Conv (3 diff. kernel sizes)
  - Both encoders utilize three skip connections to concatenate embeddings at each level

Introduction



- Optimized existing text-video retrieval model, called Dual Encoding
- Dual Encoding
  - Video encoder – pre-trained ImageNet embedding, Bi-GRU, 1D Conv (4 diff. kernel sizes)
  - Text encoder – Bag of Words, Bi-GRU, 1D Conv (3 diff. kernel sizes)
  - Both encoders utilize three skip connections to concatenate embeddings at each level
  - Hybrid space learning – employs both concept and latent features

Methods

Learning strategies

Model size reduction

Methods

Learning strategies

- PAD token in text encoder, α: space balancing hyper-parameter, Batch loss calculation, Validation performance metric, Learning rate scheduling
- Improve the model's performance

Model size reduction

Methods

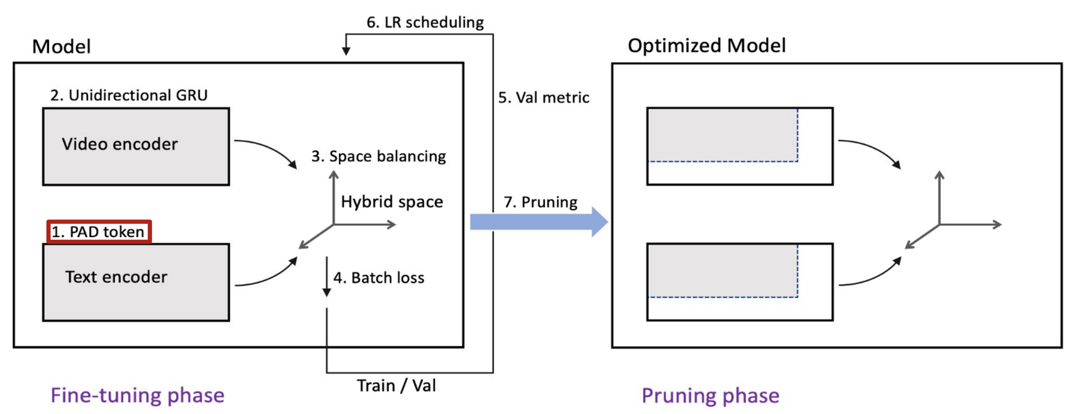
Learning strategies

- PAD token in text encoder, α: space balancing hyper-parameter, Batch loss calculation, Validation performance metric, Learning rate scheduling
- Improve the model's performance

Model size reduction

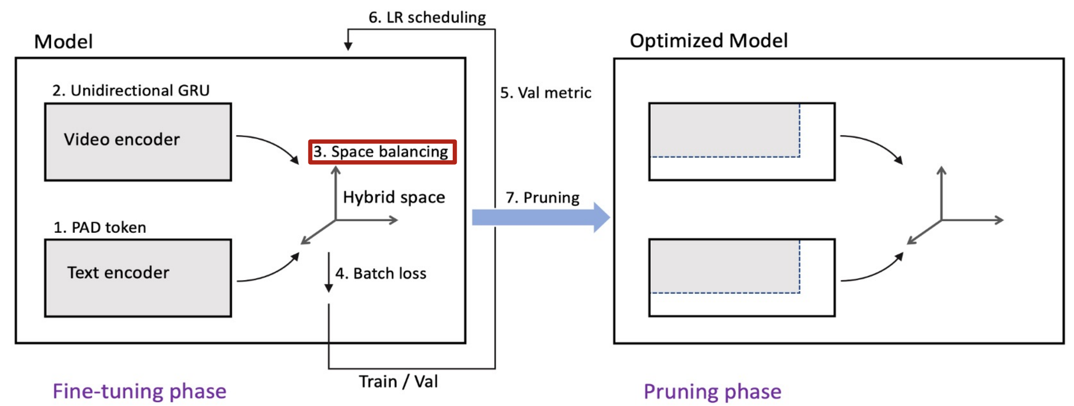
- Unidirectional GRU in video encoder, Pruning
- Reduce the model size

Learning strategies - PAD token in text encoder



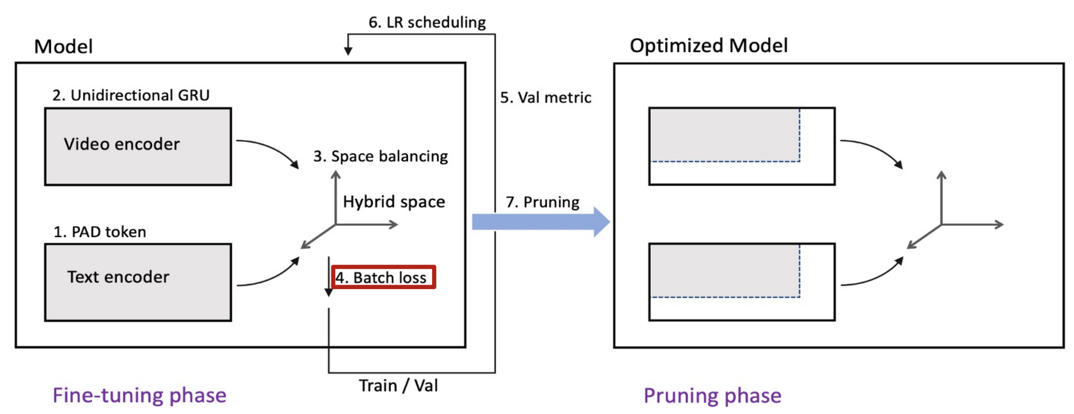
- PAD token is initialized with zeros instead of random initialization with a uniform distribution.
- These embeddings are not updated during training.

Learning strategies - α: space balancing hyper-parameter



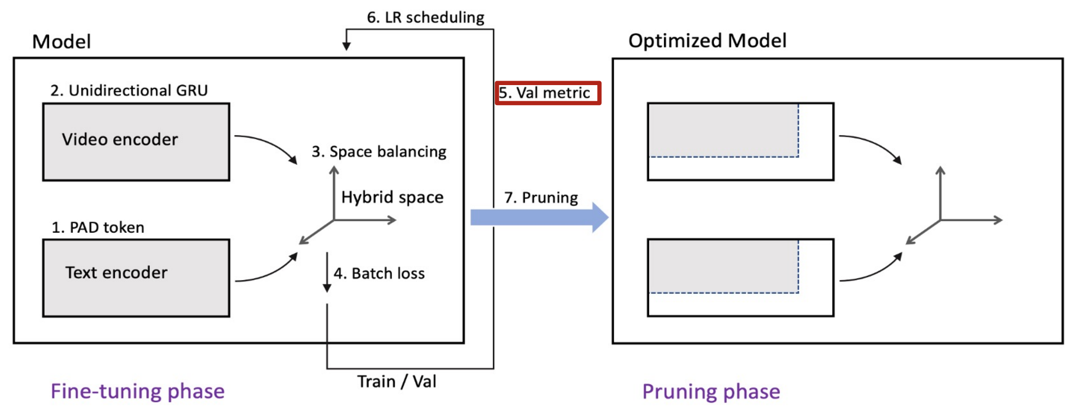
- $sim(v,s) = \alpha sim_{lat}(v,s) + (1-\alpha) sim_{con}(v,s)$
- The hyper-parameter α is adjusted to 0.2 from 0.6.
- This notes that the concept space learning is also significant.

Learning strategies - Batch loss calculation



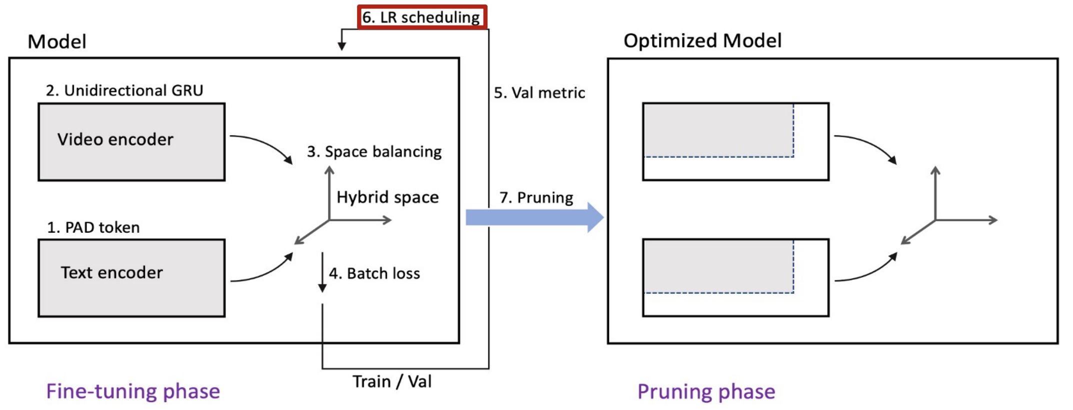
- The mean of sample loss is substituted for a sum of sample loss in batch loss calculation.
- It is well-known that the averaging removes the dependency on batch size.

Learning strategies - Validation performance metric



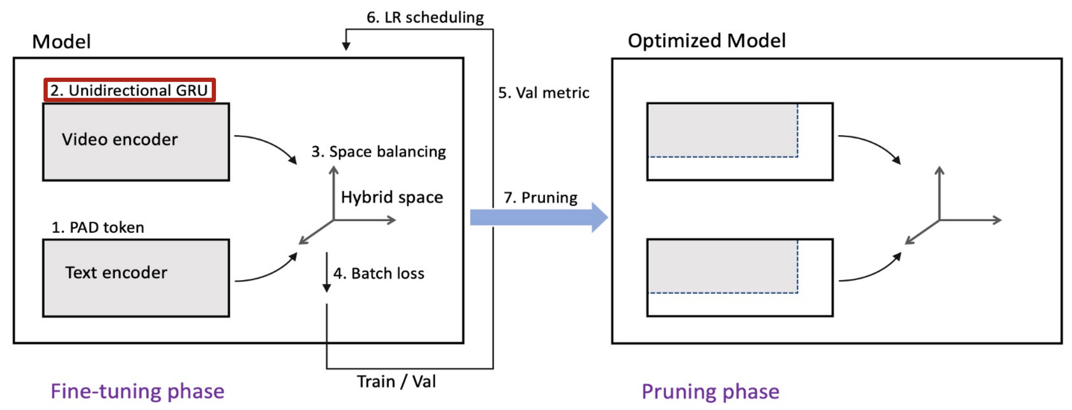
- Validation performance affects learning rate adjustment and early stopping.
- T2V Sum R is substituted for validation loss.
  - Text-to-video retrieval is usually considered more complicated than video-to-text retrieval.
  - It is a metric used to evaluate performance on a test set.

Learning strategies - Learning rate scheduling



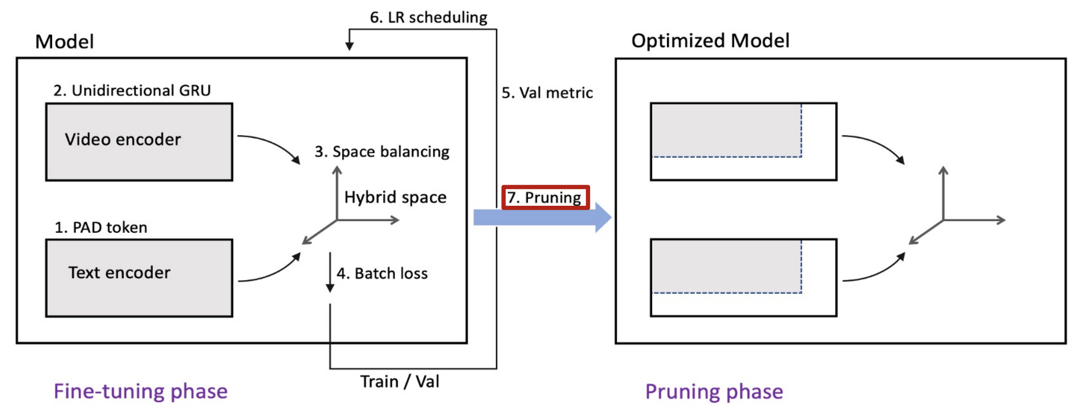
- A different method is used when decreasing the learning rate significantly.
- If the validation performance does not increase for three successive epochs,
    1. Previously saved best model state is loaded.
    2. A decay rate of 0.5 is applied to a learning rate.

Model size reduction - Unidirectional GRU



- Assumed that a high amount of duplicated information exists concerning the temporal dimension.
- Unidirectional GRU is substituted for bidirectional GRU in the video encoder.
- The model size is reduced to 0.83 of the original model.

Model size reduction - Pruning



- Global unstructured L1 pruning is used.
- Global pruning is selected because of the adaptability and efficiency of network-wide pruning.
- Unstructured pruning is preferred since the developed model is not very deep, and most of its layers and structures are important.
- L1 pruning is chosen rather than random pruning due to the credibility of its magnitude-based approach.

Experiments and result

	Size Ratio	V2T R@1	V2T R@5	V2T R@10	V2T mean r	V2T sum R	T2V R@1	T2V R@5	T2V R@10	T2V mean r	T2V sum R
Dual Encoding (DE) (baseline)	1	21.6	45.9	58.5		126	11.8	30.6	41.8		84.2
[A] DE + learning strategies w/o LR scheduling	1	20.47	45.18	57.86	46.56	123.51	12.12	32.05	43.51	113.75	87.68
[B] DE + learning strategies (DE^+)	1	20.84	46.42	58.86	43.72	126.12	12.30	32.26	43.71	116.41	88.28
[C] DE^+ + uniGRU	0.83	20.70	45.65	59.33	42.95	125.69	12.09	31.92	43.23	116.98	87.24
[D] DE^+ + prune w	0.66	20.47	45.35	58.26	45.57	124.08	12.23	31.99	43.45	118.47	87.66
[E] DE^+ + uniGRU + prune weight (DE^{++} 1)	0.66	20.57	46.19	58.63	42.30	125.38	12	31.88	43.22	117.31	87.10
[F] DE^+ + uniGRU + prune weight & bias (DE^{++} 2)	0.66	20.80	46.19	58.60	42.45	125.59	11.98	31.85	43.21	117.28	87.05

- The official split of the MSR-VTT dataset is used.
- Larger R@Ks and sum R, and smaller mean r represent better performance.
- [B] demonstrates considerable improvement in both V2T and T2V sum R compared to *DE* and [A].
- [C] displays only a minor performance decline from [B] despite the reduced model size.
- Final optimized models, [E] and [F], have smaller sizes and better overall performance than *DE*.

Conclusion

- Optimized Dual Encoding, conducting experiments on the MSR-VTT dataset.
- Several learning strategies for fine-tuning the model can improve the performance.
    - highlights the significance of a modified LR scheduling when applied altogether with suggested learning strategies.
- Changing the model architecture and applying specific pruning can minimize the model size without significantly sacrificing the performance.

Thank you!