DAR:
Augmenting Document Representations for Dense Retrieval with Interpolation and Perturbation

최성욱

목차

Abstract

- Why do we need to use this proposed model?

1. Simple and Efficiency Data Augmentation Technique

2. Just Manipulate Document Representation using Interpolation and Perturbation

Introduction

- ## Basic Augmentation

1. Generating Query From Generative Model

2. Interpolation Data From Another Dataset

Drawbacks

1. **High Cost to Generate Extra Pair**
2. Extra Training Step to train extra pair
3. Biased Variation to Query (or Document)

Introduction

- ## Proposed Augmentation

1. Visualized Labeled Embedding and Unlabeled Embedding

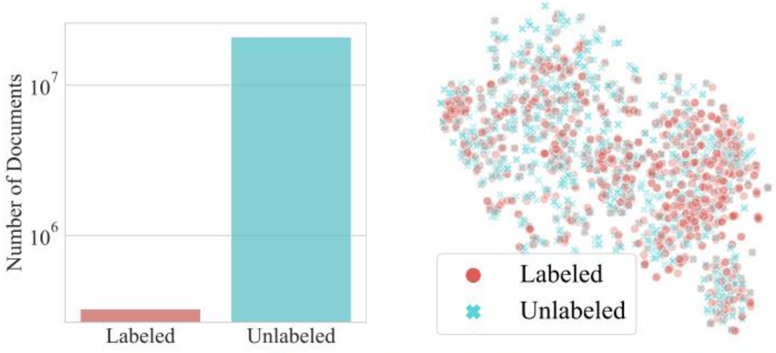


Figure 1: (Left) The number of labeled and unlabeled documents for the Natural Question dataset. (Right) T-SNE (Maaten and Hinton, 2008) visualization of randomly sampled document representations from the DPR model.

2. Interpolation two different document representations associated with the labeled query

3. Stochastically Perturbation the representation of labeled documents with a dropout mask

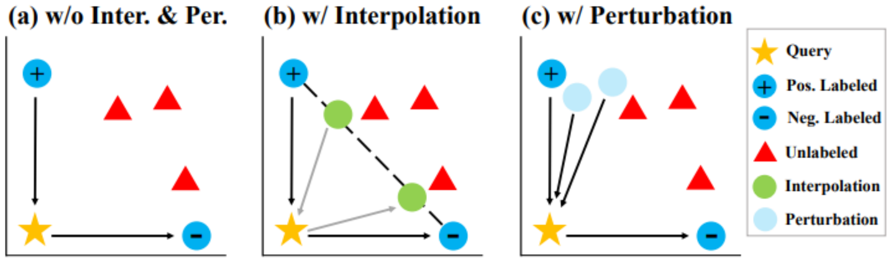


Figure 2: Our document augmenting schemes of interpolation and perturbation on a dense representation space. Pos. and Neg. denote positive and negative documents to the query.

Method

- Interpolation with Mixup

$$\tilde{\boldsymbol{d}} = \lambda \boldsymbol{d}^{+} + (1 - \lambda)\boldsymbol{d}^{-}, \quad (3)$$

where $\tilde{\boldsymbol{d}}$ is the mixed representation of positive and negative documents for the given query q, and $\lambda \in [0, 1]$.

1. Augmenting the document representation located between two labeled documents
2. Calculate Similarity $\mathrm{sim}(\boldsymbol{q}, \tilde{\boldsymbol{d}})$
3. Calculate Loss (Cross Entropy) and Add it to Original Loss

$$\min_{\theta} \sum_{(q,d^{+}) \in \tau^{+}} \sum_{(q,d^{-}) \in \tau^{-}} \mathcal{L}(f(q, d^{+}), f(q, d^{-})), \quad (2)$$

Method

- Stochastic Perturbation with Dropout

1. Randomly Mask the representation of the labeled document with dropout

2. Mask sampling using Bernoulli Distribution

3. Extra Interpolation between (perturbed document, query) and (negative document, query)

$$\{(\boldsymbol{q}, \boldsymbol{d}_i^+)\}_{i=1}^{i=n} \qquad (q, d^-)$$

Method

- Efficiency

1. Original Augmentation

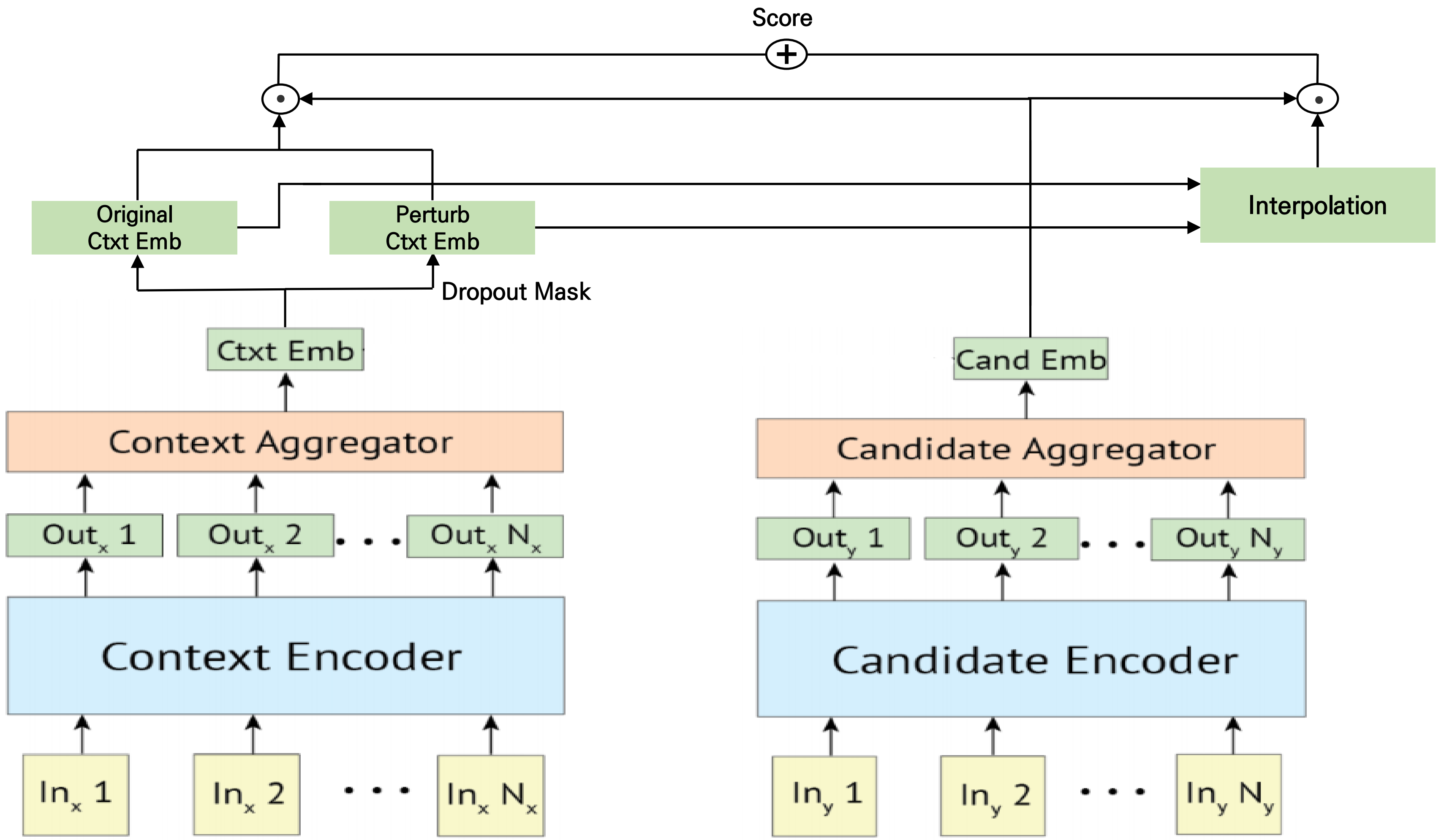
   Extra Data -> Transfer it to Model -> Extra Train Step (High Cost)

2. Just Manipulate already obtained representation of documents

3. Don't have to newly generate documents texts and to forward generated documents into model

Method

- ## Model Structure



Experiments

Experimental Setup

Passage	About 20 millions of Wikipedia Passages
Evaluate Dataset	NQ Dataset, TQA Dataset
Training Epoch	25
Batch Size	32
Negative Sampling Strategy	In-Batch Negative Sampling with Soft Negative Sample
λ	Uniform Distribution
Dropout Ratio for Perturbation	Bernoulli Distribution ($p = 0.1$)
# of Dropout Mask Sample	3 ~ 9
Learning Rate	2e-5
Optimizer	Adam

Experiments

- ## Ablation Study

1. Query Augmentation

	Natural Questions (NQ)						TriviaQA (TQA)					
	MRR	**MAP**	**T-100**	**T-20**	**T-5**	**T-1**	**MRR**	**MAP**	**T-100**	**T-20**	**T-5**	**T-1**
BM25	32.46	20.78	78.25	62.94	43.77	22.11	55.28	34.85	83.15	76.41	66.28	46.30
DPR	39.55	25.61	83.77	72.94	54.02	27.45	44.29	27.24	80.50	71.07	57.74	33.63
DPR w/ QA	40.00	24.93	83.46	72.13	55.46	27.67	46.27	28.08	80.76	71.88	59.14	35.90
DPR w/ DA	41.28	26.60	83.68	72.83	55.51	29.31	46.08	27.82	80.42	71.55	58.64	35.85
DPR w/ AR	41.18	26.04	83.60	73.41	55.51	29.11	45.13	27.57	80.65	71.68	58.09	34.52
DAR (Ours)	**42.92**	**27.12**	**84.18**	**75.04**	**57.62**	**30.42**	**47.32**	**28.70**	**81.30**	**72.66**	**59.88**	**36.94**
QAR (Ours)	**43.09**	**27.64**	**84.21**	**74.76**	**57.51**	**31.25**	**47.21**	**29.00**	**80.91**	**72.12**	**59.94**	**36.92**

2. Effectiveness of Interpolation & Perturbation

	MRR	**MAP**	**T-20**	**T-5**
DAR (Ours)	**42.92**	**27.12**	**75.04**	**57.62**
w/o Perturbation	41.26	26.19	73.68	55.37
w/o Interpolation	40.40	25.70	73.41	55.29
DPR	39.55	25.61	72.94	54.02

3. Advanced Negative Sampling Scheme

	MRR	**MAP**	**T-100**	**T-1**
DPR+HN	53.40	33.38	84.82	43.21
DAR+HN (Ours)	**54.18**	**33.71**	**85.35**	**44.18**

Conclusion

- ## Efficiency

1. Since additional Q-P pair does not need to be delivered to the LM, Training time is greatly saved

2. Effective Training Time and Memory Usage (No increase in # of Parameters)

	Time (Min.)	Memory (MiB)
DPR	19	22,071
DPR w/ QA	41	22,071
DPR w/ DA	38	22,071
DPR w/ AR	29	38,986
DAR (Ours)	21	22,071

3. Generates twice as many additional Q-P pairs as when using QA or DA