Enhancing Dual-Encoders with Question and Answer Cross-Embeddings for Answer Retrieval

Yanmeng Wang[1], **Jun Bai**[2], **Ye Wang**[1], **Jianfei Zhang**[2], **Wenge Rong**[2]
Zongcheng Ji[1], **Shaojun Wang**[1] and **Jing Xiao**[1]
[1]Ping An Technology, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
{wangyanmeng219, wangye430, jizongcheng666,
wangshaojun851, xiaojing661}@pingan.com.cn
{ba1_jun, zhangjf, w.rong}@buaa.edu.cn

Abstract

Dual-Encoders is a promising mechanism for answer retrieval in question answering (QA) systems. Currently most conventional Dual-Encoders learn the semantic representations of questions and answers merely through matching score. Researchers proposed to introduce the QA interaction features in scoring function but at the cost of low efficiency in inference stage. To keep independent encoding of questions and answers during inference stage, variational auto-encoder is further introduced to reconstruct answers (questions) from question (answer) embeddings as an auxiliary task to enhance QA interaction in representation learning in training stage. However, the needs of text generation and answer retrieval are different, which leads to hardness in training. In this work, we propose a framework to enhance the Dual-Encoders model with question answer cross-embeddings and a novel Geometry Alignment Mechanism (GAM) to align the geometry of embeddings from Dual-Encoders with that from Cross-Encoders. Extensive experimental results show that our framework significantly improves Dual-Encoders model and outperforms the state-of-the-art method on multiple answer retrieval datasets.

1 Introduction

Answer retrieval (Surdeanu et al., 2008) is an important mechanism in question answering (QA) systems to obtain answer candidates given a new question. Currently, the most widely used framework for answer retrieval task is Dual-Encoders (Seo et al., 2019; Chang et al., 2020; Cer et al., 2018), also known as "Siamese Network" (Triantafillou et al., 2017; Das et al., 2016). The Dual-Encoders model consists of two encoders to compute the embeddings of questions and answers independently, and also a predictor to estimate the relevance by a similarity score between the two embeddings.

Recently, due to the application of advanced encoding techniques, e.g., Transformer (Vaswani et al., 2017), BERT (Devlin et al., 2019), the Dual-Encoders achieved a huge boost on the overall performance (Karpukhin et al., 2020; Maillard et al., 2021). However, there remains some room to improve since the embeddings of questions and answers are encoded separately, while the cross information between questions and answers are important for answer retrieval (Yu et al., 2020).

Many efforts have been devoted in developing more powerful scoring by considering the interactions among questions and answers. For example, Xie and Ma (2019) introduced additional word-level interaction features between questions and answers for matching degree estimation. Similarly, Humeau et al. (2020) implemented attention mechanism to extract more information when computing matching score. Though such approaches improve the scoring mechanism, the overall efficiency derived from separate and off-line embeddings of questions and answers is sacrificed to some extent.

Therefore, it deserves discussing how to achieve better trade-off for maintaining the independent encoding in inference stage. To this end, the Dual-VAEs (Shen et al., 2018) is proposed by using the question-to-question and answer-to-answer reconstruction as joint training task along with the retrieval task to improve the representation learning, which maintains the independent encoding in inference stage. However, the embeddings produced by Dual-encoders or Dual-VAEs can still only preserve isolated information for questions or answers, while cross information between questions and answers is only learned through similarity score computed by two embeddings. Those embeddings preserving isolated semantics can lead to confusing results particularly when an answer can have multiple matched questions and vice versa, which is referred as one-to-many problem (Yu et al., 2020).

To address this challenge, Yu et al. (2020) further proposed Cross-VAEs by reconstructing answers from question embeddings and reconstructing ques-

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2306–2315
November 7–11, 2021.

tions from answer embeddings. In such way, the embeddings of questions or answers preserve the cross information from matched answers or questions and improve the performance in one-to-many cases. Nevertheless, both Dual-VAEs and Cross-VAEs rely on the generation sub-task to enhance the embeddings in retrieval task, while the need of text generation (the word-level joint distribution of sentences) and that of answer retrieval (the sentence-level matching distribution of QA-pairs) are different, which are suspected to conflict in joint training (Deudon, 2018). It then brings an interesting question: is it feasible to exploit the cross information in retrieval task and keep the independence of sentence encoding in inference stage.

In this research we proposed a Cross-Encoders (details in section 3.3) as an additional guidance during Dual-Encoders training besides the similarity score. The Cross-Encoders could form comprehensive representation through cross-attention to reflect the complex relations (e.g., one-to-many) between matched questions and answers. We also developed Geometry Alignment Mechanism (details in section 3.4) as the guiding way to effectively bridge the gap between Cross-Encoders and Dual-Encoders by forcing the Dual-Encoders to mimic Cross-Encoders on the geometry (i.e., semantic feature structure) in embedding space.

The contributions of this paper are in three folds: 1) Focusing on the lack of interactions in Dual-Encoders architecture, we introduce an ENhancing Dual-encoders with CROSS-Embeddings (ENDX) framework to solve this limitation, where a Cross-Encoders model is proposed to guide the training of Dual-Encoders model; 2) To achieve such enhancement in ENDX, we propose a novel Geometry Alignment Mechanism (GAM) to align the geometry of embeddings from Dual-Encoders with that from Cross-Encoders, which models the interactions between words within question and answer. This frees the Dual-Encoders from having to encode necessary information with no access to matched sentence; 3) To validate our framework, we conduct extensive experiments and show that the proposed framework significantly improves Dual-Encoders model and outperforms the state-of-art model on multiple QA datasets.

2 Related Work

Traditional answer retrieval consists of two-stage pipeline including key words matching (BM25 (Robertson and Zaragoza, 2009)) to efficiently retrieve multiple relevant passages and re-ranking by neural network to select correct answers from retrieved results. But it may fall short here as the connection between answers and questions in context is not modelled directly, while the large document where the answer locates could be not highly relevant to the question (Ahmad et al., 2019).

To address the problem in two-stage pipeline retrieval, there is growing interest in training end-to-end retrieval systems that can efficiently surface relevant results without an intermediate document retrieval phase (Karpukhin et al., 2020; Chang et al., 2020; Ahmad et al., 2019; Seo et al., 2019; Henderson et al., 2019). In recent works (Karpukhin et al., 2020; Chang et al., 2020; Maillard et al., 2021), using dense representation learned by Dual-Encoders framework outperformed BM25 in large-scale retrieval task. Dual-Encoders can encode questions and answers independently and thus enables offline processing to support efficient online response, but there exists a bottleneck that impedes the QA alignment for lack of interaction between questions and answers in their independent encoding.

Another popular way of sentence-level representation learning is Variational AutoEncoder (VAE). By encoding sentences into latent variables and reconstructing the same sentences from corresponding latent variables, VAE compacts the joint distribution of words in sentence into latent variable. Shen et al. (2018) adopted VAE in Dual-Encoders and optimized the variational lower bound and matching loss jointly. Yu et al. (2020) proposed to reconstruct questions and answers in a crossed way to improve their interaction and allow for one-to-many projection. We do not include text reconstruction into our training goal for the difference between the need of sentence representation in reconstruction and that in answer retrieval.

Our proposed framework consists of a Dual-Encoders and a Cross-Encoders. The conventional Dual-Encoders provides the system with practicality in large-scale retrieval (Karpukhin et al., 2020; Chang et al., 2020; Maillard et al., 2021), while the Cross-Encoders has interaction between question and answer to guide the training of Dual-Encoders.

3 Methodology

3.1 Problem Definition

The answer retrieval task in this work is formalized as: given a question set S_Q and an answer set

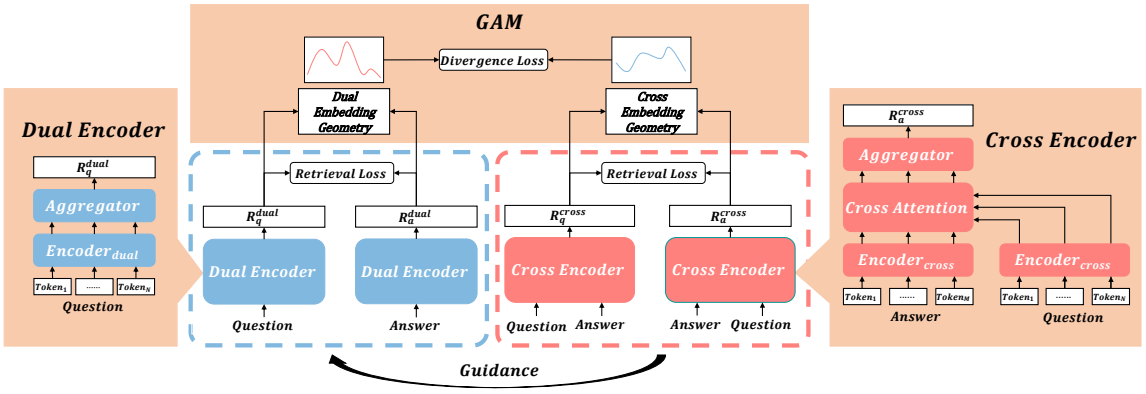


Figure 1: The overview of proposed framework that enhances Dual-Encoders with cross-embeddings, Dual-Encoders (blue) and Cross-Encoders (red) are both used for training and only Dual-Encoders is used for inference.

S_A, each sample could be represented as (q, a, y) where $q \in S_Q$ is a question, $a \in S_A$ is a sentence-level answer, and y denotes whether the answer a matches the question q or not. The target is to find the best-matched answer for the question q and a list of candidate answers $C(q) \subset S_A$.

3.2 Dual-Encoders

Our baseline model is Dual-Encoders and we refer to the sentence embedding encoded by Dual-Encoders as dual-embedding. As shown in Fig. 1, the question (answer) dual-embedding R_q^{dual} (R_a^{dual}) is processed from question (answer) text by encoder and aggregator in Dual-Encoders, where the encoder, marked as $Encoder_{dual}$ in Fig. 1, can be BERT and we employ multiple hops self-attention (Lin et al., 2017) as the aggregator in this work. The scoring function f is defined as the inner product between the dual-embeddings of question and answer: $f(q, a) = R_q^{dual} \cdot R_a^{dual}$. Intuitively, an excellent Dual-Encoders should give high scores to matched QA pairs and low scores to mismatched QA pairs. We use in-batch negatives training strategy, which is effective for learning a Dual-Encoders model (Karpukhin et al., 2020). Assuming that a mini-batch has B matched question-answer pairs, then the retrieval loss of a mini-batch is:

$$\mathcal{L}_{dual} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(R_{q_i}^{dual} \cdot R_{a_i}^{dual})}{\sum_{j=1}^{B}\exp(R_{q_i}^{dual} \cdot R_{a_j}^{dual})} \quad (1)$$

where B is the batch size; i and j are the indexes of QA pairs in a given batch.

3.3 Cross-Encoders

The cross-embeddings that involve rich question-answer interaction are obtained from the Cross-Encoders. As shown in Fig 1, the Cross-Encoders gets input from both answer and question sentences. To capture precise question-answer interaction, the matched answer (question) is used to guide the encoding of question (answer).

Let $H_q \in R^{N\times d_r}$ and $H_a \in R^{M\times d_r}$ denote the contextualized representations of words in question and answer sentences from $Encoder_{cross}$ respectively, where N and M are the number of words in question and answer sentences and d_r is the dimension of contextualized representation. A multi-head scaled dot-production attention (Vaswani et al., 2017), marked as $Cross\ Attention$ in Fig. 1, is used to refine question (answer) contextualized representation by matched answer (question). Take the refinement of question for instance, the i^{th} head is calculated as Eq. 2 and all heads are concatenated as Eq. 3 to obtain the answer-attended question representation $H_q^{'}$, then position-wise feed-forward networks (FFN) and layer normalization (LayerNorm) are used to further refine $H_q^{'}$ to obtain enhanced question contextualized representation H_q^{cross} as Eq. 4:

$$head_q^i = \text{softmax}(\frac{H_a W_q^i (H_q W_k^i)^T}{\sqrt{d_h}}) H_q W_v^i \quad (2)$$

$$H_q^{'} = [head_q^1; ...; head_q^{l_h}] W_o \quad (3)$$

$$H_q^{cross} = \text{LayerNorm}(H_q^{'} + \text{FFN}(H_q^{'})) \quad (4)$$

where $H_q^{cross} \in R^{M\times d_r}$; l_h is the number of heads; W_q^i, W_k^i, W_v^i and W_o are learnable weights. Similarly we can obtain the enhanced answer contextualized representation $H_a^{cross} \in R^{N\times d_r}$. Multi-head attention can model word-level relationships

across question and answer, and reflect the similarity between every pair of word contextualized representation across question and answer to capture the question-answer interaction and to form the comprehensive embedding of source sentence.

The sequence of H_q^{cross} and H_a^{cross} are then aggregated into fixed-length cross-embeddings R_q^{cross} and R_a^{cross}, which can precisely model the relations between questions and answers. The Cross-Encoders can be trained through loss function that is defined on a mini-batch as Eq. 5:

$$\mathcal{L}_{cross} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(R_{q_i}^{cross}\cdot R_{a_i}^{cross})}{\sum_{j=1}^{B}\exp(R_{q_i}^{cross}\cdot R_{a_j}^{cross})} \tag{5}$$

where B is the batch size; i and j are the indexes of the QA pairs in a given batch.

3.4 Geometry Alignment Mechanism

The dual-embeddings mechanism can save much response time through off-line processing while the cross-embeddings introduce early interaction and produce retrieved answer set with better relevance. To meet the gap between the dual-embeddings and cross-embeddings, regression is a direct way that can be easily thought of. However, this element-wise alignment in high dimensional space is too rigid for answer retrieval.

Inspired by the geometry-preserved dimensionality reduction for pair-wise interaction modeling proposed in SNE (Hinton and Roweis, 2002), we relax the element-wise alignment to the pair-wise alignment in the form of geometry, which is also proved to be crucial in representation learning (Passalis and Tefas, 2018). Therefore, in this research we propose the Geometry Alignment Mechanism (GAM) to align the geometry of dual-embeddings with that of cross-embeddings, which capture the question-answer interaction. Specifically, the geometry of embeddings tells who are the neighbors of a question or an answer in the embedding space. In other words, it tells which question-answer pairs, question-question pairs or answer-answer pairs are likely to be close in the feature space.

Since Dual-Encoders are not able to exploit the information from matched questions or answers, it might be difficult to accurately recreates the whole geometry of cross-embedding. Therefore we use the conditional probability converted from pair-wise dissimilarities to represent the geometry of data sample in feature space (Hinton and Roweis, 2002; Van der Maaten, 2008). The conditional probability expresses the asymmetric probability of each datapoint e_i being close to another datapoint e_j in feature space as Eq. 6:

$$p(e_j|e_i) = \frac{\exp\left(-d(e_i,e_j)\right)}{\sum_k \exp\left(-d(e_i,e_k)\right)} \tag{6}$$

where the $d(e_i, e_j)$ measures the dissimilarity between e_i and e_j.

Consequently the probability of question q_i being close to answer a_j in feature space can be described by the conditional probability $p(a_j|q_i)$. To estimate such probabilities, we can use kernel density estimation (KDE) (Scott, 1992), which replaces the negative dissimilarity function $-d(e_i, e_j)$ with a symmetric kernel function $K(e_i, e_j; \sigma^2)$ to model the similarity between e_i and e_j, where σ^2 is width. The conditional probability $p(a_j|q_i)$ of cross-embeddings $p_{cross}(a_j|q_i)$ and that of dual-embeddings $p_{dual}(a_j|q_i)$ can be estimated using a batch of samples as Eqs. 7 and 8 consequently:

$$p_{cross}(a_j|q_i) = \frac{\exp\left(K(R_{q_i}^{cross}, R_{a_j}^{cross}; 2\sigma_{c_{aq}}^2)\right)}{\sum_{k=1}^{B}\exp\left(K(R_{q_i}^{cross}, R_{a_k}^{cross}; 2\sigma_{c_{aq}}^2)\right)} \tag{7}$$

$$p_{dual}(a_j|q_i) = \frac{\exp\left(K(R_{q_i}^{dual}, R_{a_j}^{dual}; 2\sigma_{d_{aq}}^2)\right)}{\sum_{k=1}^{B}\exp\left(K(R_{q_i}^{dual}, R_{a_k}^{dual}; 2\sigma_{d_{aq}}^2)\right)} \tag{8}$$

where B is the batch size; i, j and k are the indexes of the QA pairs in a given batch.

The conditional probabilities $p(q_j|q_i)$, $p(a_j|a_i)$ can be estimated similarly. Since the conditional probability is asymmetric, $p(q_j|a_i)$ is also needed. One of the most natural choices of the kernel for kernel density estimation is Gaussian kernel defined as Eq. 9, while it suffers from the need of well-tuned width (Turlach, 1993):

$$K_{Gaussian}(e_i, e_j; \sigma) = \exp\left(-\frac{\|e_i - e_j\|_2^2}{\sigma}\right) \tag{9}$$

To alleviate the problem of domain-dependent tuning and adapt the kernel to our scoring function, we use inner product-based similarity metric as defined in Eq. 10:

$$K_{Inner}(e_i, e_j) = e_i^T e_j \tag{10}$$

In order that dual-embeddings of questions q_i and q_j can precisely model the similarity between the cross-embeddings of questions q_i and

Dataset	Training					Test			
	#Q	#A	#QA pairs	#A per Q	#Q per A	#Q	#A	#A per Q	#Q per A
ReQA SQuAD	87,355	58,934	87,599	1.00	1.48	10,539	7087	1.08	1.61
ReQA NQ	104,600	83,153	107,082	1.03	1.29	4,177	5799	1.43	1.03
ReQA HotpotQA	72,921	57,485	72,928	1.00	1.27	5,901	5745	1.00	1.03
ReQA NewsQA	71,561	39,415	74,160	1.03	1.88	4,185	2351	1.01	1.79

Table 1: Datasets statistics. #Q denotes the number of questions. #A per Q denotes the average number of matched answers for each question, and #Q per A denotes the average number of matched questions for each answer.

q_j, the conditional probabilities $p_{dual}(q_j|q_i)$ and $p_{cross}(q_j|q_i)$ should be as close as possible. Therefore, the GAM aims to learn a dual-embeddings representation that can minimize the divergence between $p_{dual}(q_j|q_i)$ and $p_{cross}(q_j|q_i)$, $p_{dual}(a_j|a_i)$ and $p_{cross}(a_j|a_i)$, $p_{dual}(a_j|q_i)$ and $p_{cross}(a_j|q_i)$ as well as $p_{dual}(q_j|a_i)$ and $p_{cross}(q_j|a_i)$. To achieve the aim of enhancement, the widely used Kullback-Leibler Divergence (KLD) is employed in this research. The loss function $\mathcal{L}_{q|q}$ defined on a mini-batch is adopted to minimize the divergence between $p_{dual}(q_j|q_i)$ and $p_{cross}(q_j|q_i)$, which can be calculated as Eq. 11:

$$\mathcal{L}_{q|q} = \frac{1}{B}\sum_{j=1}^{B}\sum_{i=1}^{B} p_{cross}(q_j|q_i)\log\frac{p_{cross}(q_j|q_i)}{p_{dual}(q_j|q_i)} \tag{11}$$

where B is the batch size; i and j are the indexes of the QA pairs in a given batch.

The same way can be used to calculate the loss functions $\mathcal{L}_{a|a}$, $\mathcal{L}_{a|q}$ and $\mathcal{L}_{q|a}$. Then the overall loss function of GAM can be defined as Eq. 12, where the hyper-parameters $\alpha_{a|q}$, $\alpha_{q|q}$, $\alpha_{q|a}$ and $\alpha_{a|a}$ are weights on different loss components:

$$\mathcal{L}_{ga} = \alpha_{a|q}\mathcal{L}_{a|q} + \alpha_{q|q}\mathcal{L}_{q|q} + \alpha_{q|a}\mathcal{L}_{q|a} + \alpha_{a|a}\mathcal{L}_{a|a} \tag{12}$$

3.5 Model Training and Inference

During training stage, we jointly train the Dual-Encoders and Cross-Encoders, and align the geometry of Dual-Encoders with that of Cross-Encoders. The overall loss function to train the full model is defined as Eq. 13, where α_{dual}, α_{cross} and α_{ga} are hyper-parameters to control the loss weight.

$$\mathcal{L} = \alpha_{dual}\mathcal{L}_{dual} + \alpha_{cross}\mathcal{L}_{cross} + \alpha_{ga}\mathcal{L}_{ga} \tag{13}$$

Since we only use the enhanced Dual-Encoders to encode questions in the inference stage while embeddings of answers are processed off-line, no extra computation is needed consequently.

4 Experiment

4.1 Datasets

Ahmad et al. (2019) introduced the Retrieval Question-Answering (ReQA) task, which focuses on sentence-level answer retrieval and establish a pipeline to transform a reading comprehension dataset to ReQA dataset. We conduct our experiments on ReQA SQuAD and ReQA NQ established from SQuAD v1.1 (Rajpurkar et al., 2016) and NQ (Kwiatkowski et al., 2019) respectively by Ahmad et al. (2019). We also use the same pipeline to process HotpotQA (Yang et al., 2018) and NewsQA (Trischler et al., 2017) datasets for more experiments. ReQA HotpotQA and ReQA NewsQA are used to denote the processed version of HotpotQA and NewsQA datasets respectively in this research. Since the original test sets of datasets above are not publicly available, the original validation sets are used as test sets. The statistics of ReQA datasets are shown in Table 1.

4.2 Evaluation Metrics

We adopt two popular metrics[1] for evaluation, i.e., mean reciprocal rank (MRR) and recall at N (R@N), which are widely used for measuring retrieval-based QA task (Ahmad et al., 2019).

MRR is the average reciprocal ranks of retrieval results, as illustrated in Eq. 14, where Q is a set of questions and $rank_i$ is the rank of the first correct answer for the i^{th} question.

$$\text{MRR} = \frac{1}{|Q|}\sum_{i}^{|Q|}\frac{1}{rank_i} \tag{14}$$

R@N is the recall score in top-N predicted subsets, as illustrated in Eq. 15, where A_i is the ranked answer list for the i^{th} question and A_i^* is the corresponding correct answer set.

$$\text{R@N} = \frac{1}{|Q|}\sum_{i}^{|Q|}\frac{|top_N(A_i)\cap A_i^*|}{|A_i^*|} \tag{15}$$

[1]https://github.com/google/retrieval-qa-eval

4.3 Compared Methods

BM25 A classical ranking method using TF-IDF like scoring function for information retrieval (Robertson and Zaragoza, 2009).

InferSent A universal sentence encoder trained with supervised natural language inference task, not in need of fine-tuning for specific retrieval task (Conneau et al., 2017).

USE-QA A multi-task pre-trained model based on the Transformer, which learns universal sentence representation through a multi-feature ranking task, a translation ranking task and a natural language inference task (Yang et al., 2020).

Dual-Encoders The vanilla Dual-Encoders train from scratch and can be implemented using different encoders. For instance, we use Dual-BERTs to denote the Dual-Encoders using BERT as encoder.

Dual-VAEs A model trained jointly with the question-to-question and answer-to-answer reconstruction tasks using VAE (Shen et al., 2018).

Cross-VAEs A model to solve one-to-many problem in answer retrieval, aligning the feature spaces of questions and answers by the question-to-answer and answer-to-question reconstruction (Yu et al., 2020).

ENDX-Encoders (Ours) The Dual-Encoders is enhanced by our ENDX framework. For instance, ENDX-BERTs is used to denote the Dual-BERTs enhanced by ENDX.

4.4 Implementation Details

We split the training sets of all datasets into new training set and validation set in a ratio of 9:1. The hyper-parameters are chosen according to the model performance (R@1) on validation set. Specifically, Dual-BERTs and ENDX-BERTs are initialized using BERT base model (Devlin et al., 2019), and the encoder of other models has 2 layers and uses 768-dim BERT token embedding as input. The cross attention modules of all ENDX-Encoders have 12 heads. We use AdamW optimizer (Loshchilov and Hutter, 2017) to train BERT-based model with 30 epochs and linearly decay the learning rate initialized as 2e-5, and train other models with 100 epochs using constant learning rate initialized as 1e-5. We set the loss weights α_{dual}, α_{cross} and α_{ga} to 0.25, 0.25 and 0.5 respectively. The loss weights $\alpha_{a|q}$ and $\alpha_{q|a}$ increase linearly from 0 to 0.5, while $\alpha_{q|q}$ and $\alpha_{a|a}$ increase linearly from 0 to 1e4 both over the first 5 epochs. The batch size of BERT-based model is set to 12, and that of other models is set to 100. Finally the parameters that perform best on validation set are used on test set.

4.5 Results and Analysis

Main Results The results on ReQA SQuAD are shown in Table 2. BM25 shows competitive performance, since keywords overlap is common in ReQA SQuAD. As a pre-trained universal sentence encoder without fine-tuning, InferSent does not perform well as the pre-training datasets are relatively small. USE-QA gets stronger performance because of the use of a more powerful encoder and a larger-scale pre-training dataset. Compared to Dual-VAEs, Cross-VAEs improves MRR, R@1 and R@5 by 1.32%, 1.07% and 2.28% respectively, while our ENDX-BERTs outperforms the current best model Cross-VAEs (Yu et al., 2020) on MRR, R@1 and R@5 by 17.88%, 15.00%, 21.60% respectively and achieves new state-of-the-art result on ReQA SQuAD.

Method	MRR	R@1	R@5
BM25	52.96	45.81	61.31
InferSent†	36.90	27.91	46.92
USE-QA†	61.23	53.16	69.93
Dual-VAEs†	61.48	55.01	68.49
Cross-VAEs†	62.29	55.60	70.05
Dual-RNNs	52.19	40.96	65.11
ENDX-RNNs	53.68(↑)	42.20(↑)	67.30(↑)
Dual-GRUs	55.24	44.39	68.00
ENDX-GRUs	58.65(↑)	48.29(↑)	70.90(↑)
Dual-LSTMs	58.77	49.26	69.79
ENDX-LSTMs	61.00(↑)	50.79(↑)	72.87(↑)
Dual-Transformers	62.58	51.51	75.99
ENDX-Transformers	63.73(↑)	53.41(↑)	76.02(↑)
Dual-BERTs	71.06	61.24	83.09
ENDX-BERTs	**73.43**(↑)	**63.94**(↑)	**85.18**(↑)

Table 2: Performance on ReQA SQuAD dataset, where the results with † are reported from (Yu et al., 2020).

Table 3 shows the performance comparison on ReQA NQ, ReQA HotpotQA and ReQA NewsQA datasets. Since the results in Table 2 have already shown the Dual-BERTs and ENDX-BERTs can significantly outperform BM25, InferSent, USE-QA, Dual-VAEs and Cross-VAEs, we only compare Dual-Encoders and ENDX-Encoders. The results in Table 2 and Table 3 both indicate the superiority of our ENDX framework which consistently outperforms Dual-Encoders with signif-

Method	ReQA NQ			ReQA HotpotQA			ReQA NewsQA		
	MRR	R@1	R@5	MRR	R@1	R@5	MRR	R@1	R@5
Dual-RNNs	41.45	26.84	60.01	22.86	13.83	32.71	22.23	12.80	32.64
ENDX-RNNs	43.57(↑)	29.29(↑)	62.11(↑)	24.20(↑)	14.69(↑)	34.79(↑)	23.33(↑)	14.39(↑)	33.14(↑)
Dual-GRUs	44.99	31.26	62.31	25.44	16.61	34.77	26.21	16.86	37.16
ENDX-GRUs	47.18(↑)	33.36(↑)	64.09(↑)	26.96(↑)	17.76(↑)	37.08(↑)	28.61(↑)	19.78(↑)	38.51(↑)
Dual-LSTMs	46.07	33.13	62.28	25.16	16.68	34.33	28.39	20.32	37.06
ENDX-LSTMs	49.29(↑)	36.18(↑)	65.25(↑)	25.78(↑)	16.66(↓)	35.52(↑)	30.01(↑)	21.42(↑)	39.55(↑)
Dual-Transformers	46.34	31.90	64.68	26.22	15.40	38.64	27.82	18.00	38.77
ENDX-Transformers	47.85(↑)	33.52(↑)	65.99(↑)	26.59(↑)	15.54(↑)	39.45(↑)	29.25(↑)	19.21(↑)	41.00(↑)
Dual-BERTs	54.80	40.58	72.66	39.04	27.13	52.91	37.35	26.64	49.36
ENDX-BERTs	**57.76**(↑)	**43.32**(↑)	**76.15**(↑)	**40.68**(↑)	**28.74**(↑)	**54.58**(↑)	**37.90**(↑)	**27.26**(↑)	**49.95**(↑)

Table 3: Performance comparison on ReQA NQ, ReQA HotpotQA and ReQA NewsQA datasets.

icant margins. For instance, on MRR, R@1 and R@5, ENDX-LSTMs outperforms Dual-LSTMs by 6.99%, 9.21%, 4.77% in ReQA NQ, and ENDX-Transformers outperforms Dual-Transformers by 5.14%, 6.72%, 5.75% in ReQA NewsQA. Compared to the powerful Dual-BERTs, our ENDX-BERT shows average relative improvements over four datasets by 3.60%, 4.86% and 2.92% on MRR, R@1 and R@5 respectively (t-test of 10 runs, p-values < 0.01).

Method	MRR	R@1	R@5
USE-QA	47.06	40.90	53.44
Cross-VAEs	48.52	44.55	53.52
Dual-BERTs	60.19	48.56	74.02
ENDX-BERTs	**64.93**	**52.23**	**81.36**

Table 4: Performance on ReQA SQuAD sub-dataset, each answer of which has at least 8 matched questions.

Performance on sub-dataset We conduct more experiments on sub-datasets of ReQA SQuAD to validate the effectiveness of our framework on coping with the one-to-many problem. The comparison results between Dual-BERTs and ENDX-BERTs on sub-datasets, in which answers have different minimum number of matched questions, are shown as Fig. 2. It is observed that ENDX-BERTs outperforms Dual-BERTs solidly. The results of the most difficult sub-dataset, in which answers have at least 8 different questions, are shown in Table 4. Compared to Dual-BERTS, USE-QA and Cross-VAEs, our proposed model prominently shows the best performance under such a significant one-to-many circumstance.

Analysis on the effects of GAM We also sample multiple questions with same answer and encode the questions by Cross-Encoders, Dual-Encoders

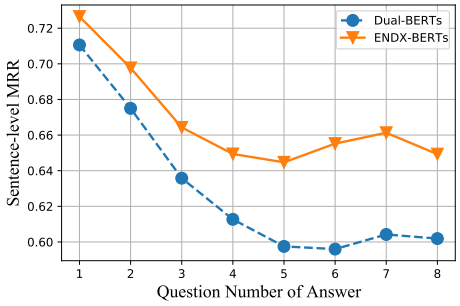
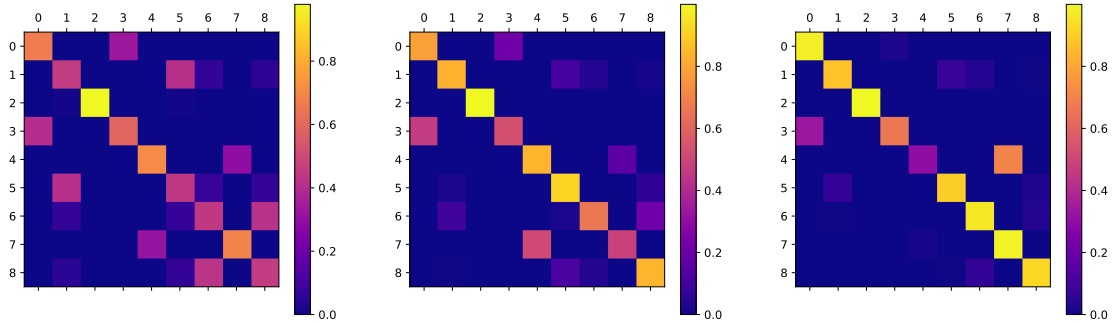


Figure 2: Comparison between Dual-BERTs and ENDX-BERTs on ReQA SQuAD sub-datasets where answer has different minimum number of matched questions.

enhanced with the proposed GAM, and basic Dual-Encoders, respectively. The question-question similarity matrices are visualized in Fig. 3. In cross-embeddings, questions could attend the matched answer which results in more accurate question representations and better capture of the correlations between questions (see Fig. 3(a)). During ENDX training, we use GAM to align the geometry of dual-embeddings with that of cross-embeddings. As shown in Fig. 3(b), dual-embeddings enhanced by GAM are able to capture more correlations in question-question similarities compared to baseline dual-embeddings (Fig. 3(c)).

Ablation study on loss function of GAM We perform the ablation study on the proposed ENDX-BERTs in ReQA NQ by removing different components of GAM loss function. As shown in Table 5, all metric scores drop significantly without optimizing $\mathcal{L}_{a|q}$ or $\mathcal{L}_{q|a}$, which indicates that $p(a_j|q_i)$ and $p(q_j|a_i)$ describe the most important parts of geometry. The reason we conjecture is that the answer retrieval task focus more on the relative distance of question-to-answer in feature space, while $\mathcal{L}_{q|q}$ and $\mathcal{L}_{a|a}$ are also helpful.



(a) Cross-embeddings (ENDX) (b) Dual-embeddings (ENDX) (c) Dual-embeddings (Baseline)

Figure 3: Question-question similarity matrices of ENDX cross-embeddings, ENDX dual-embeddings and baseline dual-embeddings, where the i^{th} row of matrix denotes the similarity between ith question and the others.

Method	MRR	R@1	R@5
Dual-BERTs	54.80	40.58	72.66
ENDX-BERTs	**57.76**	**43.32**	**76.15**
w/o $\mathcal{L}_{q\|q}$	56.37	42.01	74.65
w/o $\mathcal{L}_{a\|a}$	56.89	42.25	75.28
w/o $\mathcal{L}_{q\|a}$	56.00	42.12	73.27
w/o $\mathcal{L}_{a\|q}$	55.87	41.39	74.14

Table 5: Ablation study on ReQA NQ dataset.

Comparison with BERT$_{\text{QA}}$ We also compare the proposed ENDX-BERTs against the interaction-based model BERT$_{\text{QA}}$ (Devlin et al., 2019), which encodes concatenated sequence for every candidate QA pair. Due to the extremely large computational cost of BERT$_{\text{QA}}$, we only sample 500 QA pairs from 27 passages in ReQA SQuAD as the test set. The experimental result is shown in Table 6, where ENDX-BERTs improves MRR, R@1 and R@5 over Dual-BERTs by +4.61%, +4.59% and +5.44% respectively and only falls behind BERT$_{\text{QA}}$ by -3.38%, -4.93% and -0.81%. However, the inference runtime complexity is significantly reduced from $O(n \times m)$ to $O(n + m)$ compared to BERT$_{\text{QA}}$, where n and m are the numbers of questions and answers respectively. Therefore, the propopsed ENDX-BERTs can better balance between accuracy and efficiency for answer retrieval.

Method	MRR	R@1	R@5	average ms
Dual-BERTs	71.83	60.42	87.26	14.19
ENDX-BERTs	76.44	65.01	92.70	14.19
BERT$_{\text{QA}}$	79.82	69.94	93.51	6939.61

Table 6: Comparison with BERT$_{\text{QA}}$, where the average time (ms) to retrieve answer for one question is tested on one NVIDIA Tesla V100 GPU.

4.6 Case Study

Figure 4 shows the dual-embeddings projection (t-SNE, Van der Maaten, 2008) of 6 different questions and their shared answer. It can be seen that the dual-embeddings of our ENDX-BERTs are more compact than that of Dual-BERTs, which proves that our method can better align the questions and answers, and can produce more general representation to alleviate the one-to-many problem.

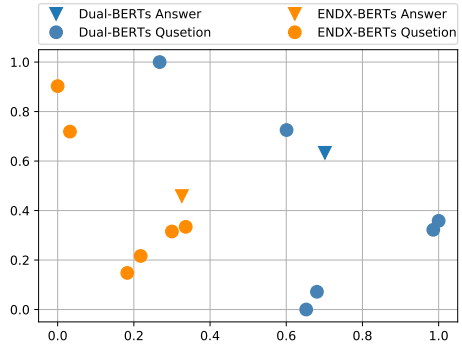


Figure 4: A case of 6 different questions sharing one answer, where the blue dot and yellow dot present the question and answer embeddings of Dual-BERTs and ENDX-BERTs in 2D space respectively.

5 CONCLUSION

In this work, we propose a framework that enhances Dual-Encoders with cross-embeddings for answer retrieval. A novel geometry alignment mechanism is introduced to align the geometry of Dual-Encoders with cross-embeddings. Extensive experimental results show that our method significantly improves Dual-Encoders model and outperforms the state-of-the-art method on multiple answer retrieval datasets.

References

Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. Reqa: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering, MRQA@EMNLP 2019, Hong Kong, China, November 4, 2019*, pages 137–146.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680.

Arpita Das, Harish Yenala, Manoj Chinnakotla, and Manish Shrivastava. 2016. Together we stand: Siamese networks for similar question retrieval. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 378–387.

Michel Deudon. 2018. Learning semantic similarity in a continuous space. In *Advances in neural information processing systems*, pages 986–997.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Matthew Henderson, Ivan Vulic, Daniela Gerz, Iñigo Casanueva, Pawel Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrksic, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5392–5404.

Geoffrey E. Hinton and Sam T. Roweis. 2002. Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems 15 [Neural Information Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada]*, pages 833–840. MIT Press.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wen-tau Yih, Barlas Oguz, Veselin Stoyanov, and Gargi Ghosh. 2021. Multi-task retrieval for knowledge-intensive tasks. *CoRR*, abs/2101.00117.

Nikolaos Passalis and Anastasios Tefas. 2018. Learning deep representations with probabilistic knowledge transfer. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pages 283–299. Springer.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

David W. Scott. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley Series in Probability and Statistics. Wiley.

Min Joon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur P. Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4430–4441.

Dinghan Shen, Yizhe Zhang, Ricardo Henao, Qinliang Su, and Lawrence Carin. 2018. Deconvolutional latent-variable model for text sequence matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 719–727.

Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. 2017. Few-shot learning through an information retrieval lens. *arXiv preprint arXiv:1707.02610*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Berwin A Turlach. 1993. Bandwidth selection in kernel density estimation: A review. In *CORE and Institut de Statistique*. Citeseer.

Laurens Van der Maaten. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NIPS*, pages 5998–6008.

Zhongbin Xie and Shuai Ma. 2019. Dual-view variational autoencoders for semi-supervised text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5306–5312. ijcai.org.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Wenhao Yu, Lingfei Wu, Qingkai Zeng, Shu Tao, Yu Deng, and Meng Jiang. 2020. Crossing variational autoencoders for answer retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5641.