JTCSE: Joint Tensor-Modulus Constraints and

Content

Abstract—Unsupervisedcontrastivelearninghasbecomeahot research topic in natural language processing. Existing works usually aim at constraining the orientation distribution of the representations of positive and negative samples in the high- dimensional semantic space in contrastive learning, but the semanticrepresentationtensorpossessesbothmodulusandorien- tationfeatures,andtheexistingworksignorethemodulusfeature oftherepresentationsandcauseinsufficientcontrastivelearning. Therefore, we first propose a training objective that is designed to impose modulus constraints on the semantic representation tensor, to strengthen the alignment between positive samples in contrastivelearning.Then,theBERT-likemodelsuffersfromthe phenomenon of sinking attention, leading to a lack of attention of Cross-Attention hidden States toCLStokensthataggregatesemanticinformation.Inresponse, we propose a cross-attention structure among the twin-tower ensemble models to enhance the model's attention to CLS token (EDFSE[1]),whichnaivelytrainsmultiplesub-encodersseparatelyandthen sumstheoutputs.Thisapproachcausesalargeinferenceoverhead. a new Joint Tensor representation b.representstheoptimizedensemblelearningframeworkJTCSEproposedin moduluscons traintandCross-attentionunsupervisedcontrastive thiswork.Itincorporatessemanticrepresentationtensormodulusconstraints learningSentenceEmbeddingrepresentationframeworkJTCSE, andjointmodelingofcross-attentionbetweensub-encoders.Thisframework which we evaluate in seven semantic text similarity computation contains only two sub-encoders. It significantly reduces inference overhead tasks,andtheexperimentalresultsshowthatJTCSE'stwin-tower whileimprovingthequalityofsentenceembeddingsrelativetoa. ensemble model and single-tower distillation model outperform the other baselines and become the current SOTA. In addition, we have conducted an extensive zero-shot downstream task based on evaluation,whichshowsthatJTCSEoutperformsotherbaselines these two PLMs, e.g., Sentence-BERT [4], ConSERT [5], and overall on more than 130 tasks. SimCSE [6]. SimCSE applies InfoNCE's [7] idea of con- We open source the code and checkpoints for this work as trastivelearning[8]bygeneratingpositivesamplesthroughthe follows: Dropout method of the BERT-like model at training time and uniformly distributing unlabeled soft-negative samples. With theappearanceofSimCSE,manyworksarebasedontheidea ofunsupervisedcontrastivelearninginSimCSEandInfoNCE. For example, ESimCSE [9] augments the positive sample in SimCSEbyconstructingproximitywordstoreplaceindividual wordsintheoriginalsample.Inaddition,ESimCSEintroduces the idea of momentum queueing in MoCo [10] to expand the scope of contrastive learning. DiffCSE [11] learns the differences between original sentences and forged sentences bygeneratingforgedsamplesthroughELECTRA[12]andthe Replaced Token Detectio task to improve the quality of sen- tence tensor representations. ArcCSE [13] generates multiple positive samples by masking the original sentences multiple

TianyuZong,HongzhuYi,YuanxiangWang,andJungangXuarewiththe SchoolofComputerScienceandTechnology,UniversityofChineseAcademy

times and constructing the positive sample triples to model ofSciences,CloudComputingandIntelligentInformationProcessingLabo- entailmentbetweensentencepairsinthetriples.InfoCSE[14] ratory(CCIP),Beijing,101408,China.E-mail:{zongtianyu20,yihongzhu23, wangyuanxiang19}@mails.ucas.ac.cn,. directly adds several EncoderLayers of Transformers [15] as Bingkang Shi is with the Institute of Information Engineering, Chi- auxiliary networks to the exterior of the BERT-like model nese Academy of Sciences, Beijing, 100085, China. E-mail: shib- and provides MLM constraints on the output of the auxil- . iary networks, while InfoNCE self-supervised constraints are 5202 yaM 7]LC.sc[2v66320.5052:viXra JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 2 applied to the ontology of the BERT-like model. SNCSE [16] phenomenon: the models disproportionately focus their at- is inspired by PromptBERT [17] and uses the Prompt tech- tention on end-of-sequence tokens at the last encoding layer. nique and different templates to generate enhanced positive We believe this attention allocation bias leads to insufficient and negative samples to improve the quality of positive and attention to the CLS token used to gather global information, negative sample pairs in SimCSE. EDFSE [1] first applies the making it difficult for the CLS token to effectively capture data augmentation method of Round Trip Translation (RTT), globalsemanticinformationandleadtolower-qualitysentence which translates the original English dataset into different embedding. languages and then into English through a translation system. The phenomenon of attention sinking is widespread in It then applies the idea of ensemble learning [18] to train BERT-like models1, which may be related to BERT pre- multiple BERT-like pre-trained models with different RTT training. All the related work is fine-tuned based on BERT- datasets and the method of SimCSE, respectively. Finally, like models, which makes it difficult to directly address the these BERT-like models are integrated into a large ensemble phenomenon. However, we further observe that the perfor- model to achieve model performance improvement. RankCSE mance of the BERT-like model on semantic textual similarity [19] also applies the same idea of ensemble learning, with computationtasksispositivelycorrelatedwiththe2-paradigm the difference that it uses the existing checkpoint SimCSE- ratio of CLS to the hidden state of other tokens in the self- base/largeastheteacherandranksthesimilarityofthecurrent attentionmoduleoftheEncoderLayers.Specifically,thelarger sample to other soft-negative samples by multiple teachers the ratio of the 2-parameter of the tensor of the hidden state and students. The multiple teachers constrain the students' corresponding to CLS to the 2-parameter of the matrix of performance through KL scatter loss, while the unsupervised the hidden states of the other tokens (defined as CLS energy InfoNCE still constrains the students. weights), the better the model performs. Therefore, we can However, there are some common problems with the above enrichthesemanticinformationofCLSpoolingbyincreasing works, and to clarify the significance and direction of this the CLS energy weights. work, we have organized the motivations as follows: Noting that cross-attention [21] has been widely employed 1) Existing works neglect the modulus feature of sen- in multimodal learning [22]–[25], we consider that in textual tence embedding representations: Starting from SimCSE, information understanding, the attention mechanism performs the common point of the above baselines is that they all in- inter-modelcross-computationtoenabletheCLStokentofuse volveunsupervisedcontrastivelearningastheprimarytraining the differentiated features of the twin encoder. Therefore, we method and employ InfoNCE as the central loss for model design a cross-attention in the proposed model to enhance the training,fine-tuningaBERT-likemodel.However,theseworks CLS energy weights, enabling the CLS pooling to aggregate invariably ignore a critical issue. Since each EncoderLayer bettersentencesemanticinformationandalleviatingthedraw- of the BERT-like model contains two LayerNorm layers for backs of attention sinking. normalizingthesamplefeatures,thehigh-dimensionalfeatures 3) Largeinferenceoverheadsandnon-autonomoustrain- ofthetextaremappedontoahypersphereduetothepresence ing remain challenges: SincebothEDFSE[1]andRankCSE of LayerNorms. This causes the text tensor representation to [19] adopt the ensemble learning training method, i.e., con- losethemodulus-lengthfeaturesandretainonlytheorientation structing a multi-tower (or twin-tower) model that unifies the features. Meanwhile, InfoNCE only constrains the alignment featuredistributionofeachencoderoutput.However,sincethe of feature representations between positive samples by cosine size of the ensemble model proposed by EDFSE is equivalent similarity and

unlabeled soft samples are distributed to the to 6 SimCSE-BERT-bases, which imposes a considerable whole hypersphere, which makes the modulus feature of the inference overhead, and RankCSE relies on the same type textbefurtherignored;therefore,thereexistsasituationinall of checkpoints to do knowledge distillation, which is not an SimCSE-typemodels,i.e.,ignorethefeaturesoftwomutually autonomous training method, the above problems are yet to positive samples are very different from each other in terms be solved. of modulus although their orientations are approximately the Based on the above discussion, we conclude that there are same in the high-dimensional space while we believe that three main motivations: The first is that the modulus of high- the orientation and mode length of features that are positive dimensional feature representations with mutually positive samples of each other should be approximately the same, it is samples need to be constrained; the second is that cross- necessary to propose a loss function for the tensor modulus attentionneedstobeintroducedtoincreasethequalityofCLS constraint. pooling, which in turn improves the model's performance on 2) More attention is needed for the CLS token: There is downstream tasks, and the third is that traditional ensemble a existing work [20] introduces the possibility that generative learning methods should be optimized to reduce the infer- language models may suffer from attentional sinking, and ence overhead and improve the quality of the output tensor. we observe that the same phenomenon exists for BERT-like Therefore, in this paper, we propose a joint semantic tensor embedding models. To the best of our knowledge, existing modulus constraint and cross-attention in ensemble model baseline models generally adopt the CLS pooling strategy, for unsupervised contrastive learning of sentence embedding, which utilizes the hidden state of CLS tokens to represent JTCSE. the semantic information of the whole sentence. However, by observing the attention score distribution, we notice that almost all baseline models suffer from an "attention sink" 1WereportthisfindingindetailintheFig.9. JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 3 According to the first motivation, we first propose an intu- a modulus-constrained training objective targeting unsu- itive training objective, i.e., the two semantic representation pervised contrastive learning, and the proposed training tensors of positive samples should have similar modulus and objective is proved to be effective through extensive orientations; in other words, the two feature representations comparative experiments and ablation experiments. of positive samples should be close to each other in terms • In order to enhance the BERT-like model's attention to oftheir distributionalpositionsin thehigh-dimensionalspace; CLS tokens, we introduce cross-attention in the twin- according to the second motivation, inspired by the methods tower ensemble model, which is jointly modeled by ofvisual-languagetoconstructanensemblemodel,wedesign multiple spatial mappings to enhance the energy weight the twin-tower model to achieve feature fusion by construct- of CLS pooling, and hence optimize the model's perfor- ing cross-attention between towers and to promote feature mance on multi-tasks. sharing and information complementarity between towers, as • Combining the tensor modulus constraints and the cross- shown in Fig. 1, which boosts the CLS energy weights in attention mechanism, our proposed twin-tower ensem- the model and optimize the performance of CLS pooling in ble model effectively reduces the inference overhead downstream tasks. Incorporating the solutions to the first two of the traditional multi-tower ensemble model EDFSE motivations, we obtain an ensemble model of the twin-tower and performs better on the semantic textual similarity trained autonomously, which solves the third motivation by computation task. both compressing the inference overhead and avoiding the • We have conducted extensive evaluations.Firstly, JTCSE discussion of non-autonomous training. performs best in seven semantic text similarity tasks, Following the existing works, we evaluate seven semantic and our proposed JTCSE and its derived models perform text similarity (STS) tasks ([26], [27], [28], [29], [30], best in the currently open-source checkpointing baseline [31], [32]) and achieve SOTA results in the baseline of in a variety of 0-shot evaluations for downstream tasks all open-source checkpoints to the best of our knowledge. in natural language processing. We conduct a detailed Compared to the EDFSE-BERT-base, our proposed JTCSE- ablation analysis to gain insight into the strengths of the BERT-base has an inference overhead close to one-third of proposed model and open source the entire code and it but outperforms it on 7 STS tasks, proving our proposed checkpoints of this work. framework's effectiveness. To have a fairer comparison, we • To promote research progress in related areas, we have propose two approaches. The first one concerns the non- open-sourced the code and saved checkpoints for this autonomouslytrainedmodelRankCSE.Weperformensemble work. learning on the other baselines and re-compare them on the ThisworkisanextensionoftheexistingworkTNCSE:Ten- 7 STS tasks, and the results show that the JTCSE is better. sor'sNormConstraintsforUnsupervisedContrastiveLearning Second, we compress the twin-tower

JTCSE into a single- of Sentence Embeddings [33], which has been accepted by tower model employing knowledge distillation with the same AAAI25 for Oral presentation. This work's main update parameter scales as the other baselines. On the 7 STS tasks, comparedtopreviousworkistheadditionofacross-attention we get the distillation model that still performs the optimal. structuretothetwinencodertoenhancetheBERT-likemodel's We report significant experimental results showing that the attention to the CLS token. This strengthens the CLS token's proposed model's performance gain does not depend on the ability to capture the global semantic information of the setting of random seeds. In addition, we have conducted a sentenceandoptimizesthemodel'sCLSpoolingperformance
broaderzero-shotevaluationbasedontheMTEB2 framework, in downstream tasks. We report the main updates to this work including more than 130 tasks such as text retrieval, text relative to the predecessor work as follows: classification, text re-ranking, bi-textmining, multilingual text semantic similarity, etc. The results show that our proposed • New Movation: We have noticed that unsupervised sen- JTCSE and the derived models are generally ahead of the tence embedding models of BERT-like models usually current open-source checkpoints. In addition, we report the suffer from attention sinking; they pay more attention to performance gains from tensor modulus constrained objective theendpunctuationorSEPtokenoftheinputsequencein and cross-attention, respectively; we visualize the attentional the last coding layer, and lack of attention to CLS token, sink phenomenon for the BERT-like model and also report coupled with the fact that all of these baselines use CLS the trend line of the near-positive correlation between CLS pooling,webelieveshouldenhancethemodel'sattention energy weights and STS task performance. In the discussion to CLS token attention to improve the quality of CLS section, we present a detailed analysis of the motivation and pooling. methodological soundness of our proposed tensor modulus • New Method: Directly optimizing the attention weight constraints. In addition to the significant experimental results, matrixmaydestroythepre-traininginformationofBERT- we report the inference overhead for different baselines and like models. For this reason, we propose the concept of visualizetheexperimentalresultsofalignmentanduniformity CLS energy weights to enhance the model's attention for unsupervised embedding representation models. to the CLS token by boosting the CLS energy weights. We summarize the main contributions as follows: Based on existing works, we find that different Encoder- • To the best of our knowledge, we are the first to propose Layers of BERT-like models focus on different features oftheinputsequence.Intuitively,differentEncoderLayers 2 in different models may also capture different semantic JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 4 features of the same sentence. Therefore, we introduce a cross-attention mechanism for feature fusion between h models.Thisapproachenrichesthesemanticinformation aggregated by the CLS token, which enhances the CLS h ■ h■ energy weight and thus optimizes the performance of CLS Pooling in downstream tasks. O • New Experimental Results: Relative to the previous ■ work TNCSE, the model JTCSE retrained in this work hasimproveditsperformanceon7STStasks;inorderto evaluateJTCSE'sgeneralizationabilitymorecomprehen- ■ sively, we have conducted an extensive 0-shot evaluation h of downstream tasks of natural language processing, our evaluation shows JTCSE achieves average performance gains across more than 130 downstream tasks compared Fig. 2. This figure represents the distribution of the positions of a pair to existing baselines, reaching new SOTA results. of positive sample semantic representation tensors h and h+ in three- • Other Updates: We enrich the insight of the tensor dimensional space and the vectors h–h+ for which they are subtracted. modulusconstrainedtrainingobjectivedesignbydecom-
Accordingtotheprincipleofsimilartriangles,whentheangleγ isspecific, the larger the modulus of h or h+, the larger the modulus of h–h+ will posing it into two sub-objectives and discussing their be,andthegreaterthevalueofbeingconstrained. significance separately. For the more extensive evaluations, we add the English part of the three multilingual STS tasks and enrich the seven STS tasks evaluated by existing work into ten. The results show that JTCSE and the derived models still generally perform best. ■■■■■■ In the subsequent sections, this paper systematically sum- ■7■O ■■■■■■ ■1 marizes the representative work on unsupervised sentence (■N ■■■■■■ embedding models in the related work section and overviews ■■ the typical applications of the cross-attention mechanism in)■W ■■■■■■ multimodal information fusion; in the method section, we ■■■■■■ ■■■■■■■■ derive the tensor modulus constraint training objective in ■■■■■■ ■■■■■■■■ detail based on the motivation of solving the problem that the ■■■■■■ ■■■■■■ ■R■V InfoNCElossfunctionneglectsthepositivesamples'modulus ■N ■ ■K■■■■■■■■ ■■■■■■■■■■■■■■■■■■■■ ■W ■ ■F alignment, and meanwhile targeting to alleviate the attention

■K sinking and enhance the CLS Pooling information density, an innovative cross-attention structure is proposed, and finally, Fig.3. Thisfigure illustratesthe binaryloss functionLTMC,with respect totherangeofvaluesofthetwoindependentvariablestandk overpartof the model architecture and loss function design are presented itsdomainofdefinition. in full; the Experiments section details the training data and evaluation tasks and reports the performance of the model on further introduces various data augmentation strategies (e.g., more than 130 tasks; the Ablation study discusses the impact randomdeletion,randominsertion,etc.[34])toenrichpositive ofthecomponentsandthereasonablenessofthelossfunction samples. Subsequent studies have continuously optimized the design; the Discussion section concerns the reasonableness of training methods for unsupervised sentence embedding based the tensor-module constrained training objectives' design and on SimCSE. For example, ESimCSE [9] combines near- detailed motivations for cross-attention design. synonym data augmentation and MoCo's [10] momentum queuingmechanismtoimprovethequalityofSimCSE'srepre- sentations;DiffCSE[11]combinesmaskedlanguagemodeling II. RELATEDWORK by introducing an additional discriminator ELECTRA [12] A. Unsupervised Sentence Embedding Approach and further optimizes the model performance by employing InfoNCE [7] (Noise Contrastive Estimation Loss) is a InfoNCE; ArcCSE [13] refers to the positive-negative-sample widely used loss function for self-supervised learning, mainly tripleconstructproposedbySentenceBERT[4]andfine-tuned for feature representation learning. The method usually uses based on SentenceBERT; SNCSE [16] employs a comparison cosine similarity to compare the similarity between positive learningstrategywithsoftnegativesamplescombinedwithbi- and negative samples and optimize the model parameters. In directional marginal loss; InfoCS [14] additionally introduces unsupervised sentence embedding training, many studies have an additional network for mask language modeling; EDFSE combined InfoNCE loss with pre-trained language models [1] employs the Round-Trip Translation data augmentation (e.g.,BERT[2],RoBERTa[3]),whichhaspushedtheprogress strategy to train multiple encoders to construct a large-scale of contrastive learning. For example, both SimCSE [6] and integration model; and the current SOTA method RankCSE ConSERT [5] utilize the idea of dropout to generate positive [19] utilizes dual-teacher ensemble learning with distillation samples, and both use cosine similarity as the only metric to techniques to train encoders. The common point of the above distinguish between positive and negative samples. ConSERT studies is that they all introduce the InfoNCE loss. However, JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 5 they only rely on the cosine similarity between embeddings tensor representations between positive and negative samples when using InfoNCE for similarity metrics, ignoring the by the constraints of the InfoNCE loss function, which con- critical factor of the modulus of the embedding tensor. strains the cosine similarity between them to be as large To this end, we address this motivation by proposing an as possible for pairs of positive samples and as uniformly unsupervised training objective incorporating embedding rep- distributed as possible for unlabeled soft negative samples. resentationmodulusconstraintstoimprovefurtherthemodel's From the perspective of positive samples, InfoNCE requires ability to detect positive and negative sample discrimination. the orientation of positive sample tensor pairs to be aligned; however, from the mathematical point of view, a tensor has the features of "magnitude" and "orientation," while InfoNCE B. Application of Cross-Attention only constrains the tensor's "orientation" but ignores the Cross-attention has been widely used in the field of mul- "magnitude." SimCSE applies InfoNCE by passing a sample timodal embedding alignment.Visual-BERT [23] encodes im- through a BERT-like model to get the corresponding tensor, age and text tokens into a multimodal sequence that is fed and due to the presence of Dropout in the model, a sample into an Encoder-Only model for joint multimodal modeling; can produce two similar features, which are positive samples VILBERT [22] uses a dual-stream structure to process visual to each other, and the features of the other samples in a mini- and linguistic information separately before feature fusion, batch are as soft-negative samples for unsupervised training. LXMERT [35] processes the inputs of the two modalities SimCSE and its derivatives use the "orientation" of the tensor separately and introduces a text mask, and introduces a cross- astheonlymetrictojudgethesimilaritybetweenpositiveand modalencoderinadditiontoaself-encoderforeachmodality, negative samples, which lacks attention to the "magnitude" of and ALIGN [36] uses contrastive learning to align the image the tensor. Therefore, we use the modulus of the tensor, i.e., andtextintoasharedembeddingspace.CLIP[37]doesnotdi- the 2-parameter, to represent the "magnitude" of the tensor, rectlyapplycross-attentionandfurtheroptimizesperformance and from an intuitive geometric perspective,

we propose a onthevisual-verbaltaskthroughasimpletwin-towerstructure constraint objective L for the modulus of the tensor TMC and intuitive multimodal contrastative learning; BridgeTower between pairs of positive samples, in order to strengthen the [24]andManagerTower[25]buildonCLIPbyintroducingan model to judge the features of positive and negative samples additionalcross-attentionnetworkbetweenthetwintowersand that do not have apparent differences in orientation as shown applying an early feature fusion strategy, and on the Visual- in Eq. 1. Linguistic Question and Answer task outperforms CLIP. ■h–h+■ Based on the research trend of multimodal learning, we L (h,h+)= . (1) TMC ■h■+■h+■ have found that the better the performance of the twin-tower ensemblemodelinavisual-linguisticalignmenttask,themore In L , h and h+ denote the features of a pair of positive parameters need to be introduced for cross-attention-based TMC samples, respectively, and ■h■, ■h+■ denotes the modulus of feature fusion to achieve complementary modeling among the tensor, which is the 2-parameter. We first construct the features. tensor representation space of a pair of positive samples and However, to our knowledge, there is no ensemble model differencevectorsaccordingtoFig.2,andweexpectthemodel based on cross-attention for feature sharing with a twin-tower to be trained with the angle cosγ and the modulus of the structureinunsupervisedsentenceembedding.Thus,itisnec- difference vectors of both as small as possible. essarytoproposeitinthisfield.Westartfromtheperspective In addition, intuitively, the larger the modulus of ■h■ and of compressing the training overhead as much as possible, ■h+■is,themorepronouncedthemodulusoftheirdifference similar to CLIP, by keeping only the basic encoder without vectors are and the more valuable the constraints are when introducing other training parameters, and based on cross- the angles γ are equal, so we establish two sub-objectives. attention to achieve feature complementarity and enhance the The first is that the modulus of the difference vectors of the representation quality of CLS Pooling. positive sample pairs should be as small as possible, and the second is that the modulus of each of the positive sample III. METHODS pairs should be as large as possible. Therefore, we construct In this section, we first introduce the proposed training Eq. 1 with the sum of the modulus of ■h■ and ■h+■ as the objective for semantic representation tensor constraints, then denominatorandthemodulusofthedifferencevectorsofboth introduce the cross-attention structure design, and finally de- pairsasthenumerator.Duringtraining,bothsub-objectivesare scribe the proposed JTCSE's overall structure and the loss optimized simultaneously; from a quantitative point of view, function's design. L has no measure and can be combined with other loss TMC functions. A. Modulus length constraints of semantic tensor representa- To more rigorously justify L TMC, we make a simple tions transformation. Firstly, according to Fig. 2 and the cosine theorem, ■h■, ■h+■, and h – h+ can construct a closed Existingmethodsfortrainingsentenceembeddingrepresen- triangle, and ■h–h+■ can be rewritten as Eq. 2: tations of unsupervised contrastive learning usually evaluate the correlation of a pair of samples by their cosine similarity. (cid:113) During training, the model is trained on the distribution of (cid:13) (cid:13)h–h+(cid:13) (cid:13)= ■h■2+■h+■2–2·■h■·■h+■·cosγ, (2) JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 6 L can be rewritten as Eq. 3: where X denotes the attention network, X ∈ {Q,K,V}, N TMC (cid:113) denotes the source sub-encoder, N ∈ {I,II} , and i denotes ■h■2+■h+■2–2·■h■·■h+■·cosγ the ordinal number of the EncoderLayer that X comes from, L TMC = ■h■+■h+■ . (3) i∈[1,12]4. We first specify the location of the cross-attention layer, as Since the tensor modulus of the samples are all larger than shown in Eq. 7: zero, there exists Eq. 4: L={i|i∈Z,1≤i≤12,imodk =0}, (7) (cid:13) (cid:13)h+(cid:13) (cid:13)=k·■h■,k ∈(0,+∞). (4) where L denotes the set of all locations where the cross- Moreover,sincecosγ takesthevalueof[–1,1],lett=cosγ, attention appears in the EncoderLayer of sub-model, and k t∈[–1,1], so Eq. 3 can be further rewritten as Eq. 5: isahyperparameterdenotingthepresenceofacross-attention √ EncoderLayer(abbreviatedasCAEL,showninFig.4a)every 1+k2–2·k·t L (k,t)= . (5) k EncoderLayers. TMC 1+k HDj andHDj denotetheoutputofthej-thEncoderLayer I II More intuitively, we visualize the binary function L TMC as when the hidden state is forward propagated within the two shown in Fig. 3. encoders. When j ∈ L, HDj, and HDj are to perform the I II It can be seen from Fig. 3 that when L TMC obtains the cross-attentioncomputation,andforEncoderI,theQj I andK Ij minimum value, k = 1 and t = 1, i.e., ■h■ = ■h+■ and networks first process HDj–1 and compute the self-attention I cosγ =1 , which is in accordance with our intended training matrix SAi, as shown in Eq. 8: I objective,meaningthatthetensorsofpositivesamplesofeach (cid:113) (cid:13) (cid:13)(Qj–1×HDj–1)·(Kj–1×HDj–1)T other should have similar modulus and should have similar SAj =softmax I I √ I I , orientations. Thus, we demonstrate that the proposed L I d TMC is consistent with our first propose motivation. (8) where d denotes the hidden state dimension. Then, the output of Vj is weighted

to obtain the context tensor CTj inside B. Designing for Cross-Attention I I Encoder I, as shown in Eq. 9: We observe that BERT-like models suffer from attention CTj =SAj ·(Vj ×HDj−1). (9) sinking, where the model disproportionately focuses on the I I I I SEPtokenorpunctuationattheendofthesentenceratherthan Meanwhile,SAi alsoweightstheoutputofVj inEncoderII I II the CLS token in deeper encoderlayers, which is detrimental to obtain the cross-attention context tensor(CACT), as shown to unsupervised sentence embedding models relying on CLS in Eq. 10: pooling. To quantify this, we define the CLS energy weight CACTj =SAj ·(Vj ×HDj−1). (10) E defined as Eq. 6: I I II II CLS The above operation is the same for Encoder II. According ■h ■ E = cls 2, (6) tothepreviousanalysis,SAj mayfocusonlocalinformation. CLS ■H ■ I − F In contrast, Vj ×HDj−1 focuses on global information, so II II which represents the ratio of the CLS token's 2-norm to the theCLStokenenrichedwithsemanticinformationinEncoder Frobenius norm of other tokens' hidden states in the context I not only extracts information from this encoder's local tensor. A higher E indicates richer semantic aggregation context but also obtains complementary information from the CLS by the CLS token and correlates with better sentence embed- global context of Encoder II. The vice versa is valid for the ding performance. To enhance E , we introduce a cross- CLS token in Encoder II. This aggregation of multi-source CLS attention mechanism within a twin-encoder architecture. This information makes the hidden state of the CLS token more mechanismenablesinteractionbetweenencodersbyusingone diversified, which enhances the E and the model's focus CLS encoder's attention weights to weigh the other's Value tensor, on the CLS token. which enriches the CLS representation with complementary semantic information without disrupting the original attention C. Model Structure Design distributionorBERT'spre-trainedknowledge,thusimproving In this subsection, we introduce the twin-tower ensemble the global representation quality of the CLS token. model JTCSE based on tensor modulus constraints and cross- We describe briefly the motivation for introducing cross- attention, as shown in Fig. 4b. The large-scale ensemble attention through the above3, and the process of computing modelEDFSE-BERTusessixSimCSE-BERT-basesfine-tuned cross-attention will be described in detail in the next step. with multi-language RTTs, which are designed to enrich In order not to cause additional training overhead, we do the distribution of textual semantic representations with an not introduce other naive network weights and only consider "intrinsic rank," which in turn enhances the model's ability the attention weights inside the EncoderLayer within both to discriminate between two similar sentences. We follow the sub-encoders. We define the network weights involved in EDFSEapproachbutcomposetheensemblemodelusingonly computing the cross-attention in each sub-encoder as Xi , twosub-encodersaugmentedwithRTTdatatoreducethehuge N inference overhead. 3IntheDiscussionsection,wewillelaborateonthemotivationsanddetails ofcross-attention. 4ThisworkisorientedtotheBERT-basemodelwith12EncoderLayers.

JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 7 ICTM Loss Add&Norm FFN Pooler Layer Pooler Layer Add&Norm Add & Output MHA MHCA ICNCE Q, K, V Q, K Pooler Pooler Q, K V Cross-Attention Encoder CAEL Layer(CAEL) Encoder Layer Encoder Layer Q, K V CAEL Encoder Layer Embedding Layer Embedding Layer Last Encoder Layer I&II Pooler Output Last Hidden State a. hP hP I II ■ ■ hP hP c. I II V b. V K, Q CAEL .... hL hL I II .... ■InfoNCE V K, Q ..... C ..A .E ..L . hL■ hL■ I II ■Hidden State Encoder Layer Encoder I Encoder II Primitive Hidden State Propagation Cross-Attention Propagation Inputed Mini-Batch Sentences ■ ■ ICTM Interaction Constraint on the Tensor Modulus Loss ICNCE Interaction Constrained InfoNCE Fig. 4. This figure shows the structure of the proposed unsupervised sentence embedding representation framework, JTCSE, which consists of two main parts: the semantic representation tensor modulus constraints and the joint modeling of subencoder cross-attention. Subfigure b. shows the overall structure ofJTCSE,whi chcontainstwosubencoders,IandII.EachisafineBERT-likemodelthatincludesanembeddinglayer,an encoder,andapoolerlayer. Before the training, we specify the cross-attention encoder layer's(CAEL) position in the encoder, the position of CAEL in both subencoders is the same. During training, a mini-batch is fed into the embedding layer of two sub-encoders simultaneously, and the hidden state output from each embedding layer goesintoitsownencoder;ifCAELisencountered,inadditiontothenormalf orwardpropagationwithineachsub-encoder,itisalsonecessarytomutually passthroughtheattentionnetwork ineachother'sEncoderLayertoachievethecomputationofcross-attention.Boththeprimitivelasthiddenstate(LHS)and thecross-attention'sLHSpassthroughtheIC-InfoNCEconstraints.TheprimitiveLHSalsopassesthr oughthepoolerlayertogetthepooleroutput,which inturnpassesthroughthetensormodulus-constrainedlossf unction.Subfigurea.representsthedetailsofCAEL,theQuery,Key,andValueweightsinMHA andMHCAareidentical.Subfigurec.representsthedetailsofICTMlossandIC-InfoNCEloss. 1) Loss of Cross-Attention and Model Continuation the temperature coefficient τ = 0.05 according to SimCSE

Training: EachencodercontainsEmbeddingLayers,Enocder- and derived work, sim(·) denotes the cosine similarity, and Layers, and PoolerLayers. We denote the two sub-encoders in ICNCE is defined as Eq. 12. JTCSE-BERT or JTCSE-RoBERTa as Encoder I and Encoder L =R·(cid:0) L (cid:0) hL,hL(cid:1) +L (cid:0) cO,cO(cid:1)(cid:1) II, respectively. Since each sub-encoder contains a Dropout ICNCE NCE I II NCE I II function, a sample will be encoded twice by each of the +(1–R)·(cid:0) L NCE(cid:0) hL II,hL I(cid:1) +L NCE(cid:0) cO II,cO I (cid:1)(cid:1) (12) two sub-encoders after entering JTCSE, resulting in a total R ∈ {0,1} denotes a binary random number. cO and cO of four tensor representations. We use hL, hL+, hL, and hL+ I II I I II II denote the outputs of the last CAEL sourced from Encoder to represent the CLS pooling of last hidden state from two I and Encoder II. During the training process, we still employ sub-encoders. Since these four features represent the same InfoNCE to continue training Encoder I and Encoder II. We sample, they are mutually positive samples. We design the employInfoNCEtofurtheroptimizetherepresentationquality interaction-constrainedInfoNCE(ICNCE)basedonInfoNCE. of the CLS token by maximizing the similarity between pos- the InfoNCE is represented as Eq. 11. itive sample pairs while uniformizing the similarity between soft negative sample pairs, with the aim of keeping the two sim(hi,h+ i) Encoders continuously trained. ICNCE is designed to allow L (cid:0) h ,h+(cid:1) =–log e τ , (11) the CLS token to focus more on the comprehensive semantic NCE i i esim(h τi,h+ i) +(cid:80) esim(h τi,h+ i) features of the input sequence, thus enhancing its ability to serve as a global representation. where i ∈ {I,II} indicates from Encoder I or Encoder II. 2) Refinement of Tensor Modulus-Constrained Training h and h+ denote the representation of the current pair of Objective: Wefindthatifthetensormodulusconstraintlossis i i positive samples, and h– denotes the representation of the computedwiththelasthiddenstateofEncoderIandEncoder i current mini-batch of other soft negative samples. We set II, the features are normalized due to the passage through JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 8 the NormLayer of the last Encoderlayer, which makes the SICKR dataset.7 Following the previous work, we take the features normalized and loses the modulus features. However, Spearman correlation in the STS-B [31] validation set as the we observe that the last layer of the network of the BERT- checkpoint-saving metric. We report the training hyperparam- like model is a FFN named PoolerLayer; the last hidden eter settings and RTT setup details in Appendix I. state regained the modulus features after passing through the PoolerLayer. Thus, we naturally adopt the Pooleroutput as the B. Tasks input of the tensor modulus constraint loss5. Following the existing work, we first evaluate the model Further, the overall loss function will be in the form of a on seven English STS tasks(STS12 [26], STS13 [27], STS14 summation of several sub-loss terms. To avoid subsequently [28],STS15[29],STS16[30],STS-B[31],SICKR[32])with optimizing the factors of the sub-loss terms, we add dynamic complementary coefficients to Eq. 1: –logsim(cid:0) hL,hL(cid:1) . the SentEval [44] package; to assess the model's performance I II on STS tasks more comprehensively, we conduct experiments Amend to Eq. 13: ontheEnglishsubtasksofthreemultilingualSTStasks(STS17 (cid:13) (cid:13)hP –hP+(cid:13) (cid:13) [45], STS22.V2 [46], STSBenchmarkMultilingual8) with the L (cid:0) h ,h+(cid:1) =–log(cid:0) sim(cid:0) hL,hL(cid:1)(cid:1) (cid:13) i j (cid:13) , MTEB [47] package. In addition, we conduct a wide range of TMC i j I II (cid:13) (cid:13)hP(cid:13) (cid:13)+(cid:13) (cid:13)hP+(cid:13) (cid:13) sentence embedding related 0-shot downstream tasks through i j (13) the MTEB package, specifically including the multilingual where i,j ∈ {I,II} and i ■= j. hP I , hP I + , hP II, and hP II+ or cross-language semantic text similarity computation tasks denote the pooler output of these hidden states, respectively. STS22.V2 [46]9. Considering the limitation of computational Intuitively, when the tensor hP i and hP i + are in similar resources,werandomlyselected45textclassificationtasks,45 directions but have significant differences in modulus, the text retrieval tasks, 15 bi-textmining tasks and the currently coefficients logsim(·) are not significantly helpful, but the available data set of 14 text re-ranking tasks, totaling more partition (cid:13) (cid:13) (cid:13)hP i –hP j+(cid:13) (cid:13) (cid:13) compensates for the loss. The two than 130 subtasks, to demonstrate the robustness of JTCSE. ■hP■+■hP+■ i j product terms in Eq. 13 are jointly constrained when both the modulus and direction differences between hP and hP+ C. Experimental Results i i are large. We report the performance of JTCSE and baselines on the Sinceweneedtooptimizethetwosub-encodersjointly,we seven STS tasks in Table I, and overall, JTCSE-BERT and defineaninteractionconstraintonthetensormodulus(ICTM), JTCSE-RoBERTa outperform the other work. Since JTCSE is denoted as Eq. 14. a twin-tower structure, for a fair comparison with the single- tower model, we follow EDFSE and distill the knowledge L ICTM =L TN(cid:0) h I,h+ II(cid:1) +L TN(cid:0) h II,h+ I (cid:1) . (14) of JTCSE through MSE loss to a naive

BERT or RoBERTa denoted as JTCSE D. JTCSE D also outperforms the other The purpose of L is to strengthen the alignment of the ICTM single-tower baselines on average on the 7 STS tasks. Com-twosub-encoderstothemodulusofthepositivesampletensor pared to the multi-tower model EDFSE, JTCSE-BERT has in the high-dimensional space. onlyone-thirdoftheinferenceoverheadofEDFSE-BERTbut Finally, we define the complete loss function for JTCSE as outperforms EDFSE-BERT on the STS tasks, which proves shown in Eq. 15: the effectiveness of the proposed modules constraint loss and L= (cid:88) L (cid:16) hL,hL+(cid:17) +L +L . (15) cross-attention. NCE i i ICNCE ICTM In addition, since RankCSE [19] is not officially open- i∈{I,II} sourced10. RankCSE uses the same type of SimCSE-base We will prove the necessity of each component of Eq. 15 [6], SimCSE-large, and DiffCSE-base [11] in constructing the in ablation experiments. In addition, the ablation experiment teacher ensemble model, which relies on pre-existing work of on pooling proves that the last hidden states outperform the the same type rather than having a BERT-base or RoBERTa- pooleroutputs.Thus,duringinferenceinJTCSE,thetwosub- base autonomously trained to obtain the that directly compar- encoders will encode the input samples separately and sum ing RankCSE with other work of the same type may lead the two obtained last hidden states as the output of the whole to fairness discussions. Therefore, in order to make a fair model without the need for pooler layer processing. comparison with RankCSE, we compare the performance on the 7 STS task by ensemble learning of JTCSE and RankCSE IV. EXPERIMENTS 7Whenweusetheofficialcodeandretrainsomeexistingopen-sourcework, A. Setup wefindthattheresultsreproducedaccordingtothedefaulthyper-references of the official code are 2%∼3% lower than the reported results, and when In JTCSE, we employ two sub-encoders previously fine-additionalunsupervisedSICKRdatasetsareadded,thereproducedresultsare tuned by the RTT training set generated by the Google Trans- barelyequaltothereportedresults,andtobefair,weaddunlabeledSICKR late system and the unsupervised SimCSE. For the training datasets when training TNCSE and JTCSE, and reported the effects of the unlabeledSICKRdatasetintheablationexperiments. dataset, we choose a 1M wiki corpus6 and an unsupervised 8 9STS22.v2contains19subtasks,whichisupdatedonSTS22byremoving 5We will discuss this design motivation in more detail in the Discussion pairswhereoneoftheentriescontainsemptysentences. section. 10All downstream experiments on RankCSE are reproduced from third- 6 partyopensourcecode JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 9 TABLEI THEL EFTTABLEREPORTSTHERESULTSOFTHEJTCSEANDBASELINEEVALUATIONONTHESEVENSTS TASKS,ANDTHERIGHTTABLEREPORTSTHE PERFORMANCEOFTHEOPEN-SOURCEREPRESEN TATIVEBASELINEONTHEENGLISHSUBTASKOFTHETHREEMULTILINGUALSTSTASKS.■AND■ DENOTERESULTSDERIVEDFROMTHEORIGINALPAPERAND[6],RESPECTIVELY. SS INDICATESAFUSIONWITHEXISTINGUNSUPERVISED CHECKPOINTKNOWLEDGE. SS INDICATESTHEBESTRESULTONTHEMAINMETRICAVG.SINCERANKCSE[19]HASNOTOFFICIALLY OPEN-SOURCEDANYCODEORCHECKPOINTS,♣DENOTESTHERESULTOFATHIRD-PARTYOPEN-SOURCECODEREPLICATION.DDENOTES DISTILLATIONTOASINGLEENCODER. Model STS12 STS13 STS14 STS15 STS16 STSB SICKR 7Avg. STS17 STS22 STSBM 3Avg. BERT-base BERT-base BERT-base■ 39.70 59.38 49.67 66.03 66.19 53.87 62.06 56.70 - - - - BERT-whitening■ 57.83 66.90 60.90 75.08 71.31 68.24 63.73 66.28 - - - - IS-BERT[38]■ 56.77 69.24 61.21 75.23 70.16 69.21 64.25 66.58 - - - - SBERT-base(Sup)■ 70.97 76.53 73.19 79.09 74.30 77.03 72.91 74.89 - - - - ConSERT[5]■ 64.64 78.49 69.07 79.72 75.95 73.97 67.31 72.74 - - - - SimCSE[6]■ 68.40 82.41 74.38 80.91 78.56 76.85 72.23 76.25 83.90 59.74 82.45 75.36 DiffCSE[11]■ 72.28 84.43 76.47 83.90 80.54 80.59 71.23 78.49 80.15 61.84 84.56 75.52 ESimCSE[9]■ 73.40 83.27 77.25 82.66 78.81 80.17 72.30 78.27 85.63 61.33 80.15 75.70 ArcCSE[13]■ 72.08 84.27 76.25 82.32 79.54 79.92 72.39 78.11 - - - - InfoCSE[14]■ 70.53 84.59 76.40 85.10 81.95 82.00 71.37 78.85 85.05 55.51 85.49 75.35 PromptBERT[17]■ 71.56 84.58 76.98 84.47 80.60 81.60 69.87 78.54 51.33 50.58 43.75 48.55 PCL[39]■ 72.84 83.81 76.52 83.06 79.32 80.01 73.38 78.42 86.32 63.10 83.83 77.75 SNCSE[16]■ 70.67 84.79 76.99 83.69 80.51 81.35 74.77 78.97 53.11 54.77 55.55 54.48 WhitenedCSE[40]■ 74.03 84.90 76.40 83.40 80.23 81.14 71.33 78.78 85.15 60.83 84.50 76.83 PromCSE[41]■ 73.03 85.18 76.70 84.19 79.69 80.62 70.00 78.49 - - - - EDFSE[1]■ 74.48 83.14 76.39 84.45 80.02 81.97 72.83 79.04 - - - - TNCSE 75.52 83.91 77.57 84.97 80.42 81.72 72.97 79.58 85.78 61.45 84.14 77.13 JTCSE 74.95 84.21 77.79 84.75 80.41 81.88 73.92 79.70 85.88 62.79 85.40 78.02 EDFSED[1]■ 74.50 83.61 76.24 84.02 80.44 81.94 74.16 79.27 - - - - TNCSED 75.42 84.64 77.62 84.92 80.50 81.79 73.52 79.77 85.36 63.79 85.41 78.19 JTCSED 75.01 84.86 77.76 84.62 80.38 82.05 74.53 79.89 85.65 63.59 85.55 78.26 RankCSE[19]♣ 74.61 85.70 78.09 84.64 81.36 81.82 74.51 80.10 85.88 62.46 62.46 70.26

RankCSE+UC 73.29 85.90 78.16 85.90 82.52 83.13 73.36 80.32 86.19 59.24 86.28 77.24 TNCSE+UC 75.79 85.27 78.67 85.99 82.01 83.16 73.01 80.56 86.51 62.05 86.35 78.30 JTCSE+UC 75.44 85.34 78.75 85.93 82.00 83.21 73.52 80.60 86.89 61.87 86.35 78.37 RankCSE+UCD 72.99 85.72 77.73 84.93 81.86 82.43 74.35 80.00 81.62 60.58 81.77 74.66 TNCSE+UCD 75.95 85.31 78.50 85.69 81.86 83.03 73.89 80.60 86.38 63.28 86.24 78.63 JTCSE+UCD 75.22 85.46 78.50 85.50 81.55 83.02 74.24 80.50 86.28 63.06 83.01 77.45 RoBERTa-base RoBERTa-base RoBERTa-base■ 40.88 58.74 49.07 65.63 61.48 58.55 61.63 56.57 - - - - RoBERTa-whitening[42]■ 46.99 63.24 57.23 71.36 68.99 61.36 62.91 61.73 - - - - SimCSE[6]■ 70.16 81.77 73.24 81.36 80.65 80.22 68.56 76.57 81.80 58.23 84.45 74.83 DiffCSE[11]■ 70.05 83.43 75.49 82.81 82.12 82.38 71.19 78.21 82.21 60.90 84.99 76.03 ESimCSE[9]■ 69.90 82.50 74.68 83.19 80.30 80.99 70.54 77.44 83.15 60.79 85.36 76.43 PromptBERT[17]■ 73.94 84.74 77.28 84.99 81.74 81.88 69.50 79.15 74.57 53.65 70.46 66.23 PCL[39]■ 71.13 82.38 75.40 83.07 81.98 81.63 69.72 77.90 81.80 61.58 85.25 76.21 SNCSE[16]■ 70.62 84.42 77.24 84.85 81.49 83.07 72.92 79.23 77.26 59.26 83.11 73.21 WhitenedCSE[40]■ 70.73 83.77 75.56 81.85 83.25 81.43 70.96 78.22 - - - - IS-CSE[43]■ 71.39 82.58 74.36 82.75 81.61 81.40 69.99 77.73 - - - - EDFSE[1] 72.67 83.00 75.69 84.07 82.01 82.53 71.92 78.84 - - - - TNCSE 74.11 84.00 76.06 84.80 81.61 82.68 73.47 79.53 84.05 62.70 83.03 76.59 JTCSE 74.92 84.22 77.08 84.69 81.39 82.60 74.03 79.94 83.73 63.78 86.33 77.95 EDFSED[1] 71.04 81.08 77.04 83.08 81.96 82.36 74.54 78.73 - - - - TNCSED 74.56 84.74 76.30 84.89 81.70 83.01 74.18 79.91 84.01 64.06 86.25 78.11 JTCSED 75.42 85.36 77.31 85.04 81.72 82.91 74.46 80.32 83.18 64.69 86.49 78.12 RankCSE[19]♣ 69.09 81.15 73.62 81.31 81.43 81.22 70.08 76.84 81.29 58.63 84.11 74.68 RankCSE+UC 74.18 84.06 77.72 83.26 79.81 81.25 72.58 78.98 82.39 60.32 85.77 76.16 TNCSE+UC 74.52 85.26 77.63 85.85 82.62 83.65 73.35 80.41 85.49 62.00 86.80 78.09 JTCSE+UC 74.57 85.73 78.17 85.78 82.73 83.73 73.52 80.61 85.30 62.87 86.82 78.33 RankCSE+UCD 68.55 82.23 73.61 81.28 81.28 80.98 71.01 76.99 82.57 60.12 84.32 75.67 TNCSE+UCD 74.14 83.86 76.09 84.07 81.59 82.90 73.55 79.46 84.04 62.72 86.07 77.61 JTCSE+UCD 74.92 85.14 77.07 84.59 81.71 83.18 74.50 80.16 83.79 64.74 86.27 78.27 withthesameunsupervisedcheckpointInfoCSE[14],denoted MTEB evaluation package. We report the experiment results asJTCSE-UCandRankCSE-UC,respectively.Inaddition,we of text classification, text re-ranking, bi-textmining and mul- distillJTCSE-UCandRankCSE-UCtoasinglenaiveencoder tilingual semantic textual similarity in Table II, Table III, toobtainJTCSE-UCDandRankCSE-UCD,respectively;the Table IV, and Table V, in which we uniformly use MTEB's experimental results show that JTCSE outperforms RankCSE default "Main result" as the evaluation metrics; For the text on both -UC and -UC D. retrieval task, due to the large number of metrics, in order In order to broadly evaluate the zero-shot performance of to evaluate the performance of each model on the retrieval JTCSE and baselines on other natural language processing task more comprehensively, we adopt the following met- tasks, we conducted over 130 zero-shot tasks based on the rics, MAP@1/5/10, MRR@1/5/10, NDCG@1/5/10, PRECI-

TABLEII THISTABLEREPORTSTHERESULTSOFZERO-SHOTTESTINGFOR45TEXTCLASSIFICATIONS,WITHTHEOPTIMALRESULTSONEACHTASKBOLDEDAND THESUB-OPTIMALRESULTSUNDERLINED. Tasks Sim~ ESim~ Diff~ Info~ SN~ Whiten~ Rank~ TN~ JT~ JT~D JT~UCD AllegroReviews 24.15 23.72 25.25 24.13 24.38 24.94 24.89 23.59 23.51 24.05 24.21 AngryTweets 42.34 41.37 42.54 41.68 44.30 41.29 42.33 41.51 40.66 40.93 41.45 ContractNLIInclusionOfVerbally 53.96 64.75 53.96 61.87 66.19 48.92 52.52 56.83 58.27 54.68 52.52 ContractNLIPermissibleAcquirement 79.21 83.15 82.58 78.65 74.16 87.08 81.46 82.58 82.58 82.58 83.15 ContractNLIPermissibleDevelopment 78.68 88.24 85.29 85.29 79.41 90.44 83.09 79.41 85.29 82.35 85.29 CUADAntiAssignmentLegalBench 84.73 82.76 80.89 82.00 79.10 80.46 83.11 83.70 81.66 82.51 81.14 CUADExclusivityLegalBench 66.40 70.47 63.39 63.78 64.04 68.50 62.86 64.04 70.60 73.10 72.31 CUADNoSolicitOfCustomersLegalBench 84.52 84.52 79.76 76.19 77.38 84.52 84.52 82.14 84.52 83.33 83.33 CUADPostTerminationServicesLegalBench 60.02 57.43 55.94 60.64 59.78 58.91 57.55 57.92 61.14 60.89 59.28 CUADTerminationForConvenienceLegalBench 80.93 79.07 79.53 83.26 67.91 77.44 80.70 77.21 84.65 84.42 84.65 CzechSoMeSentiment 45.75 44.34 46.57 46.01 47.58 47.62 43.90 47.93 47.35 47.50 47.39 GujaratiNews 40.19 40.40 40.30 39.83 40.55 41.18 39.07 39.77 40.08 39.27 38.92 HinDialect 35.92 33.35 37.50 38.67 42.60 35.75 31.84 38.09 35.58 34.82 34.80 IndonesianIdClickbait 54.26 54.39 54.09 54.56 57.57 54.15 53.44 55.73 54.90 54.92 55.39 InternationalCitizenshipQuestionsLegalBench 57.47 57.32 56.59 62.01 53.96 56.74 54.54 54.93 55.96 56.40 57.96 KLUE-TC 21.16 20.39 21.88 22.37 23.34 22.06 21.41 22.13 21.27 21.40 21.86 Language 92.56 91.40 93.83 95.04 96.05 92.80 93.13 93.22 92.23 92.42 93.28

LearnedHandsDivorceLegalBench 76.00 80.67 75.33 69.33 64.67 80.67 83.33 82.00 85.33 84.67 84.00 LearnedHandsDomesticViolenceLegalBench 78.16 73.56 78.74 70.69 72.41 75.86 75.29 74.71 81.03 79.89 77.01 LearnedHandsFamilyLegalBench 70.75 72.41 68.65 71.48 64.99 68.85 71.53 76.95 79.25 79.98 78.08 LearnedHandsHousingLegalBench 74.76 73.34 70.70 60.30 64.21 70.12 70.61 71.88 68.41 67.38 68.95 MacedonianTweetSentiment 35.77 35.66 37.44 36.77 37.95 37.12 36.50 37.85 37.36 38.01 37.98 MarathiNews 36.24 36.68 37.33 37.47 37.52 38.23 35.74 37.04 37.30 37.04 37.63 MassiveIntent 33.57 26.77 29.47 30.92 29.74 28.61 33.57 16.96 37.38 29.46 29.37 MassiveScenario 35.86 28.34 31.02 34.90 31.49 30.82 34.94 20.84 37.50 30.55 30.42 NorwegianParliament 52.46 52.60 52.25 52.83 51.33 53.24 52.89 52.35 52.88 52.69 52.64 NYSJudicialEthicsLegalBench 47.95 47.60 45.55 48.97 49.66 44.86 47.95 50.68 50.68 49.66 48.29 OPP115DataSecurityLegalBench 71.21 71.96 70.91 73.16 58.17 69.94 75.19 73.54 74.51 75.64 74.06 OPP115DoNotTrackLegalBench 81.82 86.36 78.18 90.91 80.91 80.00 81.82 90.91 91.82 90.91 87.27 OPP115InternationalAndSpecific 73.98 80.51 78.88 79.08 76.73 78.27 77.04 76.33 74.18 73.37 75.00 OPP115PolicyChangeLegalBench 87.24 87.70 86.54 88.40 83.06 88.17 84.45 86.77 89.79 89.79 89.79 OPP115ThirdPartySharingCollectionLegalBench 65.85 65.16 65.66 64.72 60.06 62.70 66.48 64.78 66.23 64.65 65.79 OPP115UserChoiceControlLegalBench 72.77 70.18 73.35 72.96 74.00 73.42 73.67 73.03 72.64 72.90 73.61 OralArgumentQuestionPurposeLegalBench 22.44 21.79 21.47 19.87 24.04 23.08 21.79 25.32 24.68 25.00 23.08 PolEmo2 34.27 32.67 37.35 35.47 36.84 34.68 34.23 33.48 33.14 31.54 33.00 PunjabiNews 65.92 65.16 64.84 64.27 67.77 62.99 63.57 65.86 63.76 66.88 66.62 Scala 50.14 50.30 49.98 50.21 50.15 50.26 50.37 50.28 50.27 50.15 50.06 SCDBPTrainingLegalBench 62.80 59.37 59.37 56.99 51.72 55.94 63.32 62.27 64.38 61.74 61.74 SentimentAnalysisHindi 38.84 39.73 40.60 39.11 39.47 39.65 40.10 38.82 38.70 38.90 39.16 SinhalaNews 35.97 35.80 35.46 37.22 39.90 35.14 34.82 34.15 33.72 33.12 34.35 SiswatiNews 71.25 71.63 73.13 74.50 73.25 73.38 72.25 73.50 71.25 71.88 71.88 SlovakMovieReviewSentiment 52.71 53.85 53.07 53.45 51.18 52.78 53.19 53.65 53.43 53.21 53.34 TamilNews 18.25 18.21 18.25 18.97 19.27 18.50 17.79 18.36 18.05 17.87 17.76 TNews 16.14 16.01 16.25 16.76 16.56 16.45 16.56 15.28 15.19 15.43 16.02 TweetEmotion 26.47 26.19 28.48 28.32 29.40 28.56 26.08 26.72 27.23 27.15 27.55 Avg.Acc 55.37 55.49 55.07 55.42 54.11 55.22 55.23 55.22 56.67 56.11 56.03 SION@1/5/10,andRECALL@1/5/10,andreporttheresults11 V. ABLATIONSTUDIES in Table VI. On these four types of tasks, JTCSE and derived A. Alignment and Uniformity models are the best overall. [48]proposestwocriticalevaluationmetricsforevaluating the quality of embedding representations: Uniformity and Alignment. Uniformity means that the embedding represen- D. Significance Test tations of a mini-batch should be distributed as uniformly as Weusetheofficialopen-sourcecode,specifyrandomseeds possible on the unit hypersphere. Alignment means that two from 1 to 5, and report the average results for 7 STS. The embedding representations that are positive samples of each BERT-likesingleencoderissusceptibletorandomseeds;how- other should be distributed as similarly as possible on the ever, our proposed framework with a twin-encoder structure unithypersphere.Bothmetricsshouldbeassmallaspossible, performs stably, and the average result is close to result in denoted as Eq. 16 and Eq. 17, respectively. Table I. We report the significance test results in Fig. 6. l align ■E (x,x+~ppos)(cid:13) (cid:13)f(x)−f(x+)(cid:13) (cid:13)2 , (16) 11ComparedtoTNCSE'sselectionofMAP@10astheevaluationmetrics where ■·■ denotes the Euclidean paradigm, f(x) denotes the on30textretrievaltasksandJTCSE'sselectionof5categorieswithatotalof embedding of the sample x being projected. 15evaluationmetricstoevaluatethemodelperformanceon45textretrieval tasks,JTCSEaccomplishedamorecomprehensiveevaluationandperformed betteronabroaderrangeoftasks. l ■log E e−t■f(x)−f(y)■2 ,t>0, (17) uniform i.i.d JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 11 TABLEIII THISTABLEREPORTST HERESULTSOFTHEZERO-SHOTEVALUATIONOFTHE14TEXTRERANKINGTASKSINTHEMTEBPA CKAGE,WHICHAREALL THERERANKINGTASKSFORTHECURRENTLYAVAILABLEDATASET.THE OPTIMALRESULTSONEACHTASKAREBOLDED,ANDTHESUB-OPTIMAL RESULTSAREUNDERLINED. Tasks Sim~ ESim~ Diff~ Info~ SN~ Whiten~ Rank~ TN~ JT~ JT~D JT~UCD Alloprof 36.67 37.81 32.07 38.52 28.20 32.04 35.68 36.25 39.71 39.63 38.49 AskUbuntuDupQuestions 51.88 52.28 52.08 52.83 45.53 51.60 53.76 50.73 52.85 52.65 54.01 CMedQAv2 13.97 14.78 15.26 17.21 11.69 15.06 14.47 14.79 14.58 14.78 15.14 ESCI 80.58 80.28 80.49 80.36 78.05 80.47 80.57 79.75 80.17 80.51 80.54 MindSmall 28.68 28.86 29.34 29.18 26.14 28.10 29.45 28.65 28.76 28.75 28.92 MMarco 2.48 3.77 3.64 4.96 2.70 4.02 3.34 2.94 4.02 4.04 4.20 NamaaMrTydi 39.88 37.00 34.29 26.69 41.05 33.48 28.62 26.33 31.42 31.38 31.89 RuBQ 27.33 24.04 24.80 23.39 20.43 25.34 22.28 18.05 23.25 23.25 23.92 SciDocsRR 67.87 70.48 70.37 71.29 58.90

67.63 69.89 70.51 69.85 69.59 71.23 StackOverflowDupQuestions 39.56 40.63 42.77 44.21 31.07 42.63 41.18 39.93 41.75 41.75 43.35 Syntec 45.65 49.60 40.28 48.99 37.39 42.25 47.51 43.86 52.56 50.93 50.85 T2 55.20 55.87 56.27 56.71 52.10 56.16 55.59 55.32 56.78 57.34 56.87 VoyageMMarco 21.60 21.41 20.90 23.57 16.50 21.52 21.09 20.46 22.07 21.78 22.69 WebLINXCandidates 7.58 9.24 7.99 9.03 6.15 7.79 9.64 8.82 9.25 8.71 10.26 Avg.MAP 37.07 37.58 36.47 37.64 32.56 36.29 36.65 35.46 37.64 37.51 38.03 TABLEIV THISTABLEREPORTSTHERESULTSOFTHEZERO-SHOTEVALUATIONOFTHE15TEXTBI-TEXTMININGTASKSINTHEMTEBPACKAGE,WHICHAREALL THECLUSTERINGTASKSFORTHECURRENTLYAVAILABLEDATASET.THEOPTIMALRESULTSONEACHTASKAREBOLDED,ANDTHESUB-OPTIMAL RESULTSAREUNDERLINED. Tasks Sim~ ESim~ Diff~ Info~ SN~ Whiten~ Rank~ TN~ JT~ JT~D JT~UCD BUCC 0.55 1.54 0.54 0.60 0.12 0.25 0.62 2.58 2.38 2.23 1.56 BUCC.v2 3.40 4.98 3.33 4.27 1.52 2.78 4.21 7.24 7.15 7.13 6.70 DiaBla 4.08 5.55 3.71 4.36 2.07 3.56 3.80 6.95 6.61 6.55 4.98 Flores 4.82 5.50 4.18 3.75 2.74 3.44 5.04 5.56 5.49 5.36 4.92 IN22Conv 1.12 1.12 1.11 1.42 1.06 1.11 1.16 1.16 1.23 1.20 1.23 IN22Gen 2.35 2.75 2.50 3.97 2.01 2.67 2.95 2.78 2.98 2.89 3.09 LinceMT 15.44 15.65 15.53 15.22 4.43 16.22 14.45 16.30 16.98 16.68 17.10 NollySenti 18.76 19.78 19.01 22.33 10.12 19.44 18.95 22.35 22.61 21.85 22.05 NorwegianCourts 87.46 87.82 88.04 90.42 83.77 88.73 85.82 90.75 90.99 90.96 90.67 NTREX 8.70 9.85 7.82 6.81 5.08 6.98 8.96 10.58 10.48 10.18 9.19 NusaTranslation 45.52 45.93 44.61 50.36 50.31 48.33 44.13 48.85 49.33 46.60 47.14 Phinc 33.15 34.80 40.41 43.80 27.58 41.79 38.13 41.43 41.33 39.53 42.40 RomaTales 2.34 2.43 3.27 3.17 3.83 3.21 2.00 4.43 3.51 4.11 3.75 Tatoeba 3.25 3.56 3.27 3.61 1.74 3.21 3.43 4.31 4.23 4.09 4.00 TbilisiCityHall 0.71 0.95 0.56 1.22 0.03 0.59 1.46 1.17 1.30 1.41 1.30 Avg.F1 15.44 16.15 15.86 17.02 13.10 16.15 15.67 17.76 17.77 17.39 17.34 where x,y ~ p and t is set to 2. In Fig. 7, we report ensemble learning does not play a central role in JTCSE data the performance of the JTCSE and distillation model JTCSE becoming SOTA on 7 STS tasks. D and the other baselines on these two metrics; the JTCSE In addition, inference overhead is a key metric for practical series models outperform the other baselines overall. applications of the models, and we report the inference overhead(GMAC13)inTableVII.Inordertoquantifytheinference B. Impact of Ensemble Learning and Analysis of Inference efficiency of each model, we define a simple metric for Efficiency characterizing the model's performance per unit of inference overhead, defined as η = Score, where Score denotes the SinceJTCSEisatwin-towerstructure,eventhoughwehave Cost model's percentage of correctness on the seven STS test sets, obtained its distillation to a single-tower model JTCSE D, to and Cost denotes the model's inference complexity, reported compare more fairly with other baselines, we used the same in Table VII. Among the multi-tower models, first notice that trainingsetexpansionmethodasJTCSEforeachbaseline12 to EDFSE adopts a naive multi-tower ensemble, which has a trainthesub-encodersandensemblelearningtheobtainedtwo vast inference overhead and thus has a low performance per sub-encoders, we report the evaluation results of the different unitinferenceoverhead;comparedtootherwin-towermodels, ensemble models on the seven STS test sets in Table VII, in JTCSE has the highest performance per unit overhead as it addition, we report the evaluation results of JTCSE's direct performs the best on the 7 STS tasks, and among the single- ensemble learning of two subencoders before training. By tower models, as JTCSE D is a SOTA model, it has the η comparison, the performance of the two sub-encoders is not highest. optimal before being trained by JTCSE, and the performance is significantly improved after training, which indicates that 12All of each baseline using the same RTT strategy, with the original trainingsetbeingWIKI1M+unlabeledSICKR.

13WeusetheThoppackagetoevaluatetheinferenceoverheadofthemodel.

 TABLEV THISTABLEREPORTSTHEEVALUATIONRESULTSOFEACHMODELONTHEMULTILINGUALORCROSS-LANGUAGETASKSTS22.V2,WHICHCONTAINS 18SUB-TESTSETS,WITHZERO-SHOTEVALUATIONSFORALLLANGUAGESORCROSS-LANGUAGESEXCEPTEN.THEOPTIMALRESULTSFOREACH SUB-TESTSETAREBOLDED,ANDTHESUB-OPTIMALRESULTSAREUNDERLINED. Tasks SimCSE ESimCSE DiffCSE InfoCSE SNCSE WhitenedCSE RankCSE TNCSE JTCSE JTCSED JTCSEUCD ar 38.33 32.48 34.94 21.08 33.58 36.08 38.16 34.75 35.16 32.77 33.15 avg 32.82 36.79 34.37 28.09 23.64 32.71 38.49 39.33 39.21 37.76 37.82 de 24.70 28.50 24.47 18.02 2.58 24.99 24.70 22.05 27.86 28.36 28.99 de-en 13.13 29.80 33.63 37.03 20.73 30.33 37.52 33.10 36.33 30.84 34.76 de-fr 35.93 32.68 38.29 2.44 25.42 31.45 37.81 35.41 32.52 40.43 35.13 de-pl 18.82 12.78 11.30 -26.67 7.08 9.58 5.67 36.71 23.13 26.02 17.68 en 59.74 61.33 61.84 55.51 54.77 60.83 62.46 61.45 62.79 63.59 63.06 es 49.23 52.14 55.03 49.06 39.98 55.16 59.91 61.34 63.54 57.28 57.75 es-en 30.44 37.84 36.83 38.53 21.28 34.14 39.37 25.96 38.77 35.18 38.00 es-it 31.48 42.50 40.91 44.44 22.54 31.27 42.43 45.70

42.56 43.80 44.14 fr 61.55 61.31 60.06 52.95 31.47 52.96 64.85 67.70 64.22 66.91 65.50 fr-pl 39.44 50.71 -5.63 16.90 16.90 16.90 39.44 50.71 28.17 28.17 28.17 it 54.67 59.89 57.61 52.94 27.64 53.46 60.43 61.60 62.37 62.05 62.68 pl 22.79 26.72 23.77 8.23 6.78 23.42 31.00 29.67 30.46 26.24 26.05 pl-en 15.44 36.41 30.43 29.48 28.67 22.82 34.44 29.15 37.36 27.74 30.62 ru 15.71 17.87 24.03 6.77 14.03 24.59 21.70 26.26 23.03 20.29 20.34 tr 28.09 31.56 29.18 24.27 16.92 28.33 30.35 30.05 33.37 31.80 31.59 zh 46.42 37.76 48.78 47.06 40.12 40.45 50.65 39.64 48.23 41.81 42.71 zh-en 4.82 9.87 13.14 27.61 15.06 11.94 12.02 16.71 15.90 16.38 20.49 Avg.Acc 32.82 36.79 34.37 28.09 23.64 32.71 38.49 39.33 39.21 37.76 37.82 TABLEVI THISTABLEREPORTSTHERESULTSOFTHEZERO-SHOTE VALUATIONOFJTCSEANDBASELINEONMTEB,WITHTHENUMBEROFTASKSANDMETRIC INDICATEDINPARENTHESES.DETAILSOFTHERESULTSAREREPORTEDINAPPENDIXII. Metrics SimCSE ESimCSE DiffCSE InfoCSE SNCSE WhinenedCSE RankCSE TNCSE JTCSE JTCSED JTCSEUCD MAP@1 7.16 8.43 7.60 6.70 2.77 7.57 6.80 7.82 8.73 8.73 8.58 MAP@5 10.04 11.32 10.38 9.28 4.16 10.42 9.48 10.76 12.01 12.03 11.96 MAP@10 10.79 12.07 11.10 9.93 4.54 11.16 10.19 11.47 12.75 12.77 12.73 MRR@1 11.81 13.12 11.99 11.14 4.50 11.84 10.61 11.92 13.59 13.62 13.46 MRR@5 16.02 17.48 16.21 15.43 6.71 16.24 14.79 16.41 18.39 18.54 18.40 MRR@10 16.90 18.38 17.09 16.21 7.27 17.12 15.63 17.30 19.30 19.41 19.27 NDCG@1 11.62 12.83 11.76 10.94 4.42 11.66 10.50 11.68 13.36 13.38 13.25 NDCG@5 13.91 15.27 14.29 13.21 5.77 14.25 13.12 14.55 16.32 16.38 16.32 NDCG@10 15.13 16.56 15.44 14.23 6.51 15.42 14.30 15.72 17.45 17.49 17.52 PRECISION@1 11.82 13.12 11.99 11.18 4.49 11.85 10.64 11.93 13.59 13.63 13.46 PRECISION@5 6.14 6.54 6.34 6.14 2.52 6.28 5.89 6.26 7.08 7.15 7.13 PRECISION@10 4.64 4.92 4.72 4.61 1.88 4.71 4.47 4.60 5.20 5.19 5.24 RECALL@1 7.16 8.43 7.60 6.70 2.77 7.57 6.80 7.82 8.73 8.73 8.58 RECALL@5 14.81 16.24 14.96 13.52 6.61 15.02 13.94 15.64 17.30 17.38 17.49 RECALL@10 19.51 20.90 19.49 17.66 9.32 19.48 18.28 20.24 21.88 21.97 22.15 45Avg. 11.83 13.04 12.06 11.12 4.95 12.04 11.03 12.27 13.71 13.76 13.70 C. Impact of Unlabeled SICKR datasets SICKR dataset to the JTCSE is the smallest, which further We have found that when reproducing SimCSE, ESimCSE, demonstrates that the stability of the JTCSE is better. and DiffCSE, we cannot reproduce the results reported in the paper using the official open-source code and default D. The Ablation of Cross-Attention Structures hyperparameters. For example, our reproduction of SimCSE In the cross-attention structure, we set CAELs at equal is only about 74%, which is far from the reported 76.25%, intervals, which means we set one CAEL for every pass so we can only improve the reproduction level by adding an through the same number of EncoderLayers.Thus, in JTCSE, unsupervisedataset,basedonwhichweroughlyimprovethe the number of CAELs can be set to 1, 2, 3, 4, 6, and 12; if results of SimCSE to about 76% in order to be fair enough to we do not set CAELs, the model degenerates to TNCSE. We conduct the subsequent experiments. In Table VIII, we report report in Table IX the effect of setting different numbers of the results we obtained by training the model with the default CAELs on the training of JTCSE. hyperparametersandtheresultsbyaddingtheSICKRdataset. When the number of CAELs is small, the features do not In addition, we report the results of training JTCSE using the interact sufficiently across models to form an effective feature original Wiki1M to demonstrate that adding the unlabelled correction but instead may introduce error features, leading to datasetdoesnotsignificantlyimprovethemodelperformance. attenuation of the model effect; when the number of CAELs Meanwhile, we find that after the optimization of the cross- is too large, the features are assimilated prematurely between attention mechanism, the sensitivity of the SICKR dataset to sub-models, and it is unable to construct an adequate two- the JTCSE is smaller than the TNCSE's. The gain of the feature space for joint modeling, which then leads to the JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 13 TABLEVII THISTABLEREPORTST HEEXPERIMENTALRESULTS,INFERENCECOMPLEXITYANDINFERENCEEFFICIENCYOFDIFFER ENTMODELSIN7STS. Model Sim~ ESim~ Diff~ Info~ Rank~ EDFSE TN~ JT~ Sim~ ESim~ Diff~ Info~ Rank~ EDFSED TN~D JT~D EnsembleModel Single-TowerModel DistilledModel STSAcc 77.97 78.51 77.89 77.05 78.22 79.04 79.58 79.70 76.25 78.27 78.49 78.85 76.04 79.27 79.77 79.89 (7Avg.) InferenceComplexity 10.90 10.90 10.90 10.90 10.90 32.70 10.90 10.90 5.40 5.40 5.40 5.40 5.40 5.40 5.40 5.40 (GMAC) InferenceEfficiency 7.15 7.20 7.15 7.07 7.18 2.42 7.30 7.31 14.12 14.49 14.54 14.60 14.08 14.68 14.77 14.79 (η) ■■■a■■ ■■■a■■ ■■■a■■ ■■■a■■ ■■■a■■ ■■ ■■ ■■ ■■ ■■ ■%■(■5■7 ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■6■L■P■&■6■(■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■(■6■L■P■&■6■(■■■a■■ ■■■a■■ ■■■a■■ ■■■a■■ ■■■a■■ ■■ ■■ ■■ ■■ ■■ ■'■L■I■I■&■6■(■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■,■Q■I■R■&■6■(■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■■ ■6■1■&■6■(■■■a■■ ■■■a■■ ■■■a■■ ■■■a■■ ■■■a■■ ■■■a■■■■■■ ■■■■■■■■a■■■■■■ ■■■■■■■■a■■■■■■ ■■■■■■■■a■■■■■■

■■■■■■■a■■■■■■ ■5■D■Q■N■&■6■(■■ ■■ ■■ ■■ ■■ ■■■a■■■■■■ ■■■■■■■a■■■■■■ ■■■■■■■a■■■■■■ ■■■■■■■a■■■■■■ ■■■■■■■a■■■■■■ ■-■7■&■6■(■■ ■■ ■■ ■■ ■■ ■■■a■■■■■■ ■■■■■■■a■■■■■■ ■■■■■■■a■■■■■■ ■■■■■■■a■■■■■■ ■-■7■&■6■(■■■' Fig.5. ThisfigurereportsthecosinesimilaritydensitydistributionplotsofdifferentmodelsontheSTS-Bdataset,wheresentencepairsareuniformlydivided into five groups, with the vertical coordinates of each subplot denoting the group and the horizontal coordinates denoting the model scoring. Each subplot shouldhaveanoverall“sub-diagonal”distribution,indicatingclosertothelabelingdistribution.Weusethesamecodetoreporttheperformanceofallmodels. TABLEVIII 2) L ICNCE Only: Although ICNCE implements angle THEIMPACTOFADDINGUNLABELLEDSICKRDATASETSONMODEL constraints for positive and negative samples between twin TRAINING. encoders and introduces cross-attention hidden state mutual supervision constraints, it does not substantially introduce a Datsets Sim~ ESim~ Diff~ Info~ Rank~ TN~ JT~ Wiki1M 74.3 75.8 75.2 75.8 77.2 79.3 79.6 new training objective, so the effect improvement is still not +SICKR 75.9 76.7 78.0 77.4 77.5 79.6 79.7 obvious. Gain↓ 1.6 0.9 2.8 1.6 0.3 0.3 0.1 3) L Only: Since the training objective of the tensor ICTM modulus feature constraints proposed in Eq. 13 is oriented to overfitting phenomenon similar to that of the single-tower PoolerOutput,butweuseCLSPoolinginourinference,there model, and loses the significance of ensemble learning. is a margin between Pooler Output and CLS Pooling, so we cannot optimize CLS pooling directly, which in turn leads to TABLEIX an insignificant performance improvement. THEEFFECTOFTHENUMBEROFCAELSONJTCSETRAINING. 4) L + L : Again, since no new training ob- NCE ICNCE jective is introduced, strengthening the continuation training NumberofCAEL 1 2 3 4 6 12 w/o of each encoder and mutual supervision of twin encoders 7STSAvg. 79.23 78.74 79.57 79.37 79.70 79.55 79.58 based only on the constraints of the tensor direction does not substantially improve the effectiveness. 5) L +L : Underthejointeffectofthecontinua- NCE ICTM E. Ablation on the Loss Function tiontrainingofeachencoderonthetensordirectionconstraint We have proposed the loss function of JTCSE in Model andthetraininggoalofthemodelengthconstraint,themodel Structure Design, which consists of three parts: L , effecthasbeenimprovedtoacertainextent;however,without NCE L , and L .In this section, we analyze the gain the introduction of the cross-attention can not effectively ICNCE ICTM from each part of the loss function and the reason for it, and alleviate the phenomenon of attention sinking of the BERT- we report the ablation results of this section in Table x. like model, so the model effect needs to be further improved. 1) L Only: Since JTCSE employs a twin encoder 6) L +L : The model effect is significantly NCE ICNCE ICTM structure, the sentence embedding is modeled through en- improved with the combined effect of cross-attention to al- semble learning. If only unsupervised training of InfoNCE is leviate attention sinking and tensor modulus constraints on performed for each encoder, which does not directly improve the training objectives. Based on the ablation results obtained the ensemble learning, and result is improved insignificantly. earlier, the role of the L , except for the cross-attention ICNCE JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 14 79.5 78.5 77.5 76.5 75.5 74.5 73.5 Sim ~ E Sim ~ Diff~ Info ~ R a n k ~ T N ~ O urs .gvA STS 7 0.16 79.6 0.14 79.4 0.12 TN~ Ours 0.10 0.08 0.06 0.04 Mean score on 7 STS Standard Deviation 0.02 –1.5 –1.0 –0.5 Uniformity Fig.6. Thisfigurereportsthemeanandvariance of the results of the significance test experiments withrandomseedselectionrangingfrom1to5. tnemngilA 80 SBERT-sup(74.89) ■■■■ 75 ■■■■ 70 ■■■■ BERT(56.70) 65 Sim~(76.25) ■■■■ SN~(78.97) 60 JT~ D(79.89) JT~(79.70) ■■■■ Info~(78.85) ESim~(78.27) 55 ■■■■ Diff~(78.49) 50 ■■■■ ■■■■ ■■■■ ■■■■ ■■■■ ■■■■ ■■■■ ■1■R■■■'■H■O ■'■H■O■H■W■H■■■/■D■\■H■U■1■R■U■P■■■6■W■D■U■W■■ ■3■R■V■L■W■L■R■Q■■■R■I■■■(■Q■F■R■G■H■U Fig.7. Thisfigurereportsuniformityandalignment metrics for JTCSE and baselines. Both of these metricsshouldbeassmallaspossibleanddistributed asclosetothebottomleftaspossibleinthisfigure. ■P■U■R■1■U■H■\■D■/■■■J■Q■L■W■H■O■H■'■■■U■H■W■I■D■■■U■R■V■Q■H■7■■■F■L■W■Q■D■P■H■6■■■I■R■■■H■O■X■G■R■0 ■[Equation: Mathematical expression: $■E■O■D■W■L■R■Q■■■R■I■■■/■D■\■H■U■1■R■U■P■■■'■H■O■H■W■L■R■Q■■■L■Q■■■(■Q■F■R■G■H■U ■-■7■&■6■(■ ■V■■■5■H■V■X■O■W■■■R■Q■■■■■■■6■7■6 ■7■1■&■6■(■ ■V■■■5■H■V■X■O■W■■■R■Q■■■■■■■6■7■6 ■■ ■■ ■■ ■■ ■■ ■3■H■U■I■R■U■P■D■Q■F■H■■■R■Q■■■W■K■H■■■■■■■6■7■6 ■0■H■D■Q■■■0■R■G■X■O■H■■■I■R■U■■■-■7■&■6■(■0■H■D■Q■■■0■R■G■X■O■H■■■I■R■U■■■7■1■&■6■(Fig. 8. This figure reports the performance of JTCSE and TNCSE on 7 STS and the average modulus length of the output hidden states after removingsomeoftheLayerNorms. TABLEX and find that no matter how much the semantics

of the input THISTABLECOMBINESEACHOFTHELOSSFUNCTIONSTOEXPLORETHE sentences differ, the modulus of their output hidden state CONTRIBUTIONOFEACHTOMODELTRAINING.NONEDENOTESADIRECT representations are always distributed in the range of 14~16, ENSEMBLEOFTWOSIMCSE-TRAINEDENCODERS.ALLEXPERIMENTS USECLSPOOLINGMETHOD. which makes the modulus constraints of the hidden state representations ineffective to be applied. LossChoice 7STSAvg. We further explore removing some of the LayerNorms None(TwinEncoderUntrained) 78.27 duringtrainingtoobtainthehiddenstatemodulusfeatures.By LNCE 78.50 analyzing the output of 100 random sentences in Wiki100M, LICNCE 78.71 LICTM 78.40 we find that removing the LayerNorm enhances the model's LNCE+LICNCE 78.55 hidden state modulus features, but this operation significantly LNCE+LICTM 79.10 damages the model's performance. Specifically, when training LICNCE+LICTM 79.62 LNCE+LICNCE+LICTM(Ours) 79.70 the TNCSE and JTCSE models, we gradually remove the LayerNorms in the penultimate 1 to 6 EncoderLayers and use the last hidden state after CLS pooling as the input for the modulus-constrained loss. Fig. 8 shows the experimen- constraints, may overlap with that of the L used to con- NCE tal results. As the number of removed LayerNorm layers tinue training. Thus, the model's performance in this setting increases, the model's performance on the seven STS tasks is close to the final. shows a systematic degradation. We hypothesize that this phenomenonstemsfromremovingLayerNorm,destroyingthe VI. DISCUSSION keyinformationlearnedinthepre-trainingstageoftheBERT- A. Reasonableness of the Tensor-Module Constrained Train- base, and seriously degrading the model's semantic extraction ing Objectives' Design capability. This finding suggests that although the presence of In our proposed training objective for tensor-constrained LayerNorm may limit the module feature of hidden states, it modulus length, we use the Pooler Outputs obtained after is crucial for maintaining the core capabilities of pre-trained the last hidden state passes through the Pooler Layer for language models. training. However, intuitively, it is more appropriate to use The above discussion illustrates that the module's features CLS pooling of the last hidden state for modulus length of the hidden state cannot be obtained by removing the Lay- constraintsconsistentwiththeinferencepoolingapproach,and erNorm.We further investigate the structure of the BERT-like inthissection,wediscusswhyPoolerOutputsareusedinstead model and find a forward neural network named Pooler Layer of the last hidden state. after the last layer of EncoderLayer.To our knowledge, even The BERT-like models are all Encoder-Only structures though the Pooler Layer contains the BERT-like pre-training and each EncoderLayer is structured like an encoder in a information, all BERT-like unsupervised sentence embedding Transformer; specifically, each EncoderLayer contains a Self- modelsdonotutilizethePoolerLayer,whichwastesvaluable Attention Mechanism module, a Forward Neural Network, pre-training knowledge. We utilize the Pooler Layer and find and two LayerNorm. LayerNorm is a normalization layer that that the Pooler Output obtained by processing the last hidden normalizes the mean and variance of different hidden states. state through the Pooler Layer is characterized by modulus. Due to LayerNorm, the forward-propagated hidden states will Specifically, for different input sentences, the corresponding lose their modulus features. We observe SimCSE-BERT-base PoolerOutputhasanextensivedistributionofmodulus,which
■@■6■/■&■> ■V■L■K■W
■V■L ■Q■D ■H■O■S■P■D■[■H ■H■F■Q■H■W■Q■H■V ■U■R■I ■O■D■X■V■L■Y
■J■Q■L■]■L■■■■ ■Q■R■L■W■Q■H■W■W■D ■V■H■U■R■F■V ■■ ■@■3■(■6■>
■>■&■/■6■@ ■W■K■L■V ■L■V ■D■Q ■H■[■D■P■S■O■H ■V■H■Q■W■H■Q■F■H
■I■R■U ■Y■L■V■X■D■O ■■■■■L■]■L■Q■J ■D■W■W■H■Q■W■L■R■Q
■V■F■R■U■H■V ■■ ■>■6■(■3■@ ■@■6■/■&■> ■V■L■K■W ■V■L ■Q■D
■H■O■S■P■D■[■H ■H■F■Q■H■W■Q■H■V ■U■R■I ■O■D■X■V■L■Y ■J■Q■L■]■L■■■■
■Q■R■L■W■Q■H■W■W■D ■V■H■U■R■F■V ■■ ■@■3■(■6■> ■@■6■/■&■> ■V■L■K■W
■V■L ■Q■D ■H■O■S■P■D■[■H ■H■F■Q■H■W■Q■H■V ■U■R■I ■O■D■X■V■L■Y
■J■Q■L■]■L■■■■ ■Q■R■L■W■Q■H■W■W■D ■V■H■U■R■F■V ■■ ■@■3■(■6■>
■@■6■/■&■> ■V■L■K■W ■V■L ■Q■D ■H■O■S■P■D■[■H ■H■F■Q■H■W■Q■H■V
■U■R■I ■O■D■X■V■L■Y ■J■Q■L■]■L■■■■ ■Q■R■L■W■Q■H■W■W■D
■V■H■U■R■F■V ■■ ■@■3■(■6■> ■@■6■/■&■> ■V■L■K■W ■V■L ■Q■D
■H■O■S■P■D■[■H ■H■F■Q■H■W■Q■H■V ■U■R■I ■O■D■X■V■L■Y ■J■Q■L■]■L■■■■
■Q■R■L■W■Q■H■W■W■D ■V■H■U■R■F■V ■■ ■@■3■(■6■> ■@■6■/■&■> ■V■L■K■W
■V■L ■Q■D ■H■O■S■P■D■[■H ■H■F■Q■H■W■Q■H■V ■U■R■I ■O■D■X■V■L■Y
■J■Q■L■]■L■■■■ ■Q■R■L■W■Q■H■W■W■D ■V■H■U■R■F■V ■■ ■@■3■(■6■>

■%■(■5■7 ■>■&■/■6■@ ■W■K■L■V ■L■V ■D■Q ■H■[■D■P■S■O■H ■V■H■Q■W■H■Q■F■H ■I■R■U ■Y■L■V■X■D■O ■■■■■■L■]■L■Q■J ■D■W■W■H■Q■W■L■R■Q ■V■F■R■U■H■V ■■ ■>■6■(■3■@ ■6■L■P■&■6■(■>■&■/■6■@ ■W■K■L■V ■L■V ■D■Q ■H■[■D■P■S■O■H ■V■H■Q■W■H■Q■F■H ■I■R■U ■Y■L■V■X■D■O ■■■■■■L■]■L■Q■J ■D■W■W■H■Q■W■L■R■Q ■V■F■R■U■H■V ■■ ■>■6■(■3■@ ■(■6■L■P■&■6■(■>■&■/■6■@ ■W■K■L■V ■L■V ■D■Q ■H■[■D■P■S■O■H ■V■H■Q■W■H■Q■F■H ■I■R■U ■Y■L■V■X■D■O ■■■■■■L■]■L■Q■J ■D■W■W■H■Q■W■L■R■Q ■V■F■R■U■H■V ■■ ■>■6■(■3■@ ■'■L■I■I■&■6■(■>■&■/■6■@ ■W■K■L■V ■L■V ■D■Q ■H■[■D■P■S■O■H ■V■H■Q■W■H■Q■F■H ■I■R■U ■Y■L■V■X■D■O ■■■■■■L■]■L■Q■J ■D■W■W■H■Q■W■L■R■Q ■V■F■R■U■H■V ■■ ■>■6■(■3■@ ■,■Q■I■R■&■6■(■>■&■/■6■@ ■W■K■L■V ■L■V ■D■Q ■H■[■D■P■S■O■H ■V■H■Q■W■H■Q■F■H ■I■R■U ■Y■L■V■X■D■O ■■■■■■L■]■L■Q■J ■D■W■W■H■Q■W■L■R■Q ■V■F■R■U■H■V ■■ ■>■6■(■3■@ ■6■1■&■6■(■>■&■/■6■@ ■W■K■L■V ■L■V ■D■Q ■H■[■D■P■S■O■H ■V■H■Q■W■H■Q■F■H ■I■R■U ■Y■L■V■X■D■O ■■■■■■L■]■L■Q■J ■D■W■W■H■Q■W■L■R■Q ■V■F■R■U■H■V ■■ ■>■6■(■3■@ ■/■D■\■H■U■■■■ ■/■D■\■H■U■■■■ ■/■D■\■H■U■■■■ ■/■D■\■H■U■■■■ ■/■D■\■H■U■■■■■■ ■/■D■\■H■U■■■■■■ ■:■K■L■Q■H■Q■H■G■&■6■(■■■■■■ ■■■■■■ ■■■■■■ ■■■■■■ ■■■■■■ ■$]■W■W■H■Q■W■L■R■Q■■■6■F■R■U■H■V Fig.9. Visualizationofatten tionscoresacrossdifferentlayersandmodels(SimCSE,ESimCSE,DiffCSE,InfoCSE,SNCSE)fortheinputse ntence:’Thisis anexamplesentenceforvisualizingattentionscores.’Eachsubplotrepresentstheaverageatte ntionweightsofaspecificlayer.Ascanbesummarizedfrom thisfigure,thereisaclearattentionsinkfortheserepr esentativemodels,whichdonotdistributetheattentionweightstothefeaturewordsinthesentence. is almost unlimited, and it makes sense to perform modulus TABLEXI length constraints on this basis. THISTABLEREPORTSTHEEFFECTOFTHEPOOLINGMETHODONTHE Therefore,basedontheabovediscussion,wefinallyusethe MODEL’SPERFORMANCEONTHE7STSTASK.AVGISTHEAVERAGEOF ALLTOKENHIDDENSTATESINTHELASTHIDDENSTATE;AVG(FL)ISTHE Pooler Output as the input of the tensor modulus constraint AVERAGEOFALLTOKENHIDDENSTATESINTHEFIRSTANDLAST instead of the last hidden state. LAYERS;ANDPOOLERISTHEOUTPUTOFTHEPOOLERLAYER.ALL CHECKPOINTSAREDERIVEDFROMOFFICIALOPENSOURCE. B. Detailed Motivations for Cross-Attention Design PoolingMethodSimESimDiffInfowhitened SN PCLPrompt JT JT-D Byvisualizingtheattentionweights,weobservethatBERT- CLS 76.3 78.3 78.578.9 78.8 79.078.4 41.1 79.779.9 like models almost universally exhibit the attention-sinking Avg 76.2 77.3 76.578.5 76.5 68.976.9 66.3 70.371.6 Avg(FL) 75.5 75.5 72.474.5 72.4 69.774.1 66.6 78.478.4 phenomenon, as shown in the Fig. 9. The BERT-like model Pooler 75.3 67.2 78.278.5 78.1 50.778.0 22.7 70.371.6 JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 16 ■■■■■■■■■ ■■■■■■■■■ ■■■■■■■■■ ■■■■■■■■■ ■■■■■■■■■ ■■■■■■■■■ ■■■■■■■■■ ■,■Q■I■R■&■6■(■(■6■L■P■&■6■(■6■1■&■6■(■'■L■I■I■&■6■(■:■K■L■Q■H■Q■H■G■&■6■(■7■1■&■6■(■-■7■&■6■(■7■1■&■6■(■■■' ■-■7■&■6■(■■■' ■ ■■■ ■■■@■6■/■&■>■■■U■R■I■■■V■W■K■J■L■H■:■■■\■J■U■H■Q■(■■■■■■■■■■■ ■(■Q■H■U■J■\■■■:■H■L■J■K■W■V ■■■■■■■■■■■ ■(■Q■H■U■J■\■■■:■H■L■J■K■W■V ■[Equation: Formula: $■F■F■X■U■D■F■\■■■7■U■H■Q■G ■■■ ■■■■■■ ■■■■■6■7■6■■■$]■F■F■X■U■D■F■\ ■■■ ■■■■■■ ■■■■■■■■■■■ ■■■■■■■■■■■ ■■■■■■■■ ■■ ■■■ ■■■■■■ ■■■■■■■■■■■ ■■■ ■■■■■■ ■■■ ■■■■■■■ ■■■■■■■■■■■ ■■■■■■■■■■■ ■■■■■■■■■■■ ■■■■■■■■■■■ ■■■6■7■6■■■$■F■F■X■U■D■F■\■■■ ■■■ ■■■ ■■■ ■■ ■■■ ■■■■■■■ ■■■ ■■■■■■■ ■■■ ■■■■■■■ ■■■ ■■■■■■■ ■■■■■■■■■■■ ■■■■■■■■■■■ ■■■■■■■■■■■ ■■■■■■■■■■■ Fig.10. Thisfigureillustratesthatthemodel’sperformanceonthe7STStaskisapproximatel ypositivelycorrelatedwiththeenergyweightoccupiedbythe CLStokeninthemodel’sall-attentionresults. usually pays too much attention to the SEP token or the Specifically, the Value of the external encoder provides punctuation at the end of the sentence instead of the CLS additional contextual information that allows the CLS token tokeninthedeepEncoderLayer.However,alltheunsupervised of the current encoder to capture a richer representation sentence embedding models use CLS pooling, and the CLS of semantic features from another encoder. The reason for token is not focused on, which is detrimental

to optimizing this is that cross-attention is computed across sub-encoders, CLS pooling14. Moreover, several representative baselines and according to [49], the following possibility exists: the demonstrate the phenomenon of attention sinking, which we current encoder's attention weights may be more concerned hypothesize may be related to BERT's pre-training. Starting with the syntactic features local to the sequence, whereas the from the perspective of boosting the attention score may external encoder's Value may contain richer global semantic disturb BERT's pre-training information. We can start from features, and combines the two types of information when a different perspective by boosting the energy contained in it is propagated forward to the next EncoderLayer of the the CLS token in a way to enhance its degree of being current encoder, thus enhancing the the information density paid attention to, which in turn enriches the global semantic of the CLS. In addition, this cross-attention design allows the information aggregated by CLS pooling. model to dynamically adjust the information sources without After Query and Key compute the attention weight matrix, changing the original attention distribution, which preserves we note that the Value is weighted, and the weighted Value is theattentionpatternofthecurrentencoderwithoutdestroying defined as the context tensor. Naturally, we define the E the pre-training information inside the current encoder, and CLS metric as Eq. 6. Intuitively, in Eq. 6, the larger the ■h ■ , introducesnewsemanticcomplementsthroughtheValueofthe cls 2 the richer the CLS token aggregates semantic information, externalencoder.Inparticular,whentheCLStokenisrequired and the better CLS pooling effectiveness should be, which toserveasaglobalrepresentationoftheentireinputsequence, we demonstrate by exploring the relationship between the theValueoftheexternalencodercanprovideahigherlevelof performanceofsomerepresentativemodelsonthe7STStasks
semanticsupport,makingitshiddenstatemorecomprehensive and the E , as shown in Fig. 10, which reports an almost and robust. CLS positive correlation between the model's performance on the Weclarifytheproposedcross-attentiondesignedtoenhance 7 STS tasks and the E CLS. E CLS and enrich the semantic information of CLS pooled To enhance the average CLS energy weight, we introduce aggregation by the above justifications. a cross-attention structure within the twin encoder inspired by multimodal information fusion. Cross-attention in twin VII. CONCLUSION encoders is a mechanism for information interaction between In this work, we introduce the unsupervised sentence em- two different tensors, compared to traditional self-attention, bedding representation framework JTCSE. In JTCSE, we first which interacts with information based only on its own input propose the training objective of tensor modulus constraints sequence and may miss some important global information to improve the alignment between positive samples in un- andleadtoimperfectCLSpooling.Incontrast,cross-attention supervised contrastive learning. Then, we introduce a cross- can achieve cross-encoder information sharing by using the attention mechanism to optimize the quality of CLS Pooling attention weights of one encoder to weigh the Value tensor of to strengthen the model's attention to CLS tokens. Through another encoder. extensive evaluations, the results show that JTCSE is the current SOTA method for seven semantic textual similarity 14SincethemaskedlanguagemodeldesignedbyBERTinthepre-training computation tasks and outperforms other models on hundreds task does not mask the sequence's CLS token, the hidden state of the CLS of zero-shot evaluation tasks for natural language processing. tokenisconsideredtoaggregateallthesequence'ssemanticinformation.We
reporttheimpactofthepoolingapproachintheTableXI In addition, we analyze the effects of important components JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 17 in JTCSE through a series of ablation experiments. In future [20] Z. Yu, Z. Wang, Y. Fu, H. Shi, K. Shaikh, and Y. C. Lin, work, we will consider generalizing tensor mode length con- "Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration," straints and cross-attention mechanisms to multimodal learn- in Proc. ICML, Vienna, Austria, Jul. 2024. [Online]. Available: ing tasks. [21] H.Lin,X.Cheng,X.Wu,andD.Shen,"CAT:Crossattentioninvision REFERENCES transformer,"inProc.IEEEICME,Taipei,Taiwan,Jul.2022,pp.1–6. [22] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task- [1] T.ZongandL.Zhang,"Anensembledistillationframeworkforsentence agnostic visiolinguistic representations for vision-and-language tasks," embeddings with multilingual round-trip translation," in Proc. AAAI, AdvancesinNeuralInformationProcessingSystems,vol.32,2019. vol.37,no.11,2023,pp.14074–14082. [23] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, [2] J.Devlin,M.-W.Chang,K.Lee,andK.Toutanova,"Bert:Pre-training "VisualBERT: A simple and performant baseline for vision and ofdeepbidirectionaltransformersforlanguageunderstanding,"inProc. language,"arXivpreprintarXiv:1908.03557,2019.[Online].Available:
NAACL-HLT,Minneapolis,MN,USA,Jun.2019,pp.4171–4186. [3] Y. Liu, M. Ott, N. Goyal, J. Du, M.

Joshi, D. Chen, O. Levy, [24] X. Xu, C. Wu, S. Rosenman, V. Lal, W. Che, and N. Duan, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly “BridgeTower: Building bridges between encoders in vision-language optimizedbertpretrainingapproach,”arXivpreprintarXiv:1907.11692, representation learning,” in Proc. AAAI, Washington, DC, USA, Feb. 2019.[Online].Available: 2023,pp.10637–10647. [4] N.ReimersandI.Gurevych,“Sentence-bert:Sentenceembeddingsusing [25] X. Xu, B. Li, C. Wu, S.-Y. Tseng, A. Bhiwandiwalla, S. Rosenman, siamesebert-networks,”inProc.EMNLP-IJCNLP,HongKong,China, V.Lal,W.Che,andN.Duan,“ManagerTower:Aggregatingtheinsights Nov.2019,pp.3982–3992. of uni-modal experts for vision-language representation learning,” in [5] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, “Consert: Proc.ACL,Toronto,Canada,Jul.2023,pp.14507–14525. A contrastive framework for self-supervised sentence representation [26] E. Agirre, D. Cer, M. Diab, and A. Gonzalez-Agirre, “Semeval-2012 transfer,” in Proc. ACL/IJCNLP, Virtual Event, Aug. 2021, pp. 5065– task 6: A pilot on semantic textual similarity,” in Proc. SEM, 5075. Montre´al, Canada, Jun. 2012, pp. 385–393. [Online]. Available: [6] T. Gao, X. Yao, and D. Chen, “Simcse: Simple contrastive learning of sentence embeddings,” in Proc. EMNLP, Punta Cana, Dominican [27] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, Republic,Nov.2021,pp.6894–6910. “*sem 2013 shared task: Semantic textual similarity,” in Proc. [7] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning SEM, Atlanta, GA, USA, Jun. 2013, pp. 32–43. [Online]. Available: with contrastive predictive coding,” CoRR, vol. abs/1807.03748, 2018. [Online].Available: [28] E.Agirre,C.Banea,C.Cardie,D.Cer,M.Diab,A.Gonzalez-Agirre, [8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “Simclr: A simple W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe, “Semeval-2014 task framework for contrastive learning of visual representations,” in Proc. 10:Multilingualsemantictextualsimilarity,”inProc.SemEval,Dublin, ICLR,VirtualEvent,May2020. Ireland,Aug.2014,pp.81–91. [9] X. Wu, C. Gao, L. Zang, J. Han, Z. Wang, and S. Hu, [29] E.Agirre,C.Banea,C.Cardie,D.Cer,M.Diab,A.Gonzalez-Agirre, “Esimcse: Enhanced sample building method for contrastive learning W.Guo,I.Lopez-Gazpio,M.Maritxalar,R.Mihalcea,G.Rigau,L.Uria, of unsupervised sentence embedding,” in Proc. COLING, Gyeongju, andJ.Wiebe,“Semeval-2015task2:Semantictextualsimilarity,english, Republic of Korea, Oct. 2022, pp. 3898–3907. [Online]. Available: spanish and pilot on interpretability,” in Proc. SemEval, Denver, CO, USA,Jun.2015,pp.252–263. [10] K.He,H.Fan,Y.Wu,S.Xie,andR.B.Girshick,“Momentumcontrast [30] E.Agirre,C.Banea,D.Cer,M.Diab,A.Gonzalez-Agirre,R.Mihalcea, for unsupervised visual representation learning,” in Proc. IEEE/CVF G. Rigau, and J. Wiebe, “Semeval-2016 task 1: Semantic textual Conf.Comput.Vis.PatternRecognit.(CVPR),Seattle,WA,USA,Jun. similarity,monolingualandcross-lingualevaluation,”inProc.SemEval, 2020,pp.9726–9735. SanDiego,CA,USA,Jun.2016,pp.497–511. [11] Y.-S.Chuang,R.Dangovski,H.Luo,Y.Zhang,S.Chang,M.Soljacic, [31] D.Cer,M.Diab,E.Agirre,I.Lopez-Gazpio,andL.Specia,“Semeval- S.-W. Li, S. Yih, Y. Kim, and J. R. Glass, “Diffcse: Difference-based 2017 task 1: Semantic textual similarity multilingual and crosslingual contrastivelearningforsentenceembeddings,”inProc.NAACL,Seattle, focusedevaluation,”inProc.SemEval,Vancouver,Canada,Aug.2017, WA,USA,Jul.2022,pp.4207–4218. pp.1–14. [12] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “ELECTRA: pre-training text encoders as discriminators rather than generators,” in [32] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and Proc. ICLR, Addis Ababa, Ethiopia, Apr. 2020. [Online]. Available: R. Zamparelli, “A sick cure for the evaluation of compositional distributional semantic models,” in Proc. LREC, Reykjavik, Iceland, [13] Y.Zhang,H.Zhu,Y.Wang,N.Xu,X.Li,andB.Zhao,“Acontrastive May2014,pp.216–223.[Online].Available: framework for learning sentence representations from pairwise and proceedings/lrec2014/summaries/363.html triple-wiseperspectiveinangularspace,”inProc.ACL,Dublin,Ireland, [33] T. Zong, B. Shi, H. Yi, and J. Xu, “TNCSE: Tensor norm constraints May2022,pp.4892–4903. for unsupervised contrastive learning of sentence embeddings,” in [14] X. Wu, C. Gao, Z. Lin, J. Han, Z. Wang, and S. Hu, “InfoCSE: Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Information-aggregated contrastive learning of sentence embeddings,” pp. 26192–26201. [Online]. Available: inFindingsoftheAssociationforComputationalLinguistics:EMNLP, v39i24.34816 AbuDhabi,UnitedArabEmirates,Dec.2022,pp.3060–3070. [34] J. W. Wei and K. Zou, “EDA: easy data augmentation techniques for [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, boosting performance on text classification tasks,” in Proc. EMNLP- A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is IJCNLP,HongKong,China,Nov.2019,pp.6381–6387. all you need,” in Advances in Neural

Information Processing [35] H. Tan and M. Bansal, “LXMERT: learning cross-modality encoder Systems (NeurIPS), Long Beach, CA, USA, Dec. 2017, pp. 5998– representations from transformers,” in Proc. EMNLP-IJCNLP, Hong 6008. [Online]. Available: Kong,China,Nov.2019,pp.5099–5110. hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html [36] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. [16] H. Wang and Y. Dou, “SNCSE: Contrastive learning for unsuper- Le, Y.-H. Sung, Z. Li, and T. Duerig, “Scaling up visual and vised sentence embedding with soft negative samples,” in Proc. ICIC, vision-language representation learning with noisy text supervision,” Zhengzhou,China,Aug.2023,pp.419–431. in Proc. ICML, Virtual Event, Jul. 2021, pp. 4904–4916. [Online]. [17] T. Jiang, J. Jiao, S. Huang, Z. Zhang, D. Wang, F. Zhuang, F. Wei, Available: H. Huang, D. Deng, and Q. Zhang, “PromptBERT: Improving bert [37] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, sentence embeddings with prompt-based fine-tuning,” in Proc. ACL, S.Agarwal,G.Sastry,A.Askell,P.Mishkin,J.Clark,G.Krueger,and Dublin,Ireland,May2022,pp.1234–1245. I.Sutskever,“Learningtransferablevisualmodelsfromnaturallanguage [18] F. Schwenker, “Ensemble methods: Foundations and algorithms [book supervision,” in Proc. ICML, Virtual Event, Jul. 2021, pp. 8748–8763. review],”IEEEComput.Intell.Mag.,vol.8,no.1,pp.77–79,2013. [Online].Available: [19] J.Liu,J.Liu,Q.Wang,J.Wang,W.Wu,Y.Xian,D.Zhao,K.Chen, [38] Y. Zhang, R. He, Z. Liu, K. H. Lim, and L. Bing, “An unsupervised and R. Yan, “Rankcse: Unsupervised sentence representations learning sentence embedding method by mutual information maximization,” in vialearningtorank,”inProc.ACL,2023,pp.13785–13802. Proc.EMNLP,Online,Nov.2020,pp.1601–1610. JOURNALOFLATEXCLASSFILES,VOL.14,NO.8,AUGUST2021 18 [39] Q. Wu, C. Tao, T. Shen, C. Xu, X. Geng, and D. Jiang, “PCL: Bingkang Shi received his Bachelor’s degree in Peer-contrastive learning with diverse augmentations for unsupervised ElectronicandInformationEngineeringfromXidian sentence embeddings,” in Proc. EMNLP, Abu Dhabi, United Arab University(XDU)in2019.Heiscurrentlypursuing Emirates,Dec.2022,pp.12052–12066. his PhD at the University of Chinese Academy of [40] W. Zhuo, Y. Sun, and X. Wang, “When does contrastive learning Sciences.HisresearchfocusesonNaturalLanguage preserve semantic similarity? a case study on sentence embeddings,” Processing(NLP),particularlymodelbiasandnatu- in Proc. EMNLP, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. rallanguagereasoning.Hisworkhasbeenpublished 3021–3035. at ICASSP 2024 and EMNLP 2023. His honors [41] Y. Jiang, L. Zhang, and W. Wang, “Improved universal sentence include an Honorable Mention in the MCM/ICM embeddings with prompt-based contrastive learning and energy-based competition and first prize in Xidian University’s learning,” in Proc. EMNLP, Abu Dhabi, United Arab Emirates, Dec. SparkCupScienceandTechnologyCompetition. 2022,pp.3021–3035. [42] J.Su,J.Cao,W.Liu,andY.Ou,“Whiteningsentencerepresentations for better semantics and faster retrieval,” CoRR, vol. abs/2103.15316, 2021.[Online].Available: [43] H.He,J.Zhang,Z.Lan,andY.Zhang,“Instancesmoothedcontrastive Yuanxiang Wang received the Bachelor’s degree learning for unsupervised sentence embedding,” in Proc. AAAI, Wash- fromtheUniversityofChineseAcademyofSciences ington,DC,USA,Feb.2023,pp.12863–12871. in2023.Heiscurrentlyworkingasaninternatthe [44] A. Conneau and D. Kiela, “Senteval: An evaluation toolkit for Cloud Computing and Intelligent Processing Lab- universal sentence representations,” in Proc. LREC, Miyazaki, Japan, oratory(CCIPLab)ofUCAS.Hisresearchinterests May 2018. [Online]. Available: arefocusedonthefieldofmultimodallargemodels, lrec2018/summaries/757.html particularly on the fine-tuning and evaluation of [45] D. M. Cer, M. T. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, thesemodels. “SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation,” in Proc. 11th Int. Workshop SemanticEvaluation(SemEval@ACL). AssociationforComputational Linguistics, 2017, pp. 1–14. [Online]. Available: 18653/v1/S17-2001 [46] X.Chen,A.Zeynali,C.Camargo,F.Flo¨ck,D.Gaffney,P.Grabowicz, Jungang Xu Jungang Xu is a full professor S.Hale,D.Jurgens,andM.Samory,“SemEval-2022task8:Multilingual in School of Computer Science and Technology, newsarticlesimilarity,”inProc.SemEval,Seattle,WA,USA,Jul.2022, University of Chinese Academy of Sciences. He pp.1094–1106. received his Ph.D. degree in Computer Applied [47] N.Muennighoff,N.Tazi,L.Magne,andN.Reimers,“MTEB:Massive Technology from Graduate University of Chinese text embedding benchmark,” in Proc. EACL, Dubrovnik, Croatia, May AcademyofSciencesin2003.Hiscurrentresearch 2023,pp.2006–2029. interests include multi-modal intelligence, natural [48] T. Wang and P. Isola, “Understanding contrastive representation languageprocessingandembodiedintelligence.He learning through alignment and uniformity on the hypersphere,” in haspublishedmorethan30papersinIEEEjournals Proc. ICML, Virtual Event, Jul. 2020, pp. 9929–9939. [Online]. andconferences. Available: [49] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, “What does BERTlookat?ananalysisofBERT’sattention,”inProc.ACLWorkshop

BlackboxNLP,Florence,Italy,Aug.2019,pp.276–286. Tianyu Zong received his Bachelor's degree from the North China University of Technology in 2020 and his Master's in Electronic Information from the University of Chinese Academy of Sciences in 2023. He is pursuing a Ph.D. at the School of Computer Science and Technology, University of ChineseAcademyofSciences.Hisresearchfocuses on natural language processing and multimodal in- formationfusion.Hehaspublishedtwofirst-author papersattheAAAIconferenceandhasbeeninvited togiveanoralpresentationatAAAI2025. Hongzhu Yi receivedhisBachelor'sdegreeinAu- tomation from Xidian University in 2023. He is currently pursuing a Ph.D degree at the Chinese AcademyofSciences.Hisresearchfocusesonlarge models, with a particular emphasis on the under- standingandgenerationofmultimodallargemodels. Heactivelyparticipatesinvariousinternationalcom- petitions,includingICASSPMEIJU2024,NeurIPS Edge LLMs Challenge 2024, CVPR Ego4D 2024, andCVPREgo4D2025.