Abstract

—Unsupervisedcontrastivelearninghasbecomeahot Add&Output Tensor Modulus Constraint research topic in natural language processing. Existing works Add&Output InfoNCE usually aim at constraining the orientation distribution of the Mutual and Self-supervised representations of positive and negative samples in the high- InfoNCE dimensional semantic space in contrastive learning, but the semanticrepresentationtensorpossessesbothmodulusandorien- ...... Q,K V V K,Q 1 2 n tationfeatures,andtheexistingworksignorethemodulusfeature 1 2 oftherepresentationsandcauseinsufficientcontrastivelearning. Q,K V V K,Q Therefore, we first propose a training objective that is designed a. b. to impose modulus constraints on the semantic representation Inputed Inputed tensor, to strengthen the alignment between positive samples in contrastivelearning.Then,theBERT-likemodelsuffersfromthe Feed Forward 1 Pretrained Encoder phenomenon of sinking attention, leading to a lack of attention Feed Forward of Cross-Attention hidden States toCLStokensthataggregatesemanticinformation.Inresponse, we propose a cross-attention structure among the twin-tower Fig. 1. Subfigure a. represents the traditional ensemble modeling approach ensemble models to enhance the model's attention to CLS token (EDFSE[1]),whichnaivelytrainsmultiplesub-encodersseparatelyandthen and optimize the quality of CLS Pooling. Combining the above sumstheoutputs.Thisapproachcausesalargeinferenceoverhead.Subfigure two motivations, we propose a new Joint Tensor representation b.representstheoptimizedensemblelearningframeworkJTCSEproposedin modulusconstraintandCross-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 [2] and RoBERTa [3], much work has been done 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: https://github.com/tianyuzong/JTCSE. Dropout method of the BERT-like model at training time and Index Terms—Unsupervised Contrastive Learning, Semantic uniformly distributing unlabeled soft-negative samples. With TextualSimilarity,Tensor-ModulusConstraints,Cross-Attention. theappearanceofSimCSE,manyworksarebasedontheidea ofunsupervisedcontrastivelearninginSimCSEandInfoNCE. For example, ESimCSE [9] augments the positive sample in I.