**Skip-Thought Vectors**

**Ryan Kiros 1, Yukun Zhu 1, Ruslan Salakhutdinov 1,2, Richard S. Zemel 1,2,Antonio Torralba 3, Raquel Urtasun 1, Sanja Fidler 1** University of Toronto 1 Canadian Institute for Advanced Research 2 Massachusetts Institute of Technology 3

**Abstract**

We describe an approach for unsupervised learning of a generic, distributed sentence encoder. Using the continuity of text from books, we train an encoder- decoder model that tries to reconstruct the surrounding sentences of an encoded passage. Sentences that share semantic and syntactic properties are thus mapped to similar vector representations. We next introduce a simple vocabulary expansion method to encode words that were not seen as part of training, allowing us to expand our vocabulary to a million words. After training our model, we extract and evaluate our vectors with linear models on 8 tasks: semantic relatedness, paraphrase detection, image-sentence ranking, question-type classification and 4 benchmark sentiment and subjectivity datasets. The end result is an off-the-shelf encoder that can produce highly generic sentence representations that are robust and perform well in practice.

**摘要**

本文描述了一种无监督学习方法，即一种通用的分布式句子编码器。利用书本中文本的连续性, 我们训练了一个编码-解码器模型, 试图重建编码段落周围的句子。因此, 可将共享语义和句法属性的句子被映射到类似的向量表示。接下来, 本文将介绍一个简单的词汇膨胀方法, 用于对未被视为训练内容的单词进行编码, 从而增加我们的词汇量到100万个单词。在模型训练结束后, 我们用线性模型对我们的向量在8项任务上进行提取和评估: 语义相关性、意译检测、图像-句子排序、问题类型分类，4基准情绪和主观性数据集。最终得到了一个可以产生高度通用的句子表示编码器,非常稳健, 并在实践中表现良好。