视觉增强词向量：我是词向量，我开眼了！

原创 橙橙子 夕小瑶的卖萌屋 2021-06-17 12:05

文 | 橙橙子

亲爱的读者，你是否被各种千亿、万亿模型的发布狂轰乱炸，应接不暇，甚至有点产生对大模型的审美疲劳？出于这个目的，今天来分享一篇研究静态词向量的小清新文章。希望大家可以在理性追热的同时，小冶情操。并且能够发现内在共性，有所启示。

论文标题：

Learning Zero-Shot Multifaceted Visually Grounded Word Embeddings via Multi-Task Training

论文链接：

https://arxiv.org/pdf/2104.07500.pdf

词向量为什么要进行视觉增强

词是自然语言表达语义的基本单元，从静态词向量word2vec[1],GloVe[2]到动态词向量ELMo[3],BERT[4]，词向量的演变进化之路就是深度学习在NLP辉煌发展历程的灵感源泉之一。在现有词向量技术的分布式假设中，有一个非常重要的概念就是“**文本上下文(Context)**”，即在相似的文本上下文中出现的词在语义表示空间中会更相似。这个理念非常成功，但是也有缺陷，它直接导致了词向量的学习过分依赖于词汇的共现关系（co-occurrences），缺乏更广泛的、来源于真实世界的知识背景。一个经典的例子是Good和Bad，与它们共现的上下文词汇经常是相似的，物理含义却截然不同。

康德曾强调过类比在科学认识活动中的重要作用，尤其是在仿生设计上。模拟和类比人类启发了神经网络、深度学习，看起来也是人工智能否通过图灵测试的关键。我们知道人类在理解词的基本概念的时候，会不由自主的和现实世界建立关联，所谓在阅读和交谈时身临其境、浮想联翩说的都是这种神奇的

能力。自然语言处理中也有一种类似的技术叫做**Grounding**，它甚至有个更高大上的名字叫**Grounded Natural Language Processing (GNLP)**，研究目的是将自然语言和外部物理世界的丰富的感知连接在一起，从而解决各种多模态问题以及反过来加强自然语言理解能力。这种感知可以是视觉信号、声音信号、运动信号等等，所以和计算机视觉、机器人技术、图形学等学科都密不可分。“Ground（ed,ing）”这个词不是很好翻译成中文，我们可能最容易联想到的就是Ground Truth（此处应该有类比）。

既然人类很擅长将视觉和语言建立关联（**Visual-Lauguage Grounding**），从而更好的理解语言。模型也可以**借助视觉信息得到更好的词向量**么？

多任务视觉Grounding

对于任意词w, 已经在文本数据上预训练好的词向量是$T_e(w) \in \mathbb{R}^d$，譬如word2vec，GloVe等。我们的目的是学习一个映射矩阵$M_{d\times c}$，将$T_e(w) \in \mathbb{R}^d$ Ground 到对应视觉强化的语义空间上，获得的Grounded词向量记作$G_e(w) \in \mathbb{R}^c$。

为了达成这一目的，论文设计了三个部分：

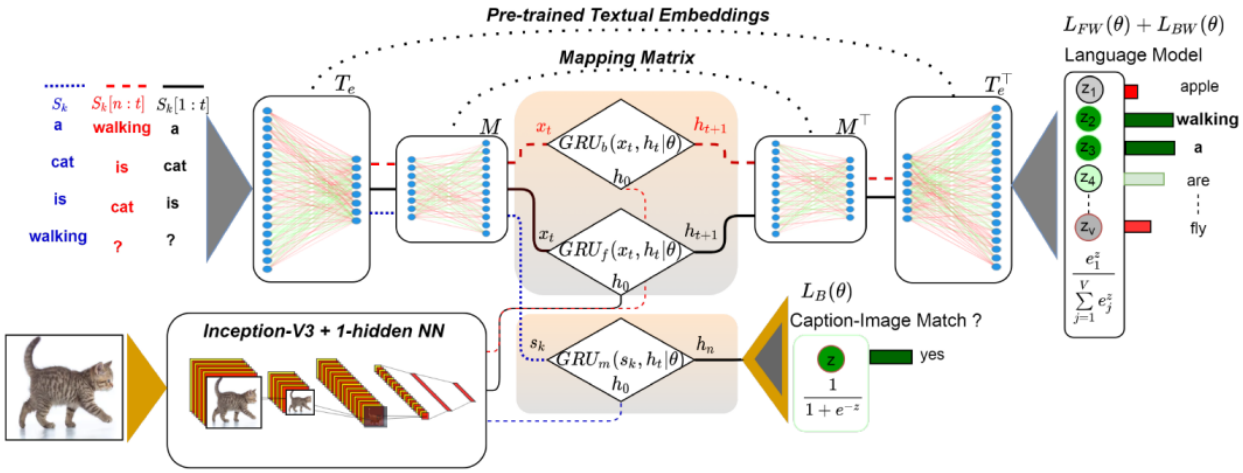

Figure 1: Our zero-shot multi-task learning approach includes: 1. Two GRU based language-model tasks in forward (GRU_f) and backward (GRU_b) directions represented by dashed red and solid black lines in the upper block. 2. A matching task predicting if the given sentence describes the image (blue dotted line, lower block). The zero-shot mapping matrix M shared by all the tasks, learns to visually ground the textual word vectors by learning a reversible mapping from textual space to grounded space.

语言模型

设图文描述数据集为$(S_k, I_k) \in D$，其中$S_k = [w_1, w_2 \dots w_n]$对应文本部分，$I_k$对应图像部分。我们使用$T_e$ 获得对应的词向量表示$S_t = [t_1, t_2 \dots t_n]$，我们接着学习一个映射矩阵$M_{d\times c}$，将这些表示 Ground 到对应的视觉强化的语义空间上。

$$G_e(S_k) = S_{t_{n\times d}} \times M_{d\times c}$$

获得的Grounded词向量记作$G_e(S_k) = [x_1, x_2 \dots x_n]$，其中$x_i \in \mathbb{R}^c$。为了达到这个目的，该文本对应的图像视觉信息融合到了语言模型的学习过程中。论文使用了GRU，这里比较巧妙地将视觉信息在线性

映射后初始化第一个hidden state h_0，相当于在语言模型的学习前有一个全局的视觉背景，我们希望GRU的门控机制可以学习到外部的视觉知识如何传播到映射矩阵M中。

同时，映射矩阵的转置也被用于进行逆向操作，即从 Grounded 空间映射回纯文本空间：

$$w_{next} = h_t \times M^\top$$

最终，前向语言模型基于图像和之前的词来逐个生成下一个词。其中，V代表词汇表的大小，B代表batch size大小，$\hat{y}_{i,c}$ 和 $y_{i,c}$分别表示预测概率和Ground Truth：

$$\mathcal{L}_{FW}(\theta) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{V} y_{i,c} log(\hat{y}_{i,c})$$

另外，论文增加了新的一个反向的GRU来加强学习能力，实现的时候将词序列逆序进行GRU建模。除两个GRU的参数不同外，其余参数都是共享的。这个设计类似于双向GRU，但是后者用在语言模型中会有会有标签泄漏的问题，所以论文这里使用了两个GRU来代替。

图文匹配

尽管基于上下文的词表示方法是获得高质量的词向量的有效途径，但是从目标设计的角度，却不见得能同时给多种视觉-语言任务（visual-language task）都带来增益，譬如图文检索任务需要模型具备两种模态的强相关性建模。所以本文也增加了一个图文匹配判定任务，试图让Grounded Embedding进一步增强图像和文字相关性能力。

虽然这里模型选的简单也很符合直觉，但是想法其实和多模预训练里使用对比学习对齐视觉和语言表示空间是类似的。这里使用了第三个GRU，同样用视觉表示来初始化h_0，这里用最后的hidden state $h_n = GRU_m(G_e(S_k), h_0|\theta)$来建模整体，负样例随机采样，优化二元交叉熵：

$$\mathcal{L}_B(\theta) = -\frac{1}{B}\sum_{i=1}^{B} y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)$$

其中，$\hat{y}$ 和 y分别表示预测概率和Ground Truth

正则化

以上任务均共享预训练好的文本词向量T_e，一个容易想到的问题是，它究竟要不要finetune呢？如果要进行更新，它们可能会极大的偏离原始向量，扰乱预先训练好的语义关系，特别是在有限的训练语料的情况下。如果完全不进行更新，由于这些词向量本身有偏，可能会很难映射到Grouded Embedding上去。为了兼容这两种情况，论文这里对T_e的学习进行了正则约束：

$$\mathcal{R}(\alpha, \beta) = \frac{\alpha}{|V|}\sum_{w \in V} |\beta - \frac{w_f \cdot w_u}{||w_f|| \times ||w_u||}|$$

其中，α控制了正则约束整体的影响，β控制调整后的词向量和最初的词向量被允许的差异程度。

最终，模型优化的是多个任务：

$$\mathcal{L}_{All}(\Theta) = \mathcal{L}_{FW}(\theta) + \mathcal{L}_{BW}(\theta) + \mathcal{L}_B(\theta) + \mathcal{R}(\alpha, \beta)$$

实验

实验训练图文训练数据选择了MS-COCO，图像的视觉信息使用训练好的Inception-V3加一层tanh非线性层来提取。预训练好的文本词向量T_e则选择使用了经典的GloVe[2] ($crawl-300d-2.2M$) 和fastText[5] ($crawl-300d-2M$)，词表大小设置为10k。

由于已经学到了文本空间向Grounded空间的映射矩阵M，对于一些不在image-text训练语料中的未登录词（Oov），也可以采取这样的映射获得对应的Grounded空间，从而获得zero-shot的能力，也是论文的卖点之一。这里设原始文本词向量为GloVe和fastText，视觉增强后的Grounded 词向量为V_GloVe和V_fastText。

如何评估词向量的好坏至今也是一个开放性问题，论文选择了intrinsic（内在评价）和 extrinsic（外在评价）两种评估方法。内在评价度量的是词向量本身的质量，忽略了它的下游任务表现。外在评价度量的是词向量在句子粒度的下游任务上的表现。

内在评估

内在评估在多种词汇相似度评估基准集合（Benchmark）上进行了测试。基线对比上，作者选择了纯文本训练的词向量和一些其他的Grounded 词向量模型。可以发现V_GloVe和V_fastText在各个benchmark上相对于纯文本预训练词向量GloVe和fastText均获得了稳定的效果提升，Spearman系数平均+6.5和+1.6。另外，实验也揭示了一些有趣的现象，SimLex999主要关注词向量之间的语义相似度，WSim353主要关注于相关性。V_Word Embedding看起来在语义相似度度量上提升的更多。

Model	RW	MEN	WSim 353	MTurk 771	SimVerb 3500	SimLex 999	Mean
GloVe	45.5	80.5	73.8	71.5	28.3	40.8	56.7
V_GloVe	52.6	**85.1**	**78.9**	73.4	37.4	51.8	**63.2**
fastText	56.1	81.5	72.2	75.1	37.8	47.1	61.6
V_fastText	**57.8**	83.6	73.9	**76.1**	**39.2**	49.0	**63.2**
Cap2Both	48.7	81.9	71.2	_	_	46.7	_
Cap2Img	52.3	84.5	75.3	_	_	51.5	_
Park et al.	_	83.8	77.5	_	_	**58.0**	_
Collell et al.	_	81.3	_	_	28.6	41.0	_

Table 1: Intrinsic evaluation. Visual grounding (denoted by 'V') improves results compared to pre-trained fastText and GloVe on all test sets.

细粒度内在评估

为了进一步研究Grounded Embedding的贡献，论文在SimLex999的多个类别数据下进行了实验，分为形容词、名词、动词，以及词的具像程度。譬如Apple（苹果）这个词是一个实体词，非常具像。而

Pressure（压力）这个词比较抽象，Conc-qx的分位数越高，代表词越具体。论文这里对比了Google hinton组在早年发的一篇Grounded 词向量的模型PictureBook[6]的结果，这个工作利用了大量图文搜索引擎日志数据来训练Grounded词向量。V_GloVe的表现并不落下风。我们可以看到之前的方法对于具体词的Grounding能力是做的比较好的，这也符合直觉，因为图文训练语料大多数都是在描述一个客观的实体。而V_Glove在抽象词的表现上要好于PictureBook，很大程度上归功于Grounding映射矩阵在zero-shot上的设计。

Model	All	Adjs	Nouns	Verbs	Conc-q1	Conc-q2	Conc-q3	Conc-q4	Hard
GloVe	40.8	62.2	42.8	19.6	43.3	41.6	42.3	40.2	27.2
V_GloVe	**51.8**	**72.1**	52.0	**35**	**53.1**	**54.8**	47.4	56.8	**38.3**
Picturebook	37.3	11.7	48.2	17.3	14.4	27.5	46.2	60.7	28.8
Picturebook+GloVe	45.5	46.2	**52.1**	22.8	36.7	41.7	**50.4**	**57.3**	32.5

Table 2: SimLex999 (Spearman's ρ) results. The best result in each category is bolded. Conc-q1 and Conc-q4 contain the most abstract and concrete words respectively. Our embeddings (V_GloVe) generalize across different word types and strongly outperform all the others on most of the categories.

外在评估

外在评估是在数年的SentEval数据集上进行测试，这种评估的优势在于不需要训练数据，而是直接把词向量进行累加平均后得到句子表示，最大程度的评估词向量空间的内在结构，并且能够发现其中存在的不规律性。我们看到V_Word Embedding大幅提升了效果，Spearman系数平均+10.0。

Model	STS12	STS13	STS14	STS15	STS16	Mean
GloVe	53.22	54.14	55.41	60.08	51.43	54.85
V_Glove	**56.17**	**62.42**	**66.57**	**69.91**	**65.56**	**64.13**
Fasttext	21.02	30.36	29.54	39.21	30.10	30.05
V_Fasttext	**31.00**	**38.63**	**42.12**	**51.74**	**38.71**	**40.44**

Table 3: Spearman correlation results for semantic similarly benchmarks. Both grounded embeddings strongly outperform their textual versions on all the tasks.

进一步分析

论文接着展示了多组词向量的最近邻结果。进一步表明Grounded 词向量可以优化纯文本向量空间，从而对齐到真实物理世界的概念上。譬如我们看bird（鸟）这个词，GloVe展示的最近邻词是turtle（乌龟）、nest（鸟巢）和squirrel（松鼠）。而V_Glove的最近邻是sparrow（麻雀）、Birds（鸟），avian（鸟类）。另一个例子是抽象程度更高的词happy（高兴），我们可以看到由于纯文本预训练词向量存在强的词共现关系的假设，会得到一些无价值的词汇，譬如everyone（所有人），always（总是）。而V_Glove得到的词更符合人类的认知：pleased（高兴），delighted（高兴）。

happy		sad		big		bird		horse		together		smart	
G	V	G	V	G	V	G	V	G	V	G	V	G	V
lucky	pleased	sadly	saddened	hard	humongous	turtle	sparrow	dog	racehorse	well	togeather	sensible	witty
everyone	delighted	shame	tragic	little	Big	nest	Birds	riding	Thoroughbred	bring	togheter	dumb	shrewd

love	merry	horrible	mournful	squirrel	avian	ponies	Horses	both	toegther	sophisticated	inteligent
always	thrilled	scared	saddening			donkey	steed	they	togather	attractive	resourceful
wish	joyful	awful	sorrowful					apart	togethor	wise	quick-witted
hope	hapy	pity	Sad					up	2gether		
		kinda	heartbreaking					them	togehter		
		sorry	heartbroken					put	togther		
								along	toghether		
								with	gether		

Table 4: Results of 10 nearest neighbors for Glove (G) and V_Glove (V). Only the differing neighbors are reported. While GloVe retrieves more related words, ours(V_Glove) focuses on similar words. Overall, V_Glove is closer to human judgment and retrieves highly semantically similar words.

结论

论文提出了一种使用视觉Grounding来增强词向量表示能力的方法。麻雀虽小，五脏俱全。论文在模型设计中使用了视觉-文本联合上下文取代纯文本上下文来进行语言模型训练，同时具备一定的zero-shot能力，其背后阐释的思想和目前火热的多模态大模型是类似的，希望对大家有所启发。

萌屋作者：橙橙子

拿过Kaggle金，水过ACM银，发过顶会Paper，捧得过多个竞赛冠军。梦想是和欣欣子存钱开店，沉迷于美食追剧和炼丹，游走于前端后端与算法，竟还有一颗想做PM的心！

作品推荐

1. **惊呆！不用一张图片，却训出个图像识别SOTA？**

参考文献

[1] Efficient Estimation of Word Representations in Vector Space https://arxiv.org/abs/1301.3781

[2] GloVe: Global Vectors for Word Representation https://www.aclweb.org/anthology/D14-1162/

[3] Deep contextualized word representations https://arxiv.org/abs/1802.05365

[4] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/abs/1810.04805

[5] Enriching Word Vectors with Subword Information https://arxiv.org/abs/1607.04606

[6] Illustrative Language Understanding:Large-Scale Visual Grounding with Image Search https://www.cs.toronto.edu/~hinton/absps/picturebook.pdf