《Learning Visual N-Granms from web data》

目录

一、论文理解

文章思想

利用弱标签的web数据，建立language与image的Visual n-gram model.能够预测与图像内容相关的phrases，这篇文章的主要贡献在于损失函数，其来源为nlp里面的n-gram模型。

关于loss函数：given an image I, assign a likelihood p(w|I) to each possible phrase (n-gram) w。we develop a novel, differentiable `loss function` that optimizes trainable parameters for `frequent n-grams`, whereas for `infrequent n-grams`, the loss is dominated by the predicted likelihood of smaller `“sub-grams”`.

相关工作

- learning from weakly supervised web data

  本文采用了和[28]一样的弱监督训练数据，但是不同于它的是不仅仅只考虑单个词，而是考虑了n-gram。数据来自图像分享网站：image-comment。

- Relating image content and language

没有采用RNN，而是采用了bilinear model，它也能根据给定图像输出phrases的概率，并把相关的phrases组合成caption。而与其他类似文章所不同的是：本文能处理大量的visual concepts，而不仅仅限制于flickr类似的数据集上面的评论内容，更加能用于实际问题。此处与[40]最相关，但是使用了一个端到端的弱监督训练方法。

- Language models

  使用了n-gram的语言模型，并使用了Jelinek Mercer smoothing[26] 。

  `n-gram models count the frequency of n-grams in a text corpus to produce a distribution over phrases or sentences, our model measures phrase likelihoods by evaluating inner products between image features and learned parameter vectors.`

数据集及预处理

- train: YFCC100M dataset

  comments:We applied a simple language detector to the dataset to select only images with English user comments, leaving a total of 30 million examples for training and testing. We preprocessed the text by removing punctuations, and we added [BEGIN] and [END] tokens at the beginning and end of each sentence.

  images:rescaling them to 256×256 pixels (using bicubic interpolation), cropping the central 224× 224, subtracting the mean pixel value of each image, and dividing by the standard deviation of the pixel values

- 1-5 grams

  the `smoothed visual n-gram models` are trained and evaluated on all n-grams in the dataset, even if these n-grams are not in the dictionary. However, whereas the probability of `indictionary n-grams` is primarily a function of parameters that are specifically tuned for those n-grams, the probability of `out-of-dictionary` n-grams is composed from the probability of smaller in-dictionary n-grams (details below).

损失函数函

- $\phi(I,\theta)$是CNN特征提取网络
- I是图像
- denote the n-gram dictionary that our model uses by D and a comment containing K words by $w \in [1,C]^K$, where C is the total number of words in the (English) language. We denote the n-gram that ends at the i-th

word of comment w by w_{i-n+1}^{i} and the ith word in comment w by w_i^i.omit the sum over all image-comment pairs in the training / test data when writing loss functions

- 预测的分布是一个 n-gram embedding matrix $E \in R^{D \times |D|}$

Naive n-gram loss

Naive n-gram loss. The naive n-gram loss is a standard multi-class logistic loss over all n-grams in the dictionary $\mathcal{D}$. The loss is summed over all n-grams that appear in the sentence w; that is, n-grams that do not appear in the dictionary are ignored:

$$\ell(\mathbf{I}, w; \theta, \mathbf{E}) = -\sum_{m=1}^{n}\sum_{i=n}^{K} \mathbb{I}\left[w_{i-m+1}^{i} \in \mathcal{D}\right] \log p_{obs}\left(w_{i-m+1}^{i} | \phi(\mathbf{I}; \theta); \mathbf{E}\right),$$

where the *observational likelihood* $p_{obs}(\cdot)$ is given by a softmax distribution over all in-dictionary n-grams w that is governed by the inner product between the image features $\phi(\mathbf{I}; \theta)$ and the n-gram embeddings:

$$p_{obs}\left(w | \phi(\mathbf{I}; \theta); \mathbf{E}\right) = \frac{\exp\left(-\mathbf{e}_w^\top \phi(\mathbf{I}; \theta)\right)}{\sum_{w' \in \mathcal{D}} \exp\left(-\mathbf{e}_{w'}^\top \phi(\mathbf{I}; \theta)\right)}.$$

The image features $\phi(\mathbf{I}; \theta)$ are produced by a convolutional network $\phi(\cdot)$, which we describe in more detail in 3.3.

由于此模型不是一个条件概率，不能进行语言模型的建模，因此加入一个back-off model[6]:

$$p(w_i^i | w_{i-n+1}^i) \propto \begin{cases} p_{obs}(w_i^i | w_{i-n+1}^i), if w_{i-n+1}^i \in D \\ \lambda p(w_i^i | w_{i-n+2}^i), otherwise \end{cases},$$

Jelinek-Mercer loss

Jelinek-Mercer (J-M) loss. The simple n-gram loss has two main disadvantages: (1) it ignores out-of-dictionary n-grams entirely during training and (2) the parameters $\mathbf{E}$ that correspond to infrequent in-dictionary words are difficult to pin down. Inspired by Jelinek-Mercer smoothing, we propose a loss function that aims to address both these issues:

$$\ell(\mathbf{I}, w; \theta, \mathbf{E}) = -\sum_{i=1}^{K} \log p\left(w_i^i | w_{i-n+1}^{i-1}, \phi(\mathbf{I}; \theta); \mathbf{E}\right),$$

where the likelihood of a word conditioned on the $(n-1)$ words appearing before it is defined as:

$$p\left(w_i^i | w_{i-n+1}^{i-1}\right) = \lambda p_{obs}\left(w_i^i | w_{i-n+1}^{i-1}\right) + (1-\lambda) p\left(w_i^i | w_{i-n+2}^{i-1}\right).$$

Herein, we removed the conditioning on $\phi(\mathbf{I}; \theta)$ and $\mathbf{E}$ for brevity. The parameter $0 \leq \lambda \leq 1$ is a smoothing

此处避免了naive 函数的两个弊端。此处提出的Jelinek-Mercer smoothing方法对于E和θ是可导的。因此loss就可以通过卷积网络回传。

训练

- CNN: residual network [23] with 34 layers
- follow [28] and perform stochastic gradient descent over outputs [4]: we only perform the forwardbackward pass for a random subset (formed by all positive n-grams in the batch) of the columns of E.(原因是全部更新的话，输出量太大)

实验

- phrase-level image tagging

  就是给定图像，输出图像内容相关的phrases（table2）；还对模型进行了复杂度的分析（table1）

Model	R@1	R@5	R@10	Accuracy
Imagenet + linear	5.0	10.7	14.5	32.7
Naive n-gram	5.5	11.6	15.1	36.4
Jelinek-Mercer	**6.2**	**13.0**	**18.1**	**42.0**

Table 2. Phrase-prediction performance on YFCC100M test set of $10,000$ images measured in terms of recall@k at three cut-off levels k (lefthand-side; see text for details) and the percentage of correctly predicted n-grams according to human raters (righthand-side) for one baseline model and two of our phrase prediction models. Higher is better.

- phrased based image retrieval

  实验表明，模型能很好的进行区分visual concepts：

  the model has learned accurate visual representations for n-grams such as “Market Street” and “street market”, as well as for “city park” and “Park City”,our model is able to distinguish visual concepts related to Washington: namely, between the state, the city, the baseball team, and the hockey team

- relating images and captions

  即给定图像检索相关phrases，注意此文用是数据具有much larger vocabularies than the baseline models

  以上两个实验任务自然是无法和专门做检索的工作效果好的。

- zero-shot transfer

  1)The performance of our models is particularly good on common classes such as those in the `aYahoo dataset` for which many examples are available in the YFCC100M dataset.

  2)The performance of our models is worse on datasets that involve fine-grained classification such as `Imagenet`, for instance, because YFCC100M contains few examples of specific, uncommon dog breeds