TensorBoard：嵌入可视化

原文链接：https://www.tensorflow.org/get_started/embedding_viz

译文链接：http://www.apache.wiki/pages/viewpage.action?pageId=10029493

贡献者：片刻 ApacheCN Apache中文网

嵌入在机器学习中普遍存在，出现在推荐系统，NLP和许多其他应用程序中。实际上，在TensorFlow的上下文中，将张量（或张量片）视为空间点是很自然的，因此几乎任何TensorFlow系统自然会产生各种嵌入。

TensorBoard有一个内置的可视化器，称为嵌入式投影仪，用于交互式可视化和分析高维数据（如嵌入）。嵌入式投影机将从您的模型检查点文件读取嵌入。虽然它对于嵌入最有用，但它会加载任何2D张量，包括您的训练权重。

要了解有关嵌入和如何训练它们的更多信息，请参阅“矢量代表词”教程。如果您对嵌入图像感兴趣，请查看 这篇文章，了解MNIST图像的有趣可视化。另一方面，如果你对文字嵌入感兴趣， 这篇文章 给出了很好的介绍。

默认情况下，嵌入投影机使用主成分分析将高维数据投影到 三维。有关PCA的视觉说明，请参阅 这篇文章。您可以使用的另一个非常有用的投影是 t-SNE。我们稍后在本教程中谈论更多的t-SNE。

如果您正在使用嵌入式工作，则可能需要将标签/图像附加到数据点。您可以通过生成包含每个点的标签的 元数据文件，并通过使用我们的Python API配置投影机，或手动构建和保存 `projector_config.pbtxt` 在与检查点文件相同的目录中来实现。

建立

有关如何运行TensorBoard的深入信息，并确保您记录了所有必要的信息，请参阅TensorBoard：可视化学习。

为了可视化您的嵌入，您需要做3件事情：

1）设置一个保存嵌入的2D张量。

```
embedding_var = tf.Variable(....)
```

2 LOG_DIR

```
saver = tf.train.Saver()
saver.save(session, os.path.join(LOG_DIR, "model.ckpt"), step)
```

3

如果你有你的嵌入相关的任何元数据（标签，图像），你可以告诉TensorBoard它通过直接存储`projector_config.pbtxt` 在`LOG_DIR`，或使用我们的API的Python。

例如，以下`projector_config.ptxt`将 `word_embedding`张量与存储在`$LOG_DIR/metadata.tsv`以下内容中的元数据相关联：

```
embeddings {
  tensor_name: 'word_embedding'
  metadata_path: '$LOG_DIR/metadata.tsv'
}
```

```
from tensorflow.contrib.tensorboard.plugins import projector

# Create randomly initialized embedding weights which will be trained.
N = 10000 # Number of items (vocab size).
D = 200 # Dimensionality of the embedding.
embedding_var = tf.Variable(tf.random_normal([N,D]),
name='word_embedding')

# Format:
tensorflow/contrib/tensorboard/plugins/projector/projector_config.proto
config = projector.ProjectorConfig()

# You can add multiple embeddings. Here we add only one.
embedding = config.embeddings.add()
embedding.tensor_name = embedding_var.name
# Link this tensor to its metadata file (e.g. labels).
embedding.metadata_path = os.path.join(LOG_DIR, 'metadata.tsv')

# Use the same LOG_DIR where you stored your checkpoint.
summary_writer = tf.summary.FileWriter(LOG_DIR)

# The next line writes a projector_config.pbtxt in the LOG_DIR.
TensorBoard will
# read this file during startup.
projector.visualize_embeddings(summary_writer, config)
```

TensorBoardLOG_DIR

```
tensorboard --logdir=LOG_DIR
```

“ ”

元数据

通常嵌入具有与之相关联的元数据（例如标签，图像）。元数据应存储在模型检查点之外的单独文件中，因为元数据不是模型的可训练参数。格式应为TSV文件（标签字符以红色显示），第一行包含列标题（以粗体显示），后续行包含元数据值：

```
Word\tFrequency
Airplane\t345
Car\t241
...
```

没有明确的密钥与主数据文件共享; 相反，假定元数据文件中的顺序与嵌入张量中的顺序相匹配。换句话说，第一行是标题信息，并且元数据文件中的第（i + 1）行对应于存储在检查点中的嵌入张量的第i行。

> 注意：如果TSV元数据文件只有一列，那么我们不期望一个标题行，并假设每一行都是嵌入的标签。我们包含这个异常，因为它与常用的“vocab文件”格式相匹配。

图片

如果您有与嵌入相关联的图像，则需要生成包含每个数据点的小缩略图的单个图像。这被称为 精灵图像。精灵应该具有相同数量的行和列，缩略图按行先顺序存储：第一个数据点放置在左上角，最后一个数据点在右下方：

0	1	2
3	4	五
6	7	

在上面的示例中注意到，最后一行不必被填充。有关sprite的具体示例，请参阅 10,000个MNIST数字（100x100）的精灵图像。

> 注意：我们目前支持高达8192px X 8192px的精灵。

构建精灵后，您需要告诉嵌入投影机在哪里找到它：

```
embedding.sprite.image_path = PATH_TO_SPRITE_IMAGE
# Specify the width and height of a single thumbnail.
embedding.sprite.single_image_dim.extend([w, h])
```

相互作用

嵌入式投影机有三个面板：

1. 顶部左侧的数据面板，您可以在其中选择运行，嵌入张量和数据列进行颜色标注。
2. 投影面板在左下角，您可以在其中选择投影类型（例如PCA，t-SNE）。
3. 右侧的检查员面板，您可以在其中搜索特定点，并查看最近邻居列表。

预测

嵌入式投影机具有降低数据集维数的三种方法：两个线性和一个非线性。每种方法都可用于创建二维或三维视图。

主成分分析减少尺寸的简单技术主要成分分析（PCA）。嵌入式投影机计算前10个主要组件。该菜单允许您将这些组件投影到两个或三个任意组合。PCA是线性投影，通常在检查全局几何时有效。

t-SNE流行的非线性降维技术是t-SNE。嵌入式投影仪提供二维和三维t-SNE视图。客户端对布局进行动画处理。因为t-SNE通常保留一些局部结构，因此对于探索当地社区和查找群集是有用的。虽然对于高维数据的可视化非常有用，但t-SNE图有时可能是神秘的或误导的。看到这篇伟大的文章，如何有效地使用t-SNE。

自定义您还可以根据文本搜索构建专门的线性投影，以便在空间中找到有意义的方向。要定义投影轴，请输入两个搜索字符串或正则表达式。该程序计算标签与这些搜索匹配的点集合的质心，并使用质心之间的差向量作为投影轴。

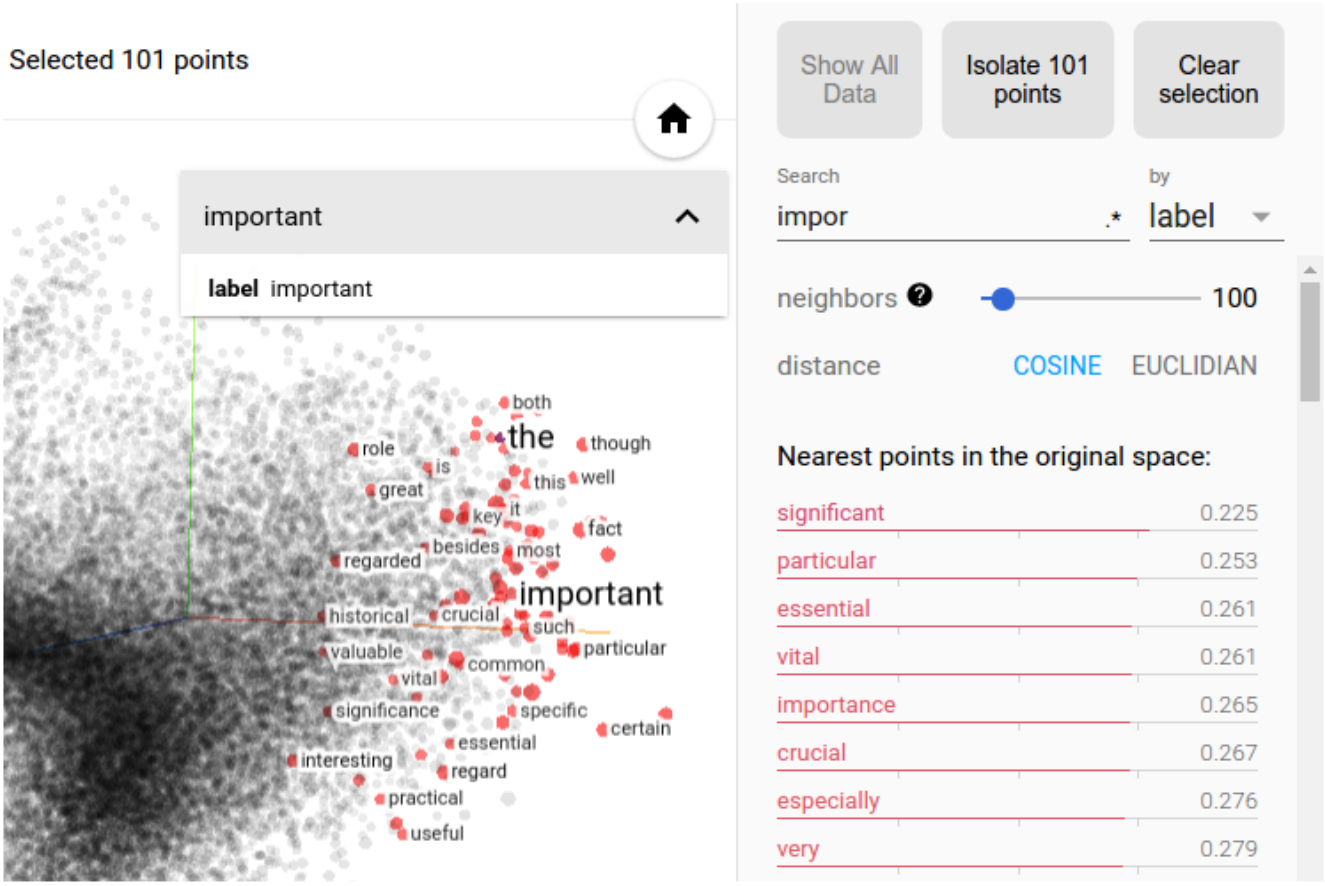
导航

要浏览数据集，您可以使用2D或3D模式浏览视图，使用自然的点击和拖动手势进行缩放，旋转和平移。单击一个点将导致右窗格显示最近邻居的显式文本列表以及当前点的距离。最近邻点自身在投影上突出显示。

缩放到集群中提供了一些信息，但是有时更有用的是将视图限制到一个点的一个子集，并仅在这些点上执行投影。为此，您可以通过多种方式选择点数：

1. 点击一个点后，也会选择最近的邻居。
2. 搜索后，选择与查询匹配的点。
3. 启用选择，点击一个点和拖动来定义一个选择球体。

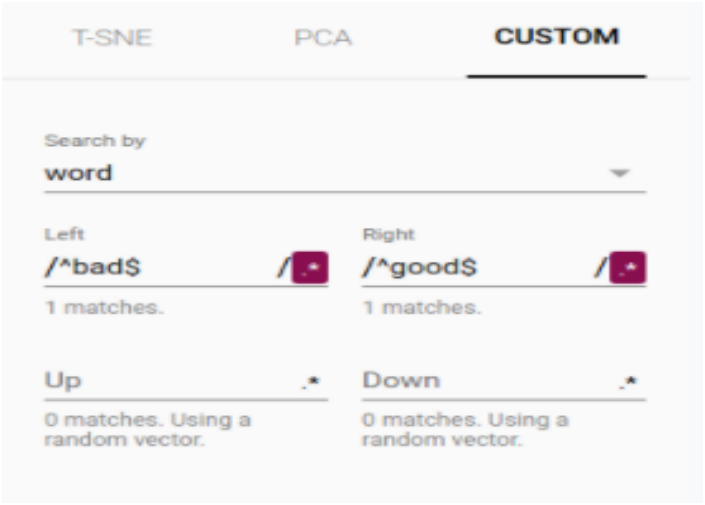
选择一组点后，您可以使用右侧的“检查器”窗格中的“隔离点”按钮隔离这些点，以进一步分析。



选择一个词嵌入数据集中最重要的邻居。

过滤与自定义投影的组合可以是强大的。在下面，我们过滤了100个最近邻居的“政治”，并把它们投射到“最佳” - “最差”向量作为x轴。y轴是随机的。

你可以看到，在右边我们有“想法”，“科学”，“视角”，“新闻”，而在左边我们有“危机”，“暴力”和“冲突”。



自定义投影控件。

将“政治”的邻居定制为"best" - "worst"向量。

协作功能

要分享您的发现，您可以使用右下角的书签面板将当前状态（包括任何投影的计算坐标）保存为小文件。然后可以将投影机指向一组一个或多个这些文件，生成下面的面板。然后，其他用户可以浏览一系列书签。