Google 概念字幕：图像字幕制作的新数据集和挑战

原创：Google谷歌开发者9 月 17 日

文 / Google AI 软件工程师 Piyush Sharma 和研究员 Radu Soricut

网络上有数十亿张图像，这有助于大众娱乐，以及向世界展示无数种主题。然而，对于有视觉障碍或由于网速太慢而无法加载图片的人士来说，其中很多视觉信息都无法获取。网站作者通过 Alt-text HTML 手动添加图像字幕，使更多人可以获取这些内容，然后我们可以使用文字转语音系统来展示对图像的自然语言描述。但是，只有很少一部分的网络图像添加了现有人工选编的 Alt-text HTML 字段。此外，虽然自动图像字幕制作有助于解决这一问题，但精准的图像字幕制作仍是一项颇具挑战性的工作，这需要提升计算机视觉和自然语言处理的现有技术水平。

注：Alt-text HTML 链接

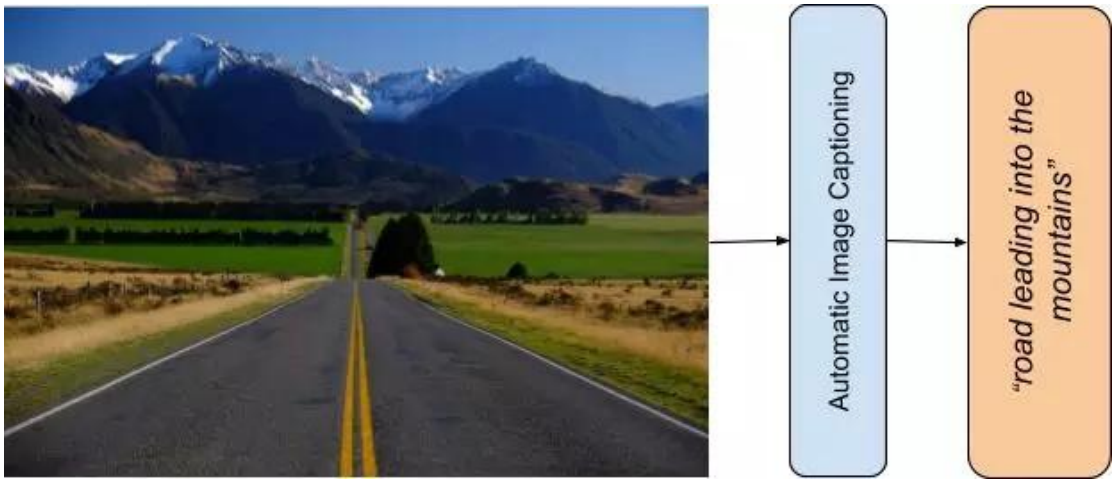
https://www.w3schools.com/tags/att_img_alt.asp

文字转语音系统链接

https://ai.googleblog.com/2017/12/tacotron-2-generating-human-like-speech.html

自动图像字幕制作链接

https://ai.googleblog.com/2014/11/a-picture-is-worth-thousand-coherent.html

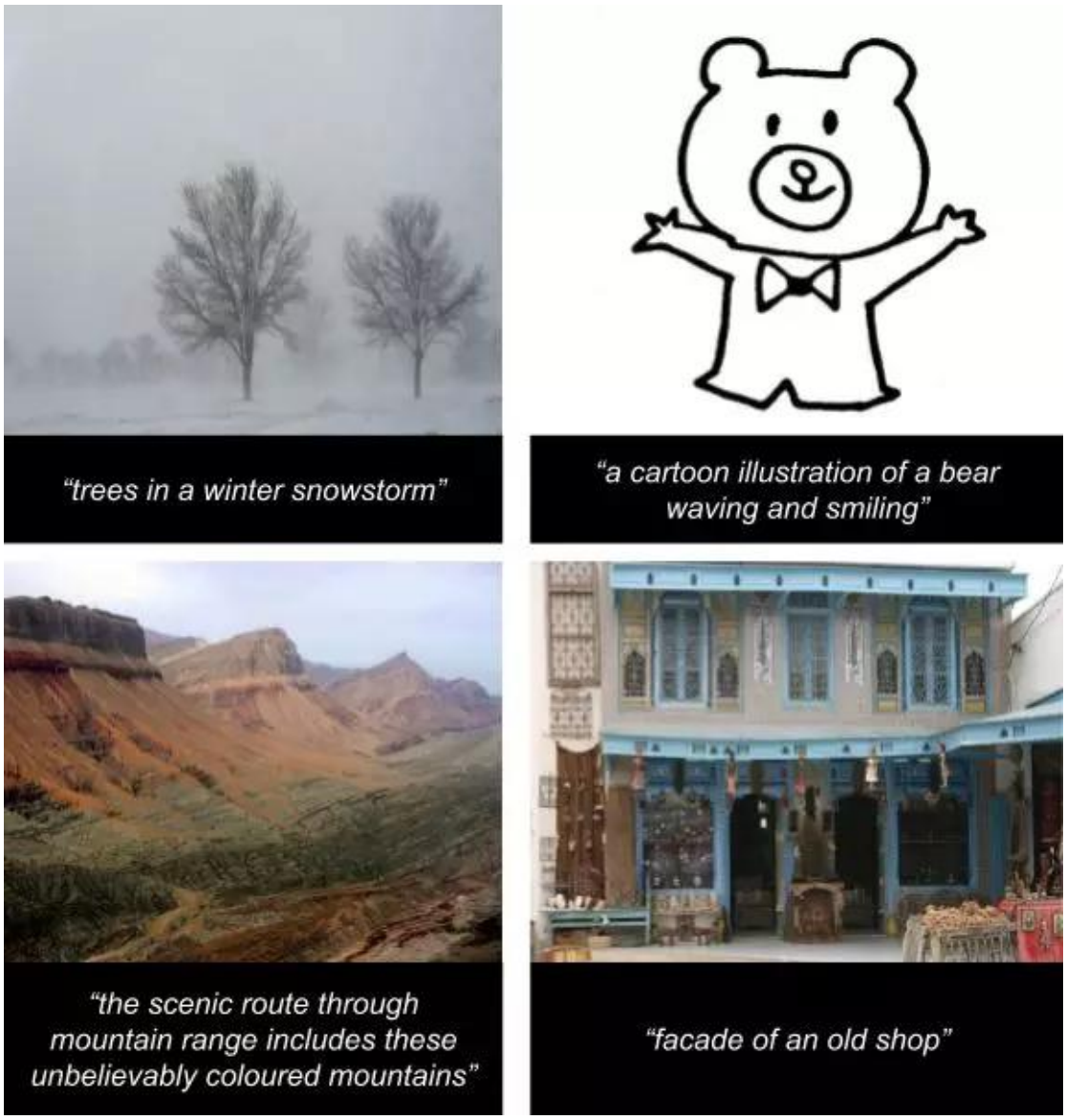


通过将图像字幕转换为文字，图像字幕制作可以帮助数百万有视觉障碍的人士。图像来自 Francis Vallance (Heritage Warrior)，在 CC BY 2.0 许可下使用

今天我们将介绍概念字幕，这是由大约 330 万图像/字幕对组成的新数据集；我们通过从数以十亿计的网页中自动提取和过滤图像字幕注解来加以创建。在 ACL 2018 发表的一篇论文中引入了 “概念字幕”，这代表在人工选编的 MS-COCO 数据集中，字幕图像增加了一个数量级。根据人类评分者的测量，机器选编的概念字幕准确率大约为 90%。此外，由于概念字幕中的图像是从网络中提取，所以与之前的数据集相比，其图像字幕风格更加多样，这便于我们更好地训练图像字幕制作模型。为了追踪图像字幕制作的进度，我们还将向机器学习社群发布概念字幕挑战，以便他们在概念字幕测试台上训练和评估自己的图像字幕制作模型。

注：论文链接
http://aclweb.org/anthology/P18-1238
MS-COCO 数据集链接

http://cocodataset.org/#home



概念字幕数据集中的图像和字幕示例
从左上角按顺时针方向开始，图片分别来自 Jonny Hunter、SigNote Cloud、Tony Hisgett 和 ResoluteSupportMedia。所有图片均在 CC BY 2.0 许可下使用

生成数据集

要生成概念字幕数据集，我们首先要从网络中获取带有 Alt-text HTML 属性的图像。我们自动筛选出带有特定属性的图像，以确保图像质量，同时避免不良内容，例如成人主题图像。然后，我们使用基

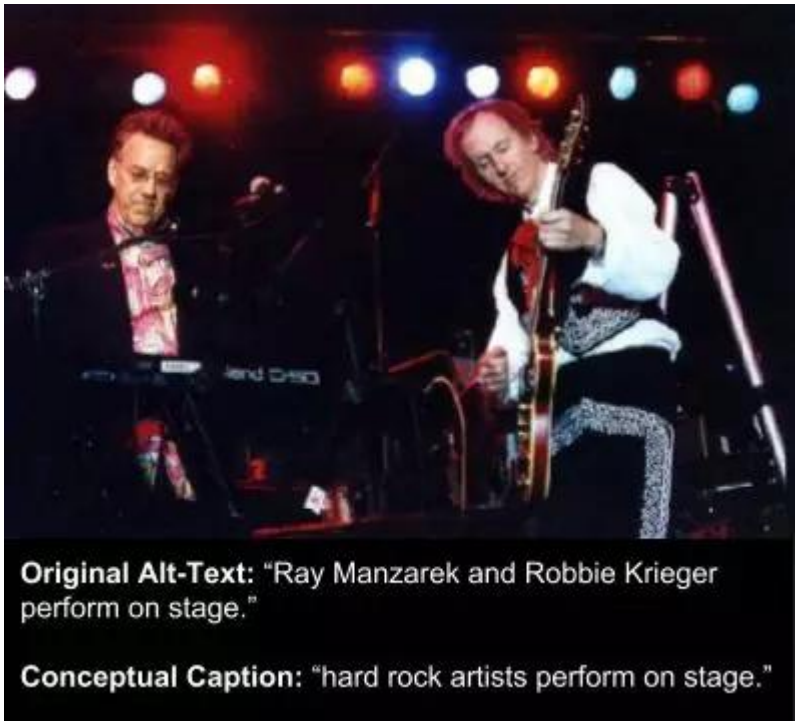
于文本的过滤方式，移除带有非描述性文本（例如 #标签、欠佳的语法或添加的语言与图像无关）的字幕；我们还舍弃带有高情感极性或成人内容的文本（如需更详细了解过滤标准，请参阅我们的论文）。我们使用现有的图像分类模型，以确保任何指定图像，在其 Alt-text（考虑词形变化）和图像分类器为该图像输出的标签之间有所重叠。

注：我们的论文链接
http://aclweb.org/anthology/P18-1238
图像分类模型链接
https://cloud.google.com/vision/
词形变化链接
http://www.aclweb.org/anthology/N15-1186

从特定名称到一般概念

虽然通过上述过滤的候选字幕往往是良好的 Alt-text 图像描述，但其中大多数都使用了专有名词（例如人物、地点、位置、组织等）。这会带来一些问题，因为图像字幕制作模型很难从输入图像像素中学会如此精细的专有名词推理，也很难同步生成自然语言描述 [1]。

为解决上述问题，我们编写了一个软件。该软件可以自动将专有名词替换为表达相同一般概念的单词，也就是使用它们的概念。在某些情况下，我们会移除专有名词以简化文本。例如，我们会替换人名（如将“前世界小姐 Priyanka Chopra 在红毯上”替换为“演员在红毯上”）、移除位置名称（将“洛杉矶演唱会上的人群”改为“演唱会上的人群”）和移除修饰语（如将“意大利美食”改为仅保留“美食”），并在有需要时，更正新组成的名词短语（如将“艺术家和艺术家”改为“艺术家”，请查看下方图示）。



图像来自 Rockoleando，在 CC BY 2.0 许可下使用

最后，我们汇总所有已解析的实体（例如，“艺术家”、“狗”、“附近” 等），并且只保留提及 100 次以上的候选类型，这一数量足以支持针对这些实体的表示学习。如此一来，我们保留了大约 1.6 万个实体概念，例如：“人”、“演员”、“艺术家”、“选手” 和 “图示”。我们保留的提及次数较少的概念包括 “法棍面包”、“缰绳”、“截止日期”、“部门” 和 “漏斗”。

最终，我们需要大约 10 亿个（英文）网页，其中包含超过 50 亿张候选图像，才能获取可供学习的简洁图像字幕数据集，其拥有超过 300 万个样本（淘汰率为 99.94%）。虽然我们可以调整控制参数，以较低的精确度在一个数量级中生成更多示例，但我们的参数还是偏向于高精确度。

数据集影响

为了测试数据集的实用性，我们在 Tensor2Tensor (T2T) 中使用 MS-COCO 数据集（使用 12 万张图像，每张图像上有 5 个人工注释的字幕）和新的概念字幕数据集（使用超过 330 万张图像，每张图像上有 1 个字幕），分别训练了基于 RNN 和基于 Transformer 的图像字幕制作模型。如需更详细了解模型架构，请参阅我们的论文。

注：RNN 链接 https://en.wikipedia.org/wiki/Recurrent_neural_network

Transformer 链接

https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

我们使用来自 Flickr30K 数据集的图像测试了这些模型（这些图像不在 MS-COCO 和概念字幕数据集的范围中），并为每个测试用例分配 3 位人类评分者来评估所产生的字幕。评估结果如下表所示。

注：Flickr30K 链接 http://web.engr.illinois.edu/~bplumme2/Flickr30kEntities/

根据这些结果，我们得出结论，在不考虑架构（即 RNN 或 Transformer）的情况下，与使用竞争方法训练的模型相比，使用概念字幕训练的模型能更好地形成一般概念。此外，我们还发现，无论使用其中哪个数据集进行训练，Transformer 模型的表现都比 RNN 模型要好。根据这些发现，我们得出的结论是，概念字幕让我们能够训练图像字幕制作模型，而且其在各种图像中的表现更佳。