神经网络

神经网络标准简化机器学习技术开发

由于没有标准的传输方式，开发人员不得不花费很多时间创建和维护用于导出和导入神经网络的转换器。

文/ Peter McGuinness，Khronos Group

存在几种用于深度学习的神经网络框架，所有这些框架都提供了距离特性和功能性。但是，在框架之间传递神经网络，会为开发人员增加额外的工作时间和工作量。由领先的硬件和软件公司组成的开放联盟 Khronos Group，创建了先进的加速标准，已经开发了 NNEF（神经网络交换格式），这是一种开放式免版税标准，它能使硬件制造商能够在训练框架和推理引擎之间，可靠地交换训练后的神经网络。

神经网络使用各种不同的框架进行训练，然后部署在类似范围广泛的推理引擎上，每个推理引擎都有自己的专有格式。这种多样性是非常可取的，但也是问题所在。开发人员必须构造专有的导出和导入程序，以便跨不同的推理引擎部署网络。对于负责构建、训练和部署网络的研究人员、开发人员和数据科学家而言，这是多余且不必要的工作。

问题始于可用的训练框架的数量：Caffe、TensorFlow、Chainer、Theano、Caffe2 和 PyTorch 等。尽管它们为开发人员提供了不同的功能和优化，但是每个训练框架对于在其中开发的网络也都有自己的格式，这使得框架之间的转换既麻烦又费时。

图1：神经网络交换表允许硬件制造商在训练框架和推理引擎之间，可靠地交换训练后的神经网络。

推理阶段也有类似的问题。在部署经过训练的网络之前需要转换器。如图 1 所示，其中每个推理引擎都需要针对每个训练框架的导入程序。这种转移过程为开发人员带来了额外的工作，却对创建或实施已部署的产品和系统没有额外的好处。

没有标准的传输方式，开发人员不得不花费很多时间创建和维护用于导出和导入神经网络的转换器，而且更长的开发时间也意味着浪费金钱。也许更重要的是，碎片化也威胁着机器学习领域的创新。通过花费不必要的时间来重新格式化导入和导出软件，开发人员将无法将更多的时间和精力在嵌入式视觉和推理中的机器学习方面，做出实际和具体的进步。

尽管这是一个复杂的问题，但是针对碎片化的解决方案却非常简单：将经过训练的神经网络，以用于神经网络的 PDF 形式，传输到推理引擎的简化过程。就像可移植文档格式（PDF）已经可以轻松地传输文档一样，NNEF 允许开发人员、研究人员和数据科学家轻松地将其网络从培训框架传输到推理引擎，而不必花费额外的时间进行翻译或添加导出程序。

通过提供机器学习生态系统所有部分都可以依赖的全面、可扩展、并且得到良好支持的标准，通用传输标准将减少浪费在传输和翻译上的时间，并最终使该行业向前迈进，实现真正的实施目标。

通过描述受过训练的网络，以及它们独立于训练框架和推理引擎的权重，NNEF 使数据科学家和工程师能够将经过训练的网络，从它们选择的训练框架转移到各种推理引擎中。这种交换的自由，再加上时间和开发成本的节省，将为开发人员提供急需的时间和手段，使其专注于创新，而不是忙于没有附加值的工作，从而有助于刺激新兴机器学习技术的发展。

Khronos Group 允许行业成员以直接贡献者或顾问的身份参与 NNEF 的开发，从而制定满足整个行业需求的透明标准。