## FeatGraph：图神经网络系统的灵活高效后端

### 1 Motivate：

图网络包含了很多稀疏和稠密的运算，稀疏是指其邻阶矩阵是非常庞大和稀疏的，对节点进行聚合需要乘这样一个邻阶矩阵，那就是一个超大的稀疏运算，论文中指出：在GNN model中稀疏计算大概占上60%。

论文里指出的已有的一些工作还存在的不足：

* 一些厂商提供的算子lib，例如MKL、cuSPARSE，针对图神经网络的加速支持还不好，具体表现为只支持图网络用到的一小部分算子的加速。
* 已有的一些图的算法，BFS、PageRank的节点都是和标量绑定的，但是图网络处理的问题节点与tensor绑定。
* 现有的图处理系统将边计算作为调度的单元，而GNN中关于边有着大量的稀疏计算.
* 之前的一些工作关心了在图网络中通过图分割的方式提高cache利用率的问题，但是没有考虑到feature这个维度。

本文通过张量编译解决问题。张量编译在IR中将计算定义与计算调度分开，以便专注于调度部分的性能优化。现有的张量编译器（TVM等）也是面向稠密矩阵的，本文扩展TVM以便于支持广义SpMM和SDDMM，并通过设计两粒度编程接口允许users在每个顶/边上自定义feature计算。

### 2 主要工作

#### 2.1 什么是SpMM和SDMM



以GCN为例，在以顶点为中心的计算中，稀疏的邻接矩阵A要与相对稠密的特征矩阵Xv做运算，就是SpMM（sparse-dense matrix multiplication）。



类似的，在边上的计算的做法是将源顶点与目的顶点的feature点乘再乘以邻接矩阵A，就是SDDMM（sampled dense-dense matrix multiplication）。

#### 2.2 系统设计与优化

1