doi:10.1631/FITEE.1800469

题目：用于语音识别的二值神经网络

概要：近年来，在语音识别的声学建模中，深度神经网络(DNNs)明显优于高斯混合模型。然而，推断阶段巨大的计算量使其难以部署在低功耗的嵌入式模型上。为此，稀疏性和低精度定点量化技术被广泛使用。为降低推理阶段计算量，本文开发了用于语音识别的二进制神经网络，并实现了高速的二值矩阵乘法。在中央处理器(CPU)和图形处理单元(GPU)上，二值矩阵乘法的运行速度是浮点矩阵乘法的 5~7 倍。针对大规模连续语音识别的声学建模，提出多种二值神经网络及相关模型优化算法。为提高二值模型的精度，探索了从浮点模型到二值模型的知识蒸馏技术。在标准的 Switchboard 语音识别任务上，该二值神经网络模型比浮点神经网络模型速度提高 3~4 倍。借助知识蒸馏技术，二值深度神经网络或卷积神经网络相对其浮点神经网络的词错误率增加可以保持在 15%以内。若只二值化卷积神经网络的卷积层，词错误率增加几乎可忽略。

关键词：语音识别；二值神经网络；二值矩阵乘法；知识蒸馏；位 1 计数