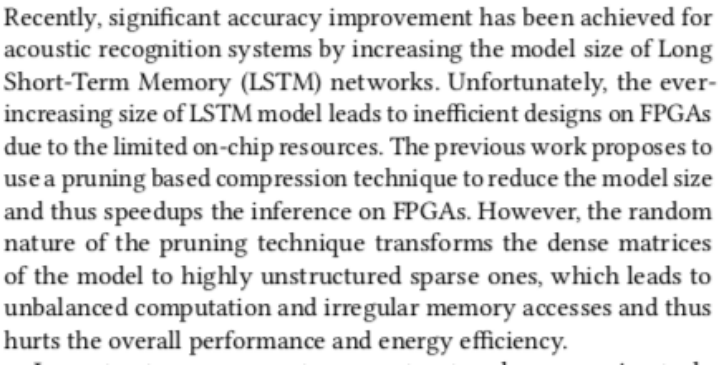
SIAM-IMA Etymo workshop - text extraction

Steven Elsworth

June 13, 2018

Text extraction


Recently, significant accuracy improvement has been achieved for acoustic recognition systems by increasing the model size of Long Short-Term Memory (LSTM) networks. Unfortunately, the ever-increasing size of LSTM model leads to inefficient designs on FPGAs due to the limited on-chip resources. The previous work proposes to use a pruning based compression technique to reduce the model size and thus speedups the inference on FPGAs. However, the random nature of the pruning technique transforms the dense matrices of the model to highly unstructured sparse ones, which leads to unbalanced computation and irregular memory accesses and thus hurts the overall performance and energy efficiency.

Recently, significant accuracy improvement has been achieved for acoustic recognition systems by increasing the model size of Long Short-Term Memory (LSTM) networks. Unfortunately, the ever-increasing size of LSTM model leads to inefficient designs on FPGAs due to the limited on-chip resources. The previous work proposes to use a pruning based compression technique to reduce the model size and thus speedups the inference on FPGAs. However, the random nature of the pruning technique transforms the dense matrices of the model to highly unstructured sparse ones, which leads to unbalanced computation and irregular memory accesses and thus hurts the overall performance and energy efficiency.

Methods

- pdftotxt
- Textract (Backend pdftotxt)
- PyPDF2
- pdfminer
- pyocr (backend tesseract)

See Jupyter Notebook.

Problems faced

Give examples of :

- Equation conversion
- Figure conversion
- White space
- Merged words
- Page Columns
- Time