MDH+ATF: Code Generation & Optimization for Deep-Learning Computations

Universität Münster

Ari Rasch, Richard Schulze, …

a.rasch@uni-muenster.de

Generation

Let T and T' be two arbitrary types. A function $h : T[\,N_1\,]\ldots[\,N_d\,] \to T'$ on d-dimensional arrays is called a *Multi-Dimensional Homomorphism (MDH)* iff there exist *combine operators* $\circledast_1, \ldots, \circledast_d : T' \times T' \to T'$, such that for each $k \in [1, d]$ and arbitrary, concatenated input array $a +\!\!+_k b$ in dimension k:

$$h(\; a +\!\!+_k b \;) \;=\; h(a) \;\circledast_k\; h(b)$$

DL computations can be expressed as **MDH functions**, and **GPU/CPU/… code generated** and **optimized** according to **MDH formalism** [1]

[1] Rasch, *(De/Re)-Composition of Data-Parallel Computations via Multi-Dimensional Homomorphisms*, **TOPLAS'24**

MDHs can be uniformly expressed via our `md_hom` higher-order function:

$$\texttt{md\_hom}(\; f,\; (\circledast_1,\ldots,\circledast_D)\;)(\;a\;) \;:=\; \underset{i_1 \in I_1}{\circledast_1} \;\ldots\; \underset{i_D \in I_D}{\circledast_D} \; f(\; a[i_1,\ldots,i_D]\;)$$

Convolutions

```
CONV<…> = out_view<…>( O:(n,p,…)->(n,p,q,k) ) o
          md_hom<…>( *, (++,++,++,++ , +,+,+) ) o
            inp_view<…>( I:(n,p,…)->(n,p+r,q+s,c) , F:(n,p,…)->(k,r,s,c) )
```

Linear Algebra

```
MatMul<…> = out_view<…>( C:(i,j,k)->(i,j) ) o
            md_hom<…>( *, (++, ++, +) ) o
              inp_view<…>( A:(i,j,k)->(i,k), B:(i,j,k)->(k,j) )
```

…

Point-Wise OPs

```
BiasAdd<NHWC><…> = out_view( O:(n,h,w,c)->(n,h,w,c) ) o
                   md_hom<…>( +, (++,++,++,++) ) o
                     inp_view<…>( I1:(n,h,w,c)->(n,h,w,c),I2:(n,h,w,c)->(c) )
```

```
out_view(...) ∘
   md_hom(...) ∘
      inp_view(...)
```

Code Generation

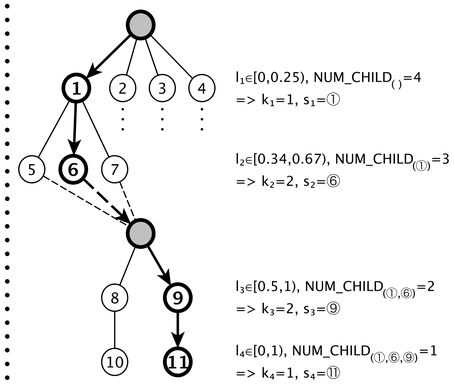
(auto-tunable)

NVIDIA CUDA · OpenMP · OpenCL™

Optimization

Our **Auto-Tuning Framework (ATF)** is a **general-purpose approach** that **automatically optimizes** (*auto-tunes*) programs with ***constrained tuning parameters*** [2]

[2] Rasch, Schulze, Steuwer, Gorlatch, *Efficient Auto-Tuning of Parallel Programs with Interdependent Tuning Parameters via Auto-Tuning Framework (ATF)*, **TACO'21**



CoT (Chain-of-Trees)
A new search space structure for constrained tuning parameters

```
#atf::tp name       /* name       */
        range      /* range      */
        constraint /* constraint */
```

We extend the traditional definition of *tuning parameters by a* ***parameter constraint***.

ATF efficiently
generates / stores / explores
the search spaces of
constrained tuning parameters

up to 2.67x speedups over *NVIDIA cuBLAS*

up to 3.5x speedups over NVIDIA cuDNN

MDH+ATF achieves on **GPUs, CPUs, …**
often higher **Performance & Portability & Producitivity**
than well-performing
hand- and **machine-optimized approaches** [1]

up to 9.01x faster than *Intel oneDNN*

up to 3.01x faster than TVM

Highlights only!