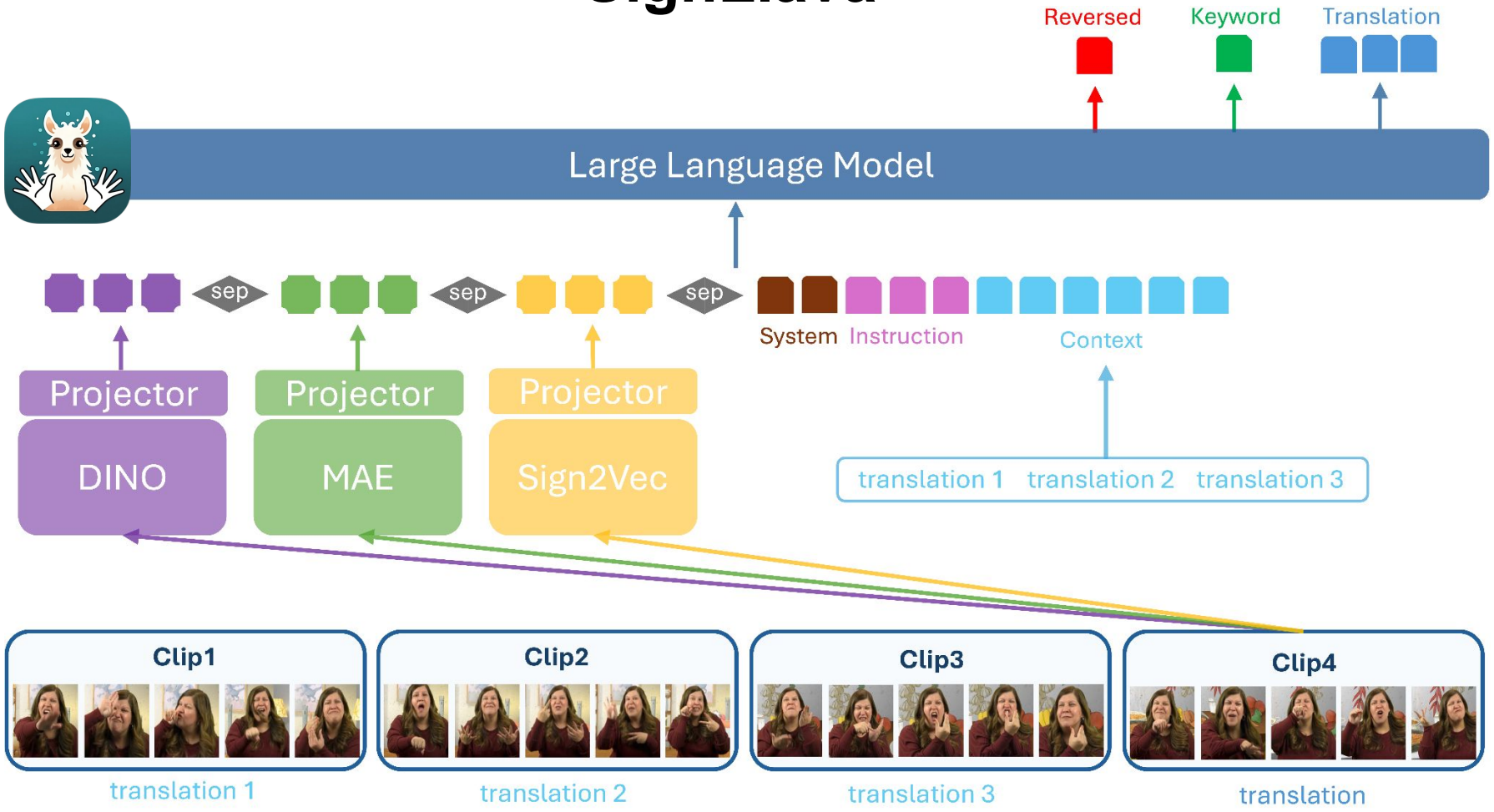
SignLlava
LUMI hackathon 2024

Tomas Zelezny, Jakub Straka, Matej Sieber

SignLlava


Reversed
Keyword
Translation
Large Language Model
sep
sep
sep
System
Instruction
Context
Projector
Projector
Projector
DINO
MAE
Sign2Vec
translation 1
translation 2
translation 3
Clip1
Clip2
Clip3
Clip4
translation 1
translation 2
translation 3
translation

Hackathon goals

- Optimization of our model based on Llama3
- Multi-node training
- Move to Llama-70b instead of Llama-8b
- Dealing with some existing problems
  - Correctly use flash-attention (1Torch was not compiled with flash attention)
  - Correctly use memory efficient attention (1Torch was not compiled with memory efficient attention)
  - SDPA attention implementation warnings for multi-node ROCM
- Moving our codes to latest Llama3.1 version

Dealing with some existing problems

- Correctly use flash-attention (1Torch was not compiled with flash attention)
- Correctly use memory efficient attention (1Torch was not compiled with memory efficient attention)
- Slow model loading

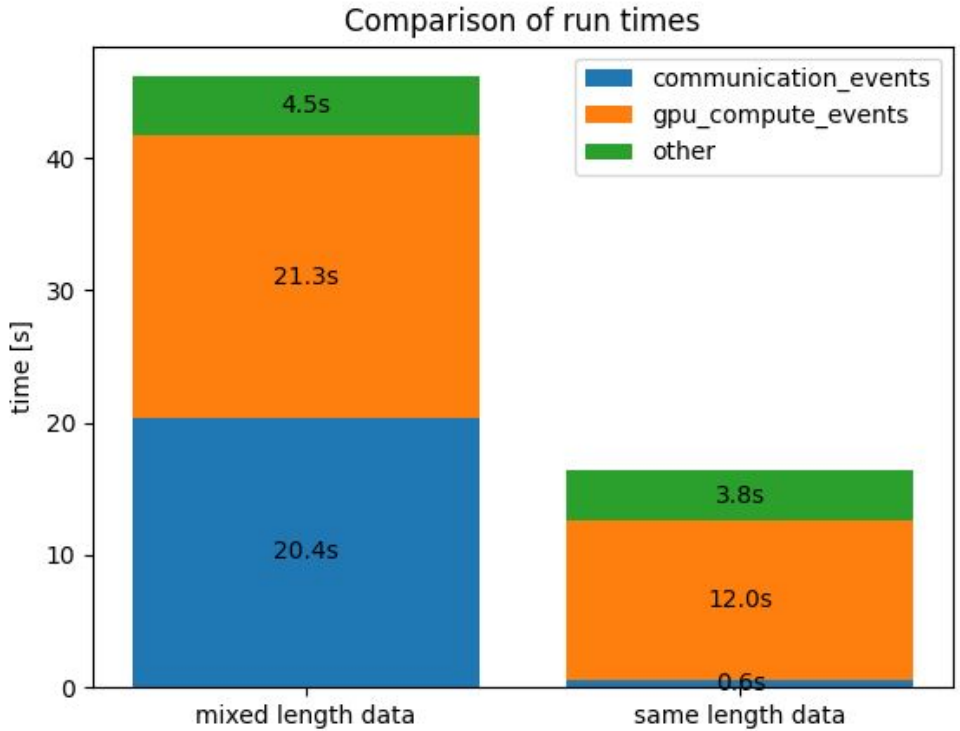
- Solved with new singularity image rocm6.1.3 + venv

Profiling of our model based on Llama3 8B

- **Tools:** rocprof v1 , omnitrace, torch.profiler
- Estabilished workflow setup
- Better understanding of the execution

- **Possible improvement:** Refactor of dataloader

Load balancing issue



Using Llama-70b instead of Llama-8b

- Successfully training with both 1-node and multi-node setups
- ZeRO3 parallelism
- Core parameter: stage3_prefetch_bucket_size
- Found bug in official release accelerate==1.0.1

Multi-node training

- Successfully training without deepspeed launcher
  - Llama3-8B
  - Llama3-70B
- Deepspeed launcher doesn’t work yet
- A lot of sanity-check experiments
  - Weak spots in our code
  - Batch size and prefetch-bucket-size importance
- Not profiled yet

Moving to latest Llama3.1 version

What is left to do

- Improve our dataloader for GPU workload balance
- Profile the 70B model
- Profile multiple nodes
- Estimate scalability and best setup for our use-case

Highlights of this week

- New singularity image
- Discovery of load balance issue
- Llama3-70B model training
- Multi-node setups