基于LLM大模型构建中文对话引擎：NEMO+CHATGLM

NVIDIA企业级开发者社区 李奕澎

AGENDA

- NVIDIA NeMo toolkit介绍
- 中文LLM大语言模型ChatGLM介绍
- 代码实践：NeMo结合ChatGLM快速构建中文场景的对话引擎

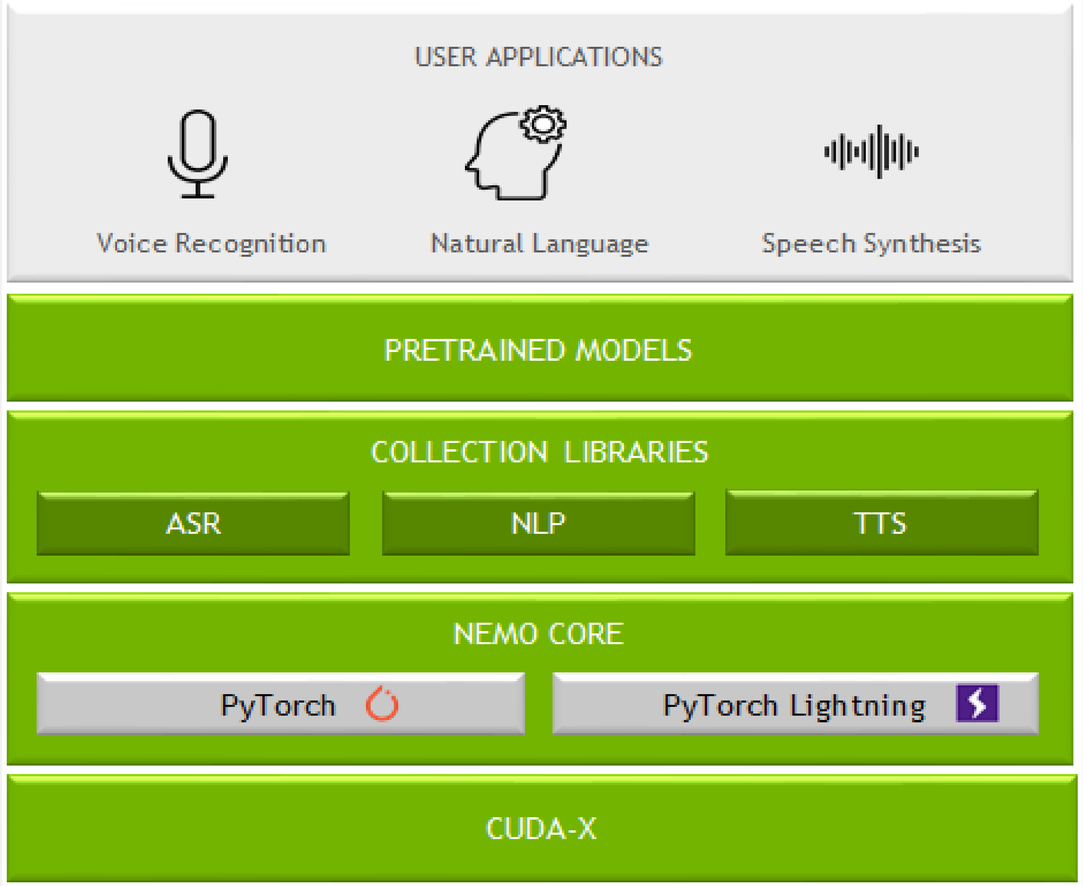
NVIDIA NEMO TOOLKIT

用于构建 SOTA 模型的对话式AI工具库

- 可以构建基于深度学习的语音和语言理解模型
- 集成NLP自然语言处理， ASR自动语音识别， TTS语音合成
- 多语言支持
- 完全开源，简单易用的APIs
- 整合PyTorch & PyTorch Lightning深度学习框架
- 基于GPU并行加速计算框架CUDA
- 多卡分布式训练混合式精度计算加速训练过程
- 100+ NGC预训练模型拿来即用

https://ngc.nvidia.com/catalog/containers/nvidia:nemo

https://github.com/NVIDIA/NeMo



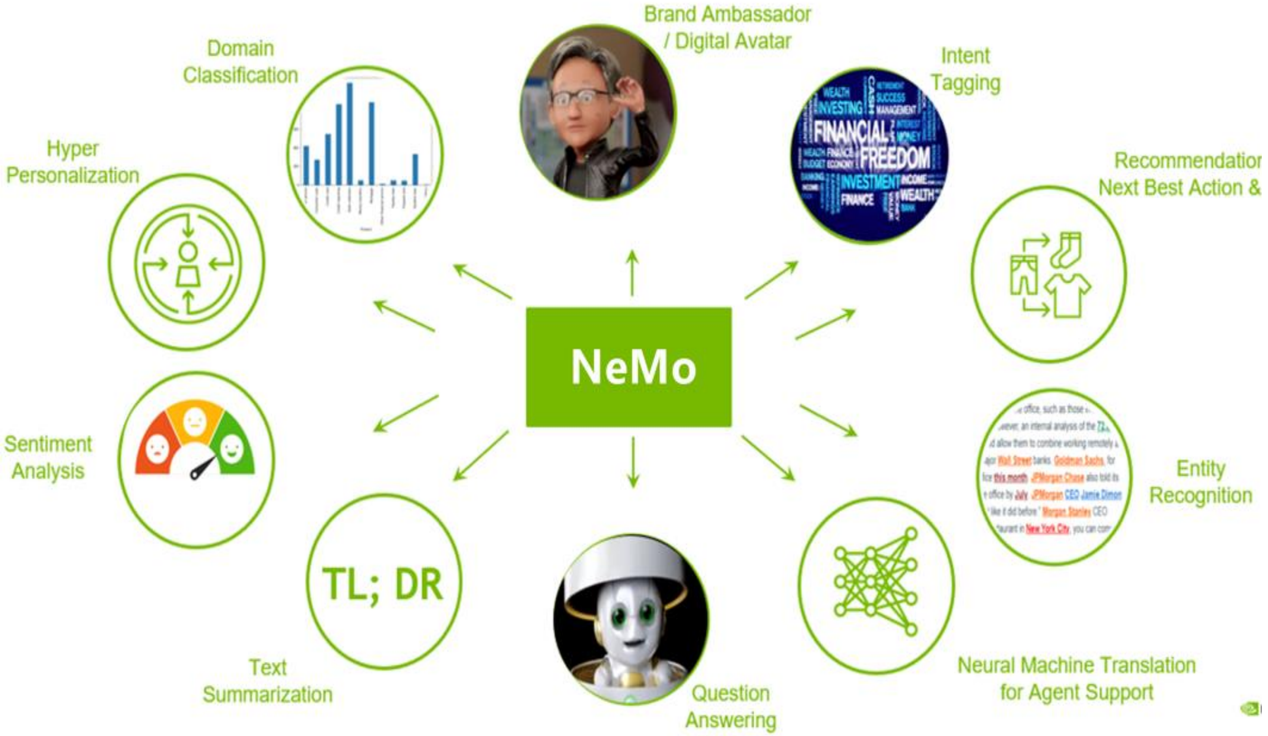
ASR Support in Multiple Languages

- Catalan
- French
- German
- Italian
- Mandarin
- Polish
- Russian
- Spanish

NLP Support in Multiple Languages

- Spanish
- Russian
- Mandarin
- German
- French

NeMo能做什么


NeMo
Brand Ambassador / Digital Avatar
Intent Tagging
Recommendation Next Best Action &
Entity Recognition
Neural Machine Translation for Agent Support
Question Answering
Text Summarization
TL; DR
Sentiment Analysis
Hyper Personalization
Domain Classification


NVIDIA Omniverse Avatar
for Project Tokkio
FEATURED
CHEESEBURGER
ENTREES
SIDES
NVIDIA

NeMo 推理代码示例

3行代码语音识别ASR

```
import nemo.collections.asr as nemo_asr
asr_model = nemo_asr.models.EncDecCTCModel.from_pretrained(model_name="stt_zh_citrinet_512")
result = asr_model.transcribe(paths2audio_files = ["/root/a1.wav"])
print(result)
```

3行代码机器翻译NMT

```
import nemo.collections.nlp as nemo_nlp
nmt_model = nemo_nlp.models.MTEncDecModel.from_pretrained(model_name="nmt_zh_en_transformer6x6")
result = nmt_model.translate(text)
print(result)
```

3行代码语音合成TTS

```
from nemo.collections.tts.models import FastSpeech2HifiGanE2EModel
model = FastSpeech2HifiGanE2EModel.restore_from("/home/nvidia/ms_chatbot/tts_en_e2e_fastspeech2hifigan.nemo")
tokens = model.parse(response)
audio = model.convert_text_to_waveform(tokens=tokens)
```

```
import IPython
IPython.display.Audio(audio.to('cpu').detach().numpy(), rate=22050)
```

中文LLM大语言模型ChatGLM介绍

ChatGLM-6B 是一个开源的、支持中英双语的大语言模型，使用了和 ChatGPT 相似的技术，针对中文问答和对话进行了优化，结合模型量化技术，用户可以在消费级的显卡上进行本地部署。ChatGLM2-6B 是ChatGLM-6B 的第二代版本，更强大的性能、更长的上下文、更高效的推理、更开放开源。

THUDM/chatglm2-6b
Updated Jul 20 · 2.45M · 1.54k

THUDM/chatglm-6b
Updated Jul 19 · 985k · 2.58k

THUDM/chatglm-6b-int8
Updated May 15 · 11.8k · 62

THUDM/chatglm2-6b-int4
Updated 21 days ago · 45.6k · 160

THUDM/chatglm-6b-int4-qe
Feature Extraction · Updated Apr 14 · 4.19k · 79

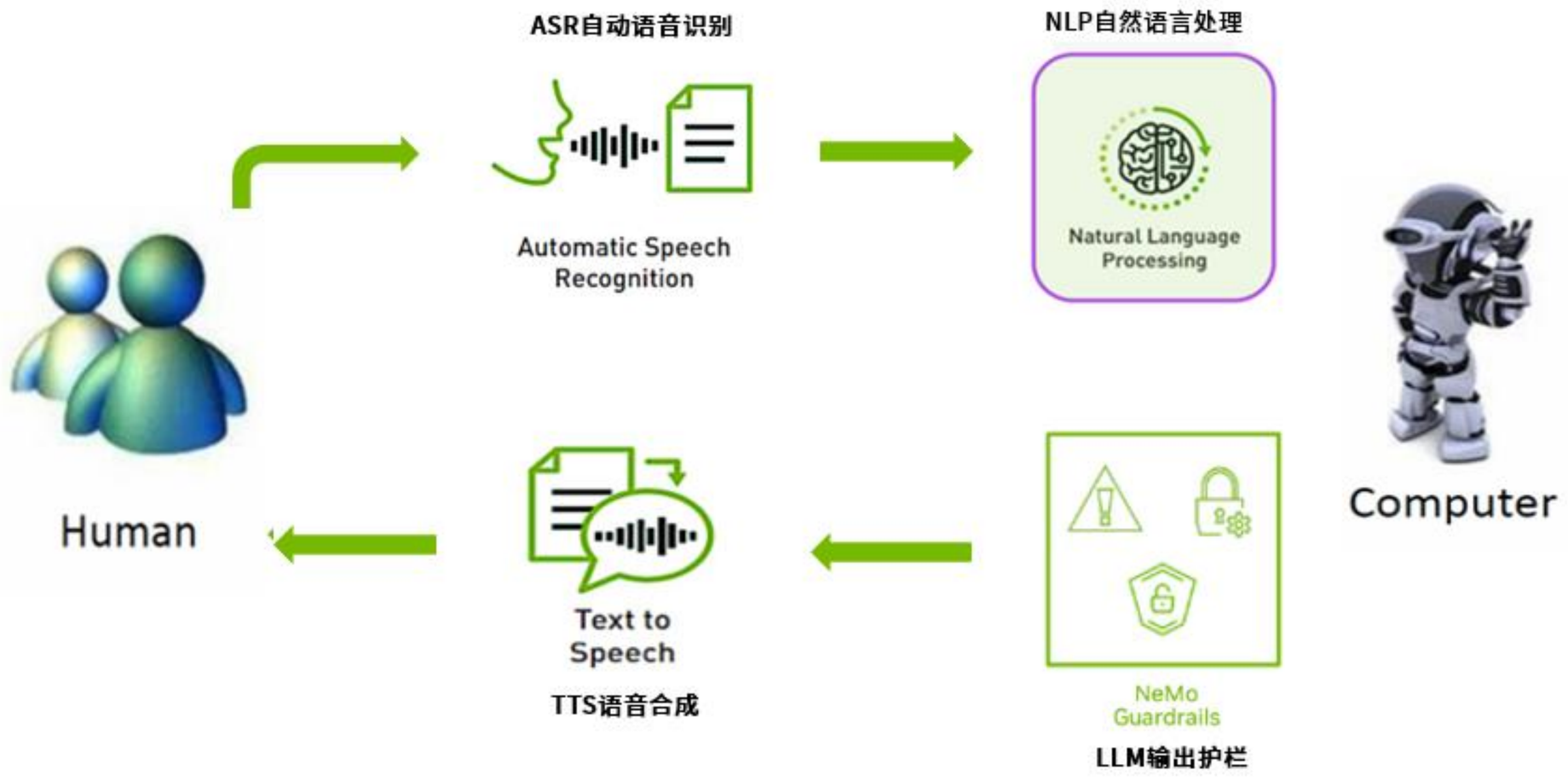
THUDM/glm-10b
Feature Extraction · Updated Mar 2 · 1.45k · 30

代码调用 ChatGLM2-6B 模型来生成对话

```
from transformers import AutoTokenizer, AutoModel
tokenizer = AutoTokenizer.from_pretrained("THUDM/chatglm2-6b", trust_remote_code=True)
model = AutoModel.from_pretrained("THUDM/chatglm2-6b", trust_remote_code=True, device='cuda')
model = model.eval()
response, history = model.chat(tokenizer, "你好", history=[])
print(response)
```

👋!我是人工智能助手 ChatGLM2-6B,很高兴见到你,欢迎问我任何问题。

NeMo+ChatGLM对话式AI-人机交互



NeMo对话式AI-人机交互demo

```
import nemo.collections.asr as nemo_asr
citrinet = nemo_asr.models.EncDecCTCModel.from_pretrained(model_name="stt_zh_citrinet_512")
asr_result = citrinet.transcribe(paths2audio_files=["tianqi.wav"])
asr_result[0]
```

'今天天气如何'

ASR语音识别
构建机器的“耳朵“实现听写的过程

```
from transformers import AutoTokenizer, AutoModel
tokenizer = AutoTokenizer.from_pretrained("THUDM/chatglm2-6b", trust_remote_code=True)
model = AutoModel.from_pretrained("THUDM/chatglm2-6b", trust_remote_code=True).quantize(8).cuda()
model = model.eval()
response, history = model.chat(tokenizer, asr_result+"10个字以内", history=[])
print(response)
```

阳光明媚

NLP+LLM
构建机器的“ 大脑“理解文字聊天对话

```
from nemo.collections.tts.models import FastSpeech2HifiGanE2EModel
model = FastSpeech2HifiGanE2EModel.restore_from("/home/nvidia/ms_chatbot/tts_en_e2e_fastspeech2hifigan.nemo")
tokens = model.parse(response)
audio = model.convert_text_to_waveform(tokens=tokens)
```

```
import IPython
IPython.display.Audio(audio.to('cpu').detach().numpy(), rate=22050)
```

0:00 / 0:01

TTS语音合成
构建机器的“嘴巴“把文字用声音说出来

THANK YOU!

NVIDIA®