

Инструкция по установке экземпляра

TextAI.NTR



Инструкция по установке экземпляра ПО, предоставленного для проверки

Требования к системе

Ubuntu 22.04 или новее.

GPU Nvidia A100 80G

GPU Nvidia T4 16G

Процедура установки

Для установки и запуска системы необходимы файл `docker-compose.yaml` и `docker` образы `backend` и `frontend`. Их можно загрузить с корпоративного Docker реестра `registry.ntrlab.ru`.

Установка NVIDIA драйверов

Поскольку образы используют NVIDIA CUDA, нужно установить Docker, NVIDIA драйверы и NVIDIA Container Toolkit на хостовой машине.

Обновите систему:

```
sudo apt update && sudo apt upgrade -y
```

Проверьте доступные версии драйверов NVIDIA с помощью команды:

```
apt search nvidia-driver
```

Установите актуальную версию драйвера и перезагрузите систему:

```
sudo apt install -y nvidia-driver-<version>  
sudo reboot
```

Убедитесь, что драйвер установлен.. Вы должны увидеть таблицу с информацией о GPU:

```
nvidia-smi
```

Логин в корпоративном Docker реестре `registry.ntrlab.ru`

Используйте предоставленный логин/пароль

```
docker login registry.ntrlab.ru
```

Установка Docker

Установите зависимости:

```
sudo apt install -y apt-transport-https ca-certificates curl software-properties-common  
docker-compose-plugin
```

Добавьте официальный ключ GPG Docker:

```
curl -fsSL https://download.docker.com/linux/ubuntu/gpg | sudo gpg --dearmor -o  
/usr/share/keyrings/docker-archive-keyring.gpg
```

Добавьте репозиторий Docker:

```
echo "deb [arch=$(dpkg --print-architecture)
signed-by=/usr/share/keyrings/docker-archive-keyring.gpg]
https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable" | sudo tee
/etc/apt/sources.list.d/docker.list > /dev/null
```

Установите Docker:

```
sudo apt update
```

```
sudo apt install -y docker-ce docker-ce-cli containerd.io
```

Проверьте, что Docker работает:

```
sudo systemctl status docker
```

```
sudo docker --version
```

Установка NVIDIA Container Toolkit

Добавьте репозиторий NVIDIA Container Toolkit:

```
curl -fsSL https://nvidia.github.io/libnvidia-container/gpgkey | sudo gpg --dearmor -o
/usr/share/keyrings/nvidia-container-toolkit-keyring.gpg
```

```
echo "deb [signed-by=/usr/share/keyrings/nvidia-container-toolkit-keyring.gpg]
https://nvidia.github.io/libnvidia-container/stable/ubuntu18.04/$(dpkg --print-architecture) /" |
sudo tee /etc/apt/sources.list.d/nvidia-container-toolkit.list
```

Установите toolkit:

```
sudo apt update
```

```
sudo apt install -y nvidia-container-toolkit
```

Настройте Docker для использования NVIDIA runtime:

```
sudo nvidia-ctk runtime configure --runtime=docker
```

```
sudo systemctl restart docker
```

Проверьте, что GPU доступен для Docker. Вы должны увидеть вывод `nvidia-smi` из контейнера с доступными GPU на хостовой машине:

```
sudo docker run --rm --gpus all nvidia/cuda:12.2.0-runtime-ubuntu22.04 nvidia-smi
```

Загрузка большой языковой модели

```
sudo apt install -y python3-pip
```

```
pip3 install huggingface_hub
```

```
huggingface-cli download hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4 --local-dir
/path/to/models/llama-3.1-70b
```

Запуск контейнеров

Для запуска системы используйте приложенный `docker-compose.yaml`:

```
docker compose -f <путь к файлу docker-compose.yaml> up -d
```

Контакты технических специалистов

Чурилов Илья Дмитриевич, ведущий разработчик
почтовый адрес: ichurilov@ntrlab.com

Мынка Владимир Александрович, разработчик
почтовый адрес: vmynka@ntr.ai

Матенков Андрей Николаевич, специалист по установке и настройке LLM
почтовый адрес: amatenkov@ntr.ai

Козаченко Александр Игоревич, менеджер проекта
почтовый адрес: akozachenko@ntr.ai