10-414/714 – Deep Learning Systems: Algorithms and Implementation

Customizing Pretrained models

Fall 2025

Tianqi Chen (this time) and Tim Dettmers
Carnegie Mellon University

Outline

Image generation models

Language models

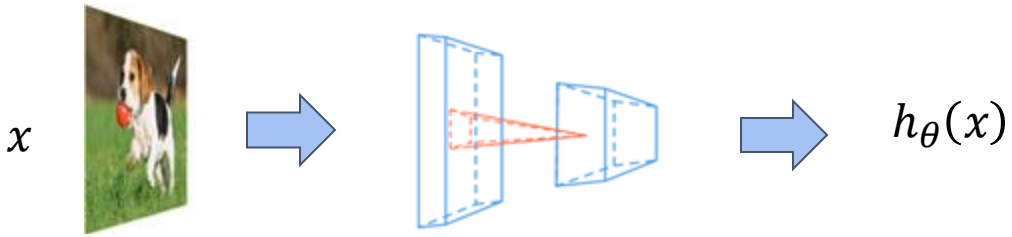
Outline

Image generation models

Language models

Recap: Element of Machine Learning

1. The hypothesis class:



2. The loss function:

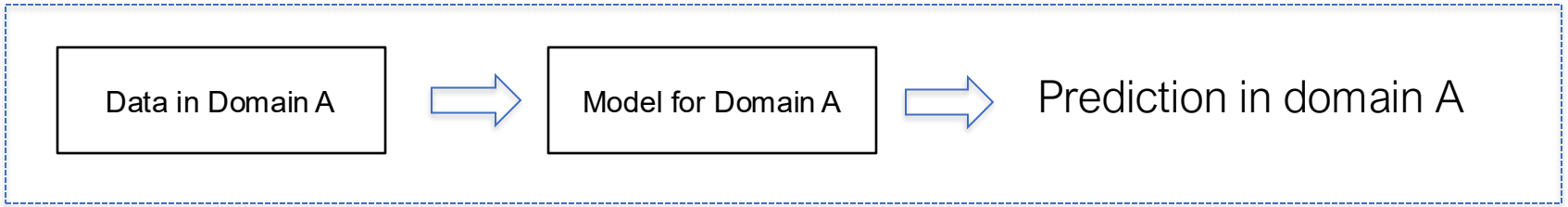
$$l(h_\theta(x), y) = -h_y(x) + \log \sum_{j=1}^{k} \exp\left(h_j(x)\right)$$

3. An optimization method:

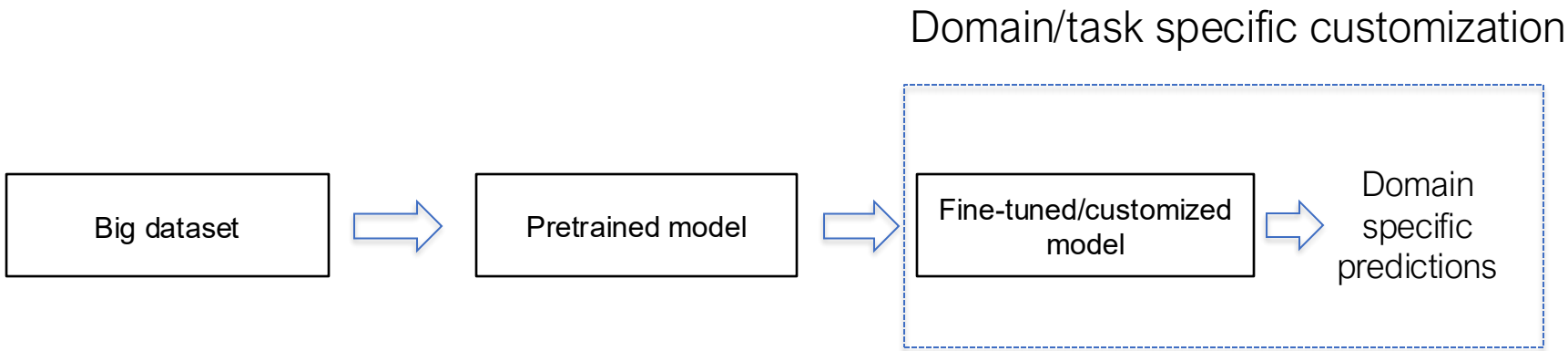
$$\theta := \theta - \frac{\alpha}{B} \sum_{i=1}^{B} \nabla_\theta \ell\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right)$$

Making use of pretrained model

Per domain approach

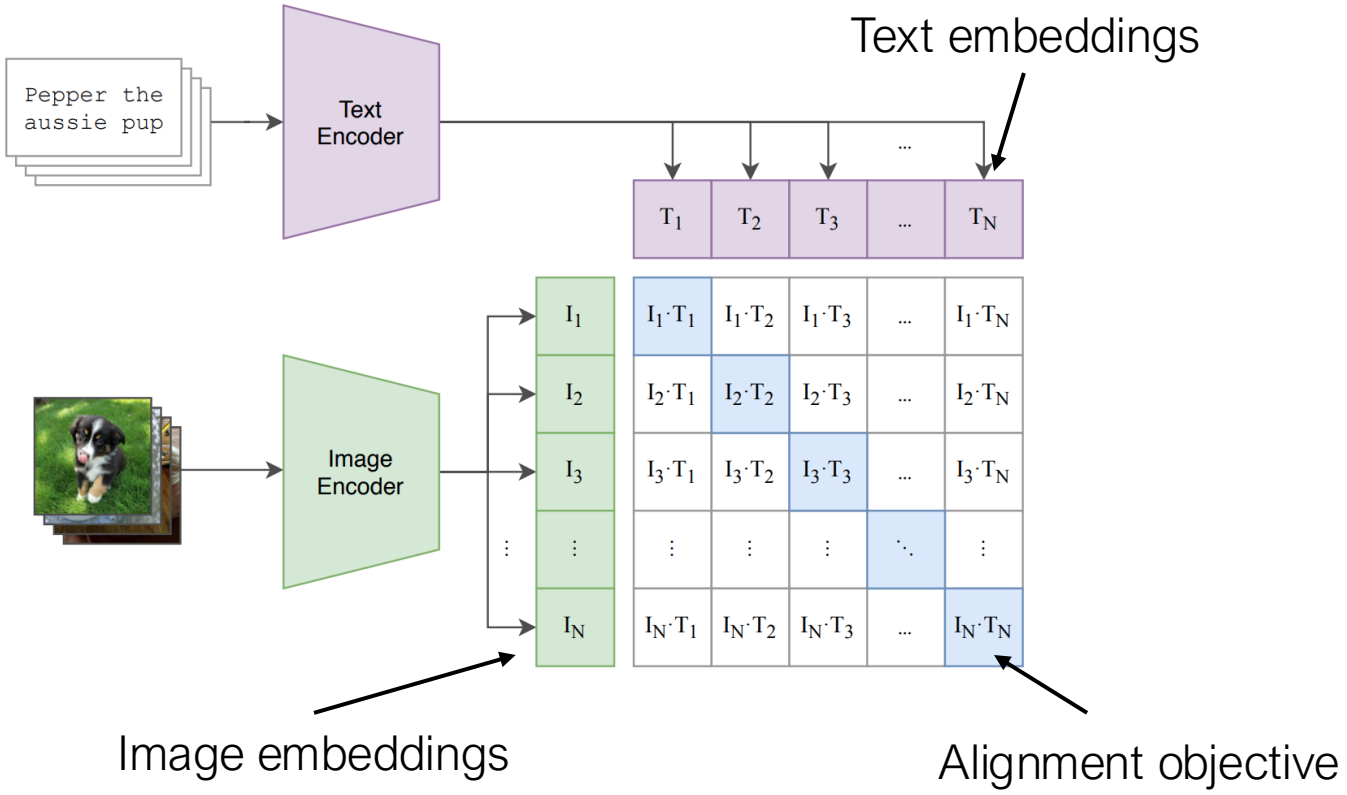


Leveraging pretraining



What are typical ways to leverage a pretrained model?

CLIP: Image and Text Embedding



These embeddings can be used in many downstream tasks

Radford et.al Learning Transferable Visual Models From Natural Language Supervision

Recap: Diffusion Model

$q(x_0)$  $q(x_T)$

Generation process: $dx_t = -[\frac{1}{2}\beta(t)x_t - \frac{\beta(t)\epsilon_\theta(x_t, t)}{\sigma_t}]\,dt + \sqrt{\beta(t)}\,dW_t$

Noise prediction

What if we have want to generate
Image from input texts?

Image source: https://cvpr2022-tutorial-diffusion-models.github.io/

Adding Control Condition to Generation

$q(x_0)$  $q(x_T)$

Generation process: $dx_t = -[\frac{1}{2}\beta(t)x_t - \frac{\beta(t)\epsilon_\theta(x_t, \tau_\theta(c), t)}{\sigma_t}]\,dt + \sqrt{\beta(t)}\,dW_t$

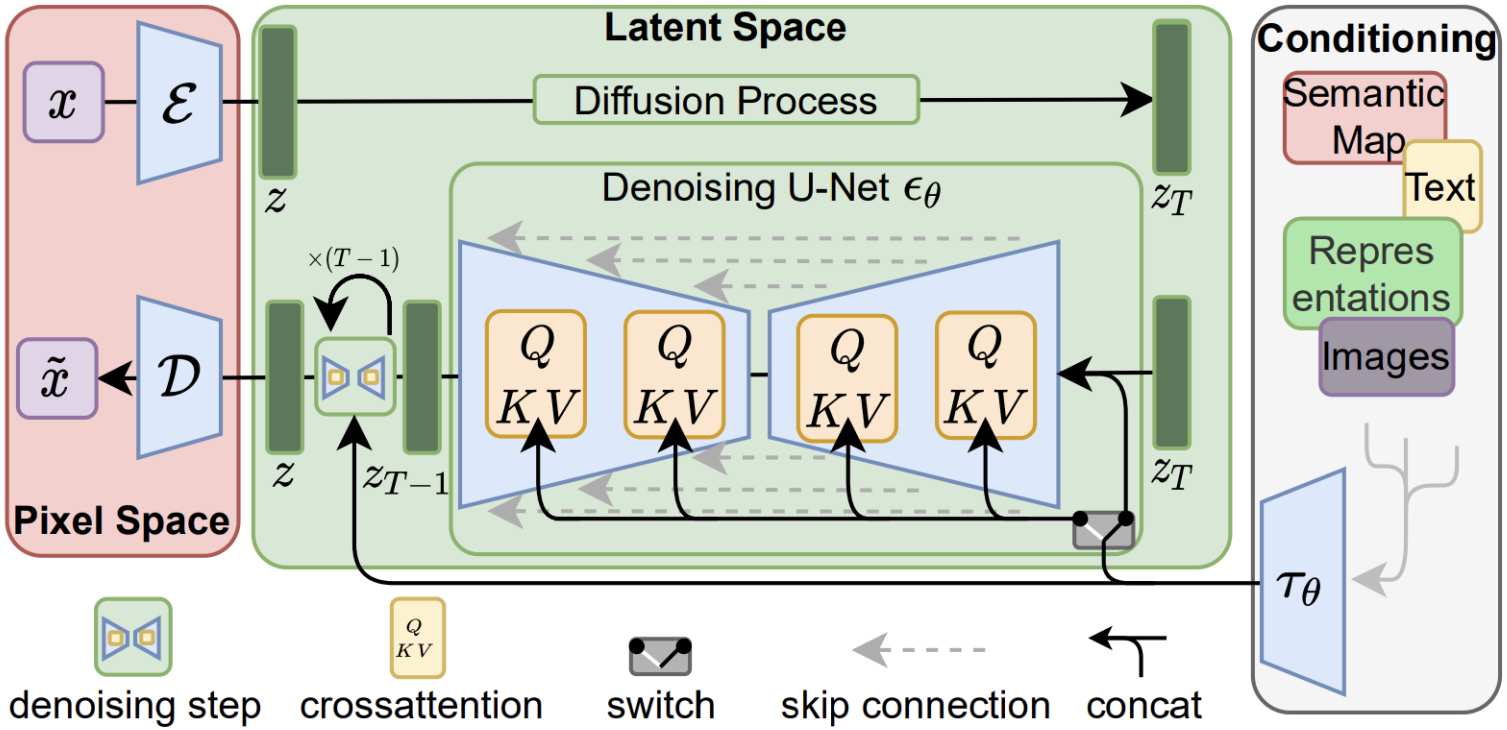
Noise prediction

$\tau_\theta(c)$ Embedding of extra condition

Use CLIP embedding for text input!

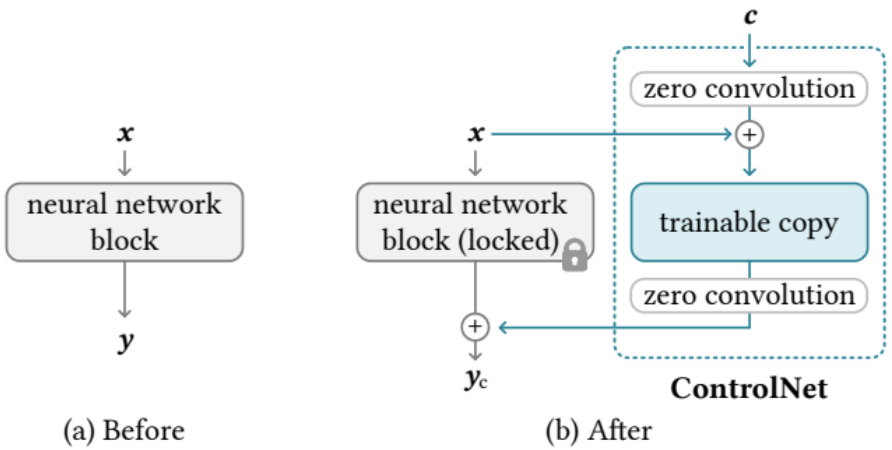
Image source: https://cvpr2022-tutorial-diffusion-models.github.io/

Latent Space Diffusion Models



Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", CVPR 2022

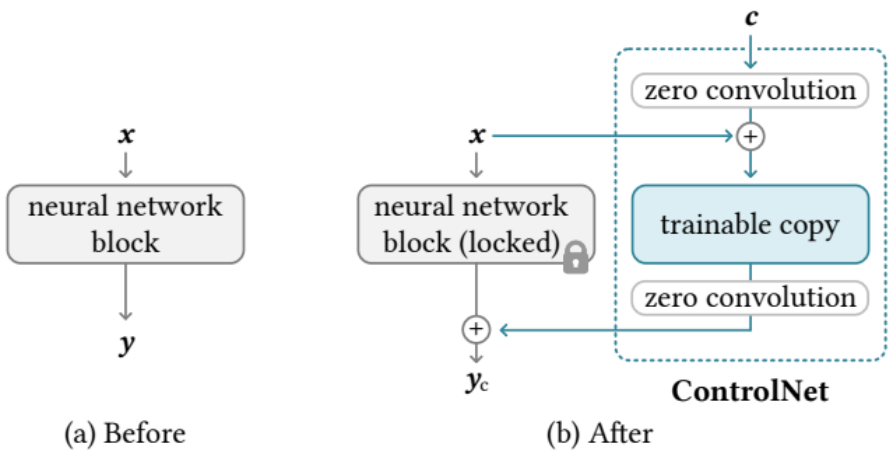
Control Net

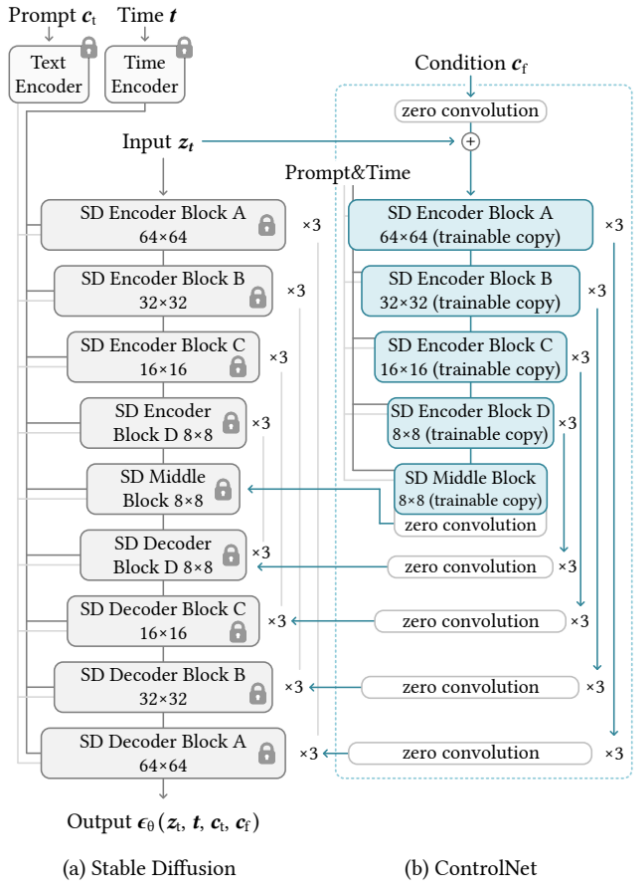


Initialized as original model
(due to zero weight convolution)

Lvmin Zhang, Anyi Rao, Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models.

Control Net applied to Stable Diffusion





Lvmin Zhang, Anyi Rao, Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models.

Control Net

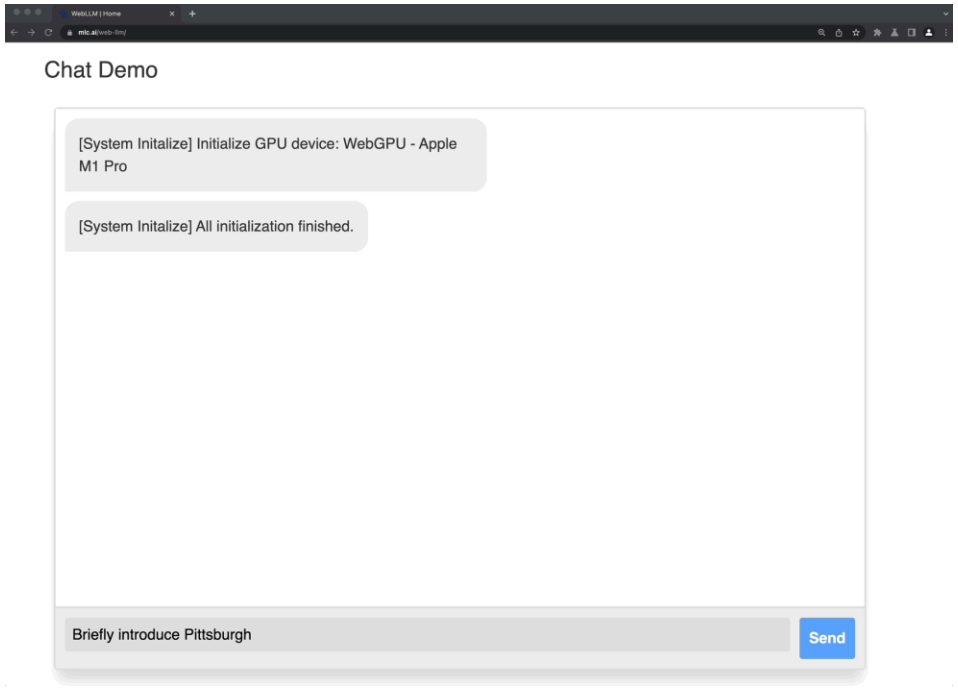
Lvmin Zhang, Anyi Rao, Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models.

Outline

Image generation models

Language models

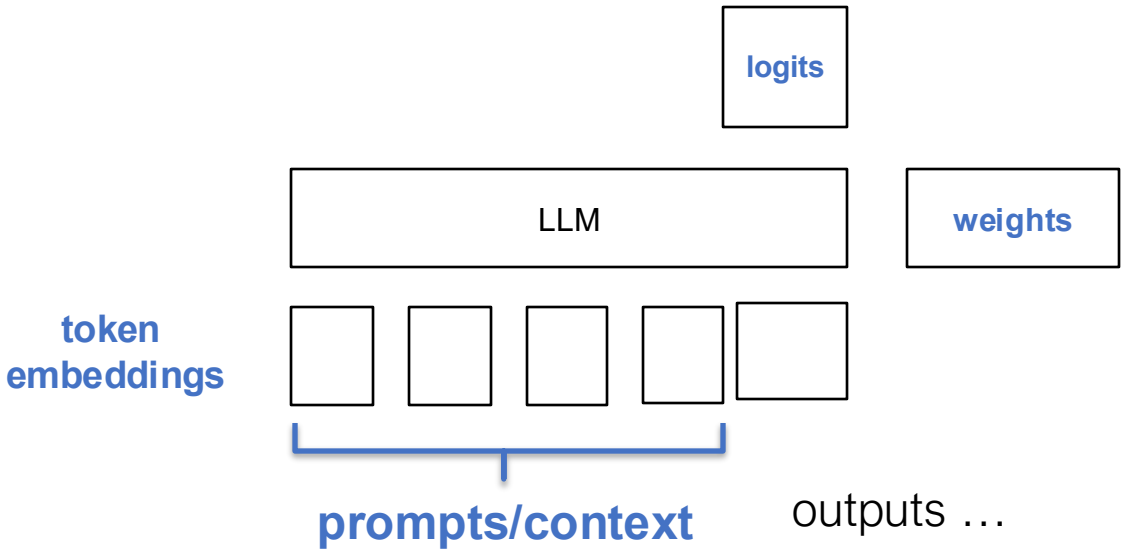
LLMs comes with a lot of parameters



Llama-70B would consume 320GB VRAM to just to store parameters in fp32

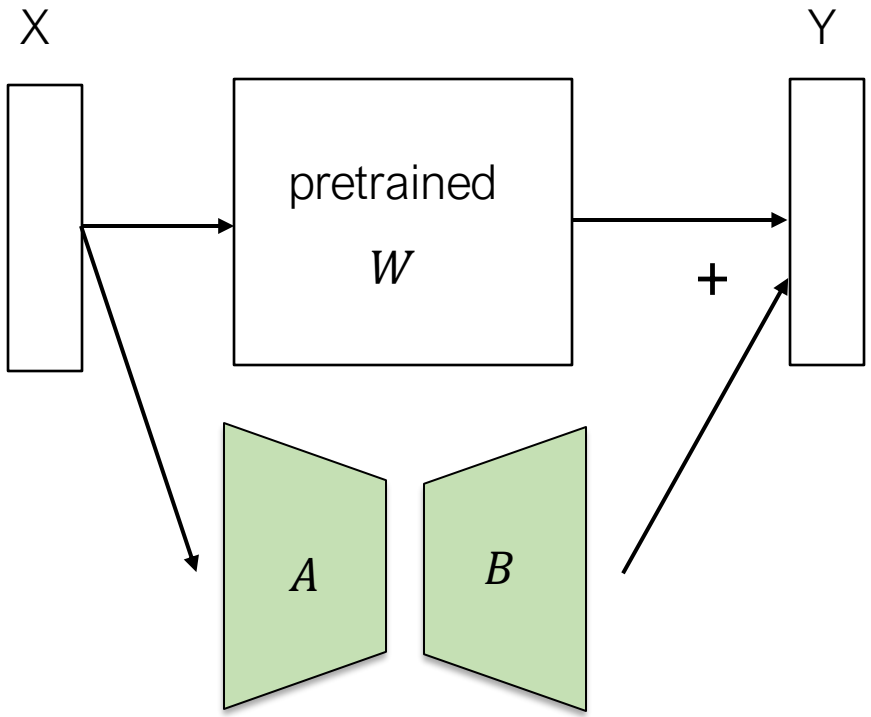
How do we ship domain specific model variants?

Language Models: Places that are Customizable



Each colored components can be customized/post-processed

LoRA: Low-Rank Adaptation



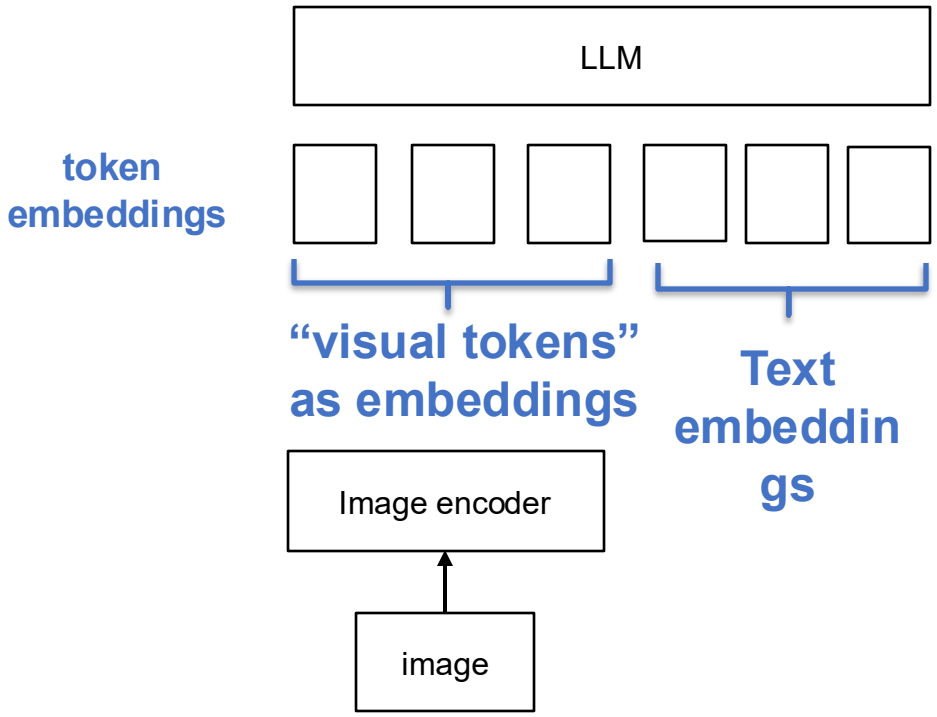
Fix pretrained weight

A and B are low rank matrices

Hu et.al. LoRA: Low-Rank Adaptation of Large Language Models

Also applies to diffusion models

Multi-Modality and Embedding Mapping



Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User: What is unusual about this image?

LLaVA: The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

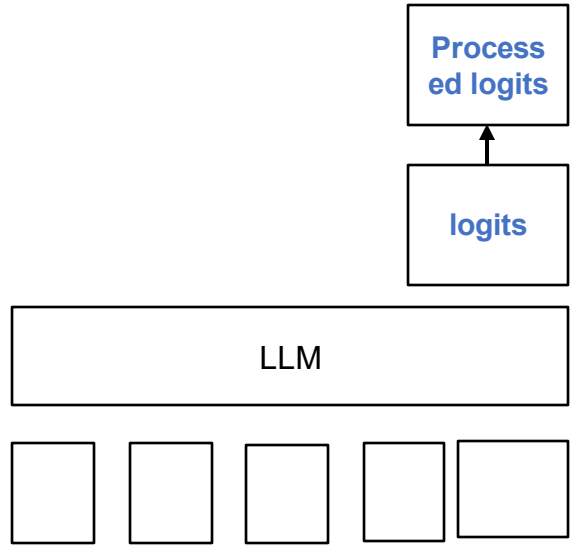
Liu and Li et.al. Visual Instruction Tuning: LLaVA (Large Language-and-Vision Assistant) built towards GPT-4V level capabilities

Zhu and Chen et.al. MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models

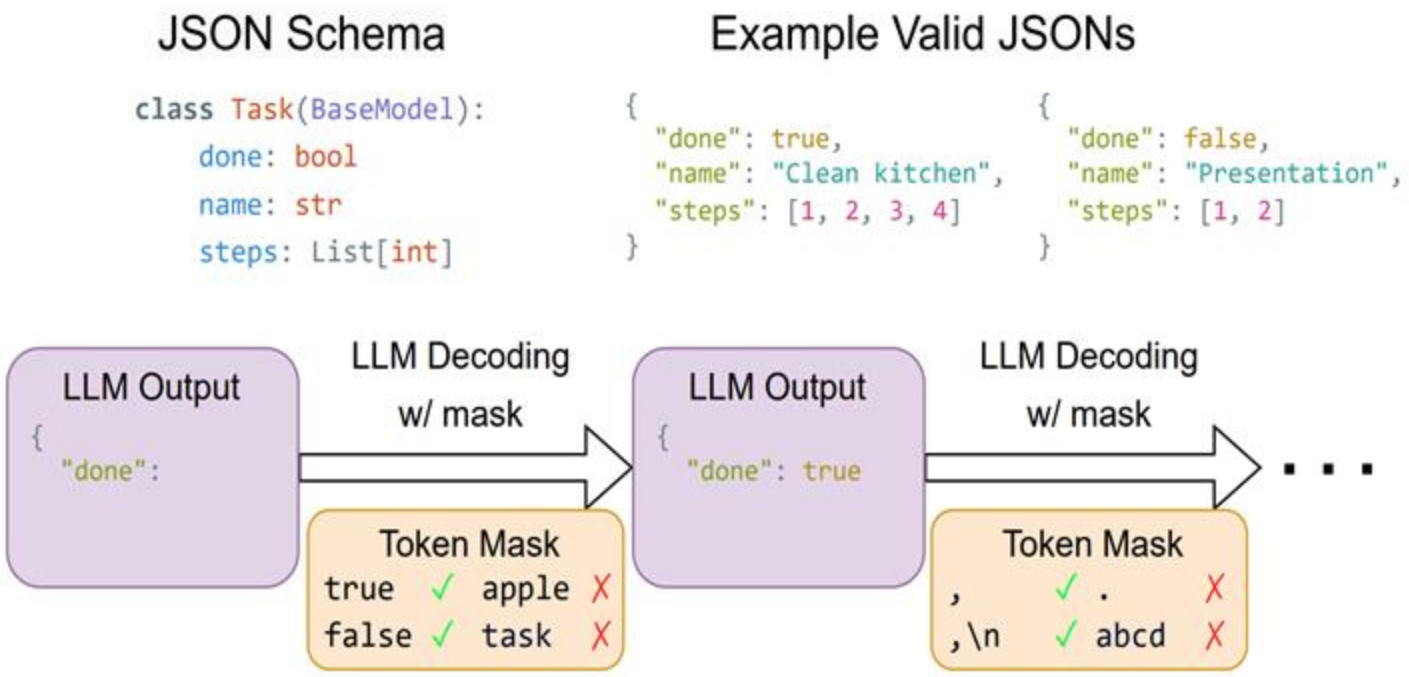
Postprocessing Logits

Adding penalties for repetitions

Forcing output to follow a grammar (e.g. JSON)



Generating Structured Outputs via Constraint Decoding



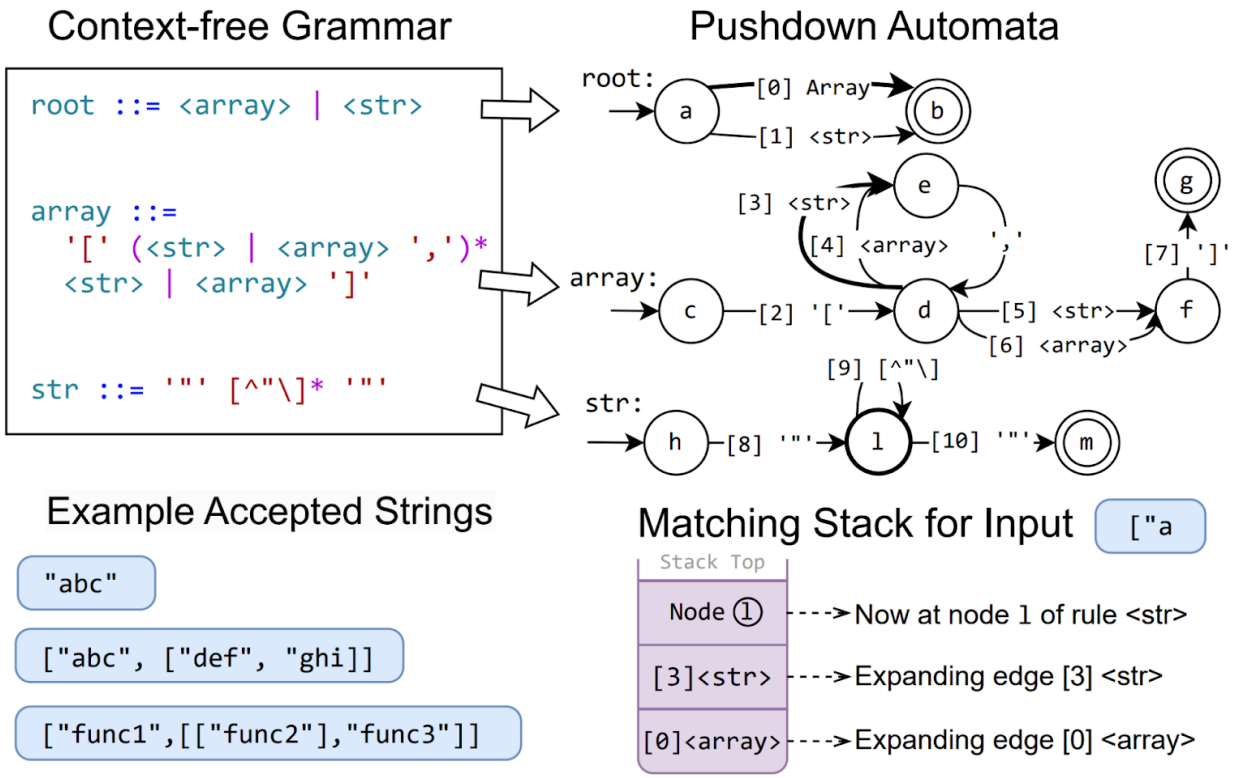
Challenges

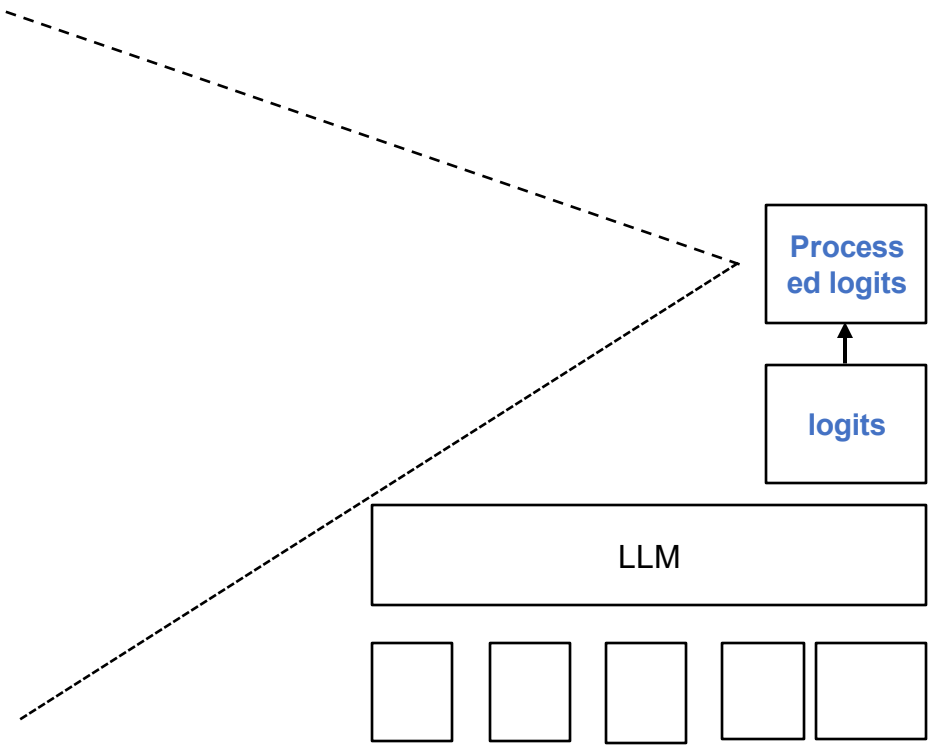
Large vocabulary size

Infinite possible grammar state for complex grammars

GPUs getting faster (and mask generation on CPU cannot keep up with them)

Context-free Grammar

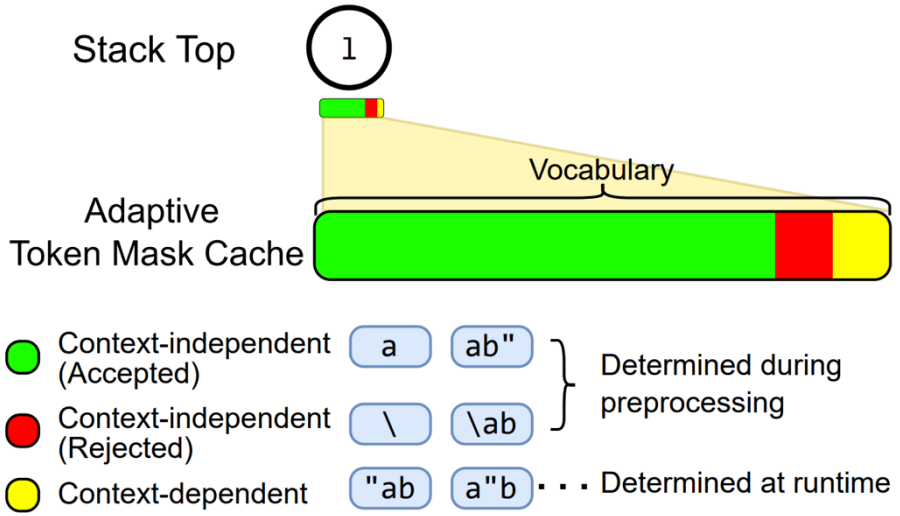




XGrammar: Efficient and Flexible Grammar Engine

Key insight

- **Context-independent tokens:** Most tokens (>99%) can be determined **ahead of time by only looking at top of the stack**
- **Context-dependent tokens** we still check outlier tokens at runtime to ensure full coverage.
- At runtime, we first retrieve the pre-computed token mask from the parsing state, then check the context-dependent tokens efficiently.



Outline

Image generation models

Language models