Golang ile Generative AI Pratikleri

Gokonf ‘24

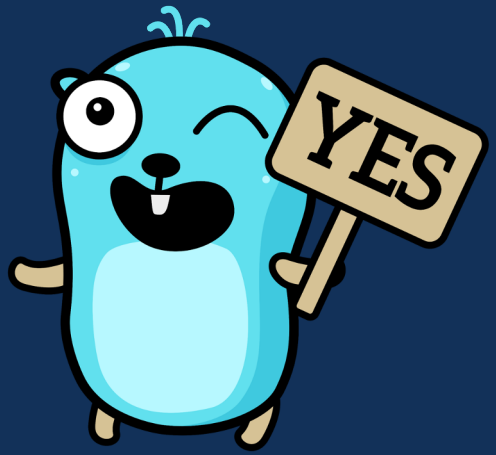
17 Şubat 2024
Yağmur YILDIZ

Akış

01 Neden Go & GenAI?

02 GenAI 101

03 Use Case: Txt2Img



Akış

01 Neden Go & GenAI?

02 GenAI 101

03 Use Case: Txt2Img

ÇIKARIM (INFERENCE)

- Eğitimli yapay zeka modelini canlı canlı kullanarak çıktı elde etmek
- Bir model 30-40GB
- Bir cycle 5-20 sn
- Aynı anda birden fazla istek

17 Şubat 2024
Yağmur YILDIZ

GO'NUN AVANTAJLARI

- Performance
- Efficiency
- Concurrency
- Scalability

\+ Bulut tabanlı çözümlere için mükemmel olması

Akış

01 Neden Go & GenAI?

02 GenAI 101

03 Use Case: Txt2Img

POPÜLER MODELLER

Kapalı Kaynak

ChatGPT

Midjourney

DALL·E

runway

IIElevenLabs

ANTHROP\C

Gemini

Açık Kaynak

MISTRAL AI_

BLOOM

Whisper

LLaMA by Meta

Cody

Meta AI AudioCraft

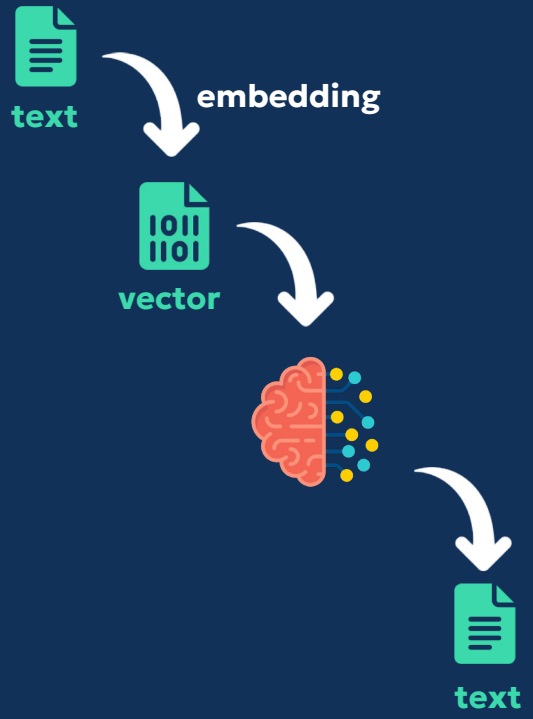
OpenAI Jukebox

17 Şubat 2024
Yağmur YILDIZ

BÜYÜK DİL MODELLERİ (LLMs)

Milyarlarca kelimelik büyük veriyle eğitilen derin öğrenme modelleridir.

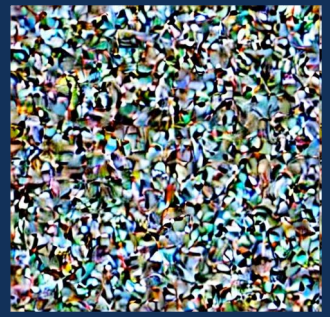
- **Transformers**
- **BERT (Bidirectional Encoder Representations Tr.)**
- **GPT (Gen Pre-Training Tr.)**
- **LaMDA (LM for Dialogue App.)**



17 Şubat 2024
Yağmur YILDIZ

GÖRSEL ÜRETME

Daha çok diffusion modelleri kullanılan bu alanda amaç inputun ögelerini anlamlandırarak bir gürültünün o ögeleri barındıran bir kompozisyona dönüşmesidir.



Latent Noise

Generated Image

- Text2Image
- Image2Image
- Audio2Image
- Video2Image

17 Şubat 2024
Yağmur YILDIZ

STABLE DIFFUSION MODELLERİ

Version	Release Date	Resolution	Parameters	Prompts Technology	Strengths	Weaknesses
1.4	08.2022	512x512	860M	CLIP	beginner-friendly, a little more artistic driven	long prompts, lower resolution
1.5	10. 2022	512x512	860M	CLIP	beginner-friendly, stronger portrait generation	long prompts, lower resolution
2.0	11.2022	768x768	-	OpenCLIP	shorter prompts, richer colors	aggressive NSFW filtering
2.1	12.2022	768x768	-	OpenCLIP	shorter prompts, richer colors	more “censored”, celeb filtered
XL 1.0	01.2023	1024x1024	3.5B	OpenCLIP & CLIP	shorter prompts, high resolution	requires GPU

SAMPLER, SCHEDULER, SEED

diversity, quality, speed, convergence

Sampler

Olasılık uzayında modelin nasıl çalışacağını belirler. Euler, DDIM, DDPM

Scheduler

Modelin her sampleda nasıl yakınsayacağını belirler. Linear, PNDM, Karras

Seed

Üretim sürecine başlarken kullanılan random değerdir.

17 Şubat 2024
Yağmur YILDIZ

SAMPLING STEP, CFG SCALE, SIZE

quality, creativity, computing power

Sampling Step

Samplerın kaç kez çalışacağını belirler. 30-75 arası tercih edilir.

CFG Scale

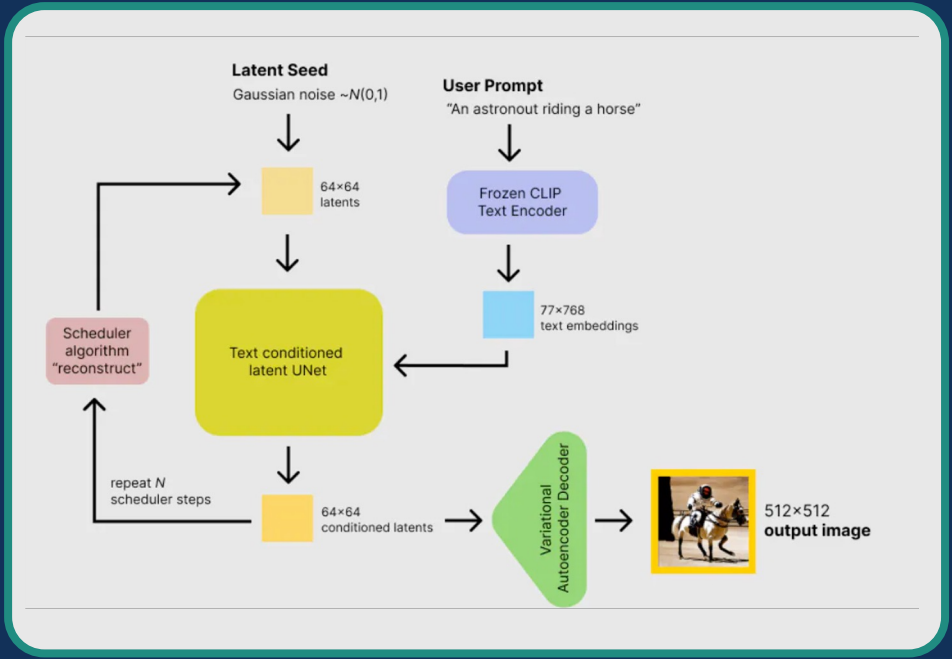
Promptun generation üzerindeki etkisini belirler. 7-10 arası tercih edilir.

Size

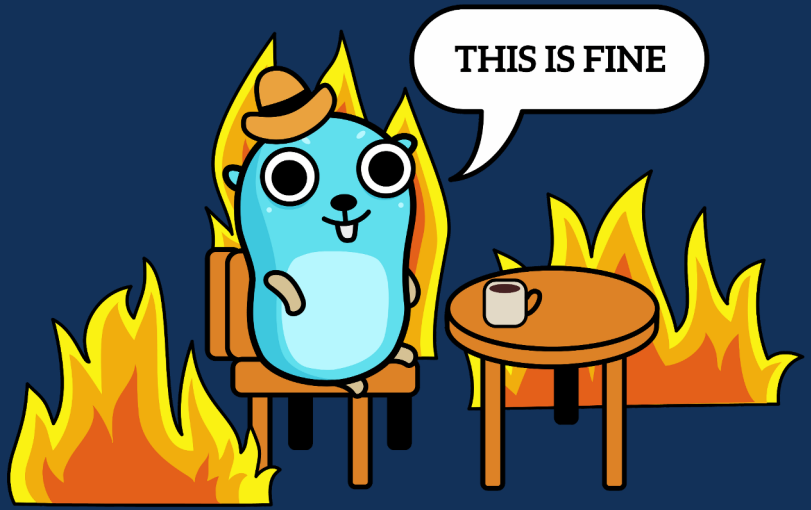
Üretilecek resmin WxH boyutudur. Her modelin baz bir boyutu vardır.

17 Şubat 2024
Yağmur YILDIZ

STABLE DIFFUSION ALGORITHM



17 Şubat 2024
Yağmur YILDIZ



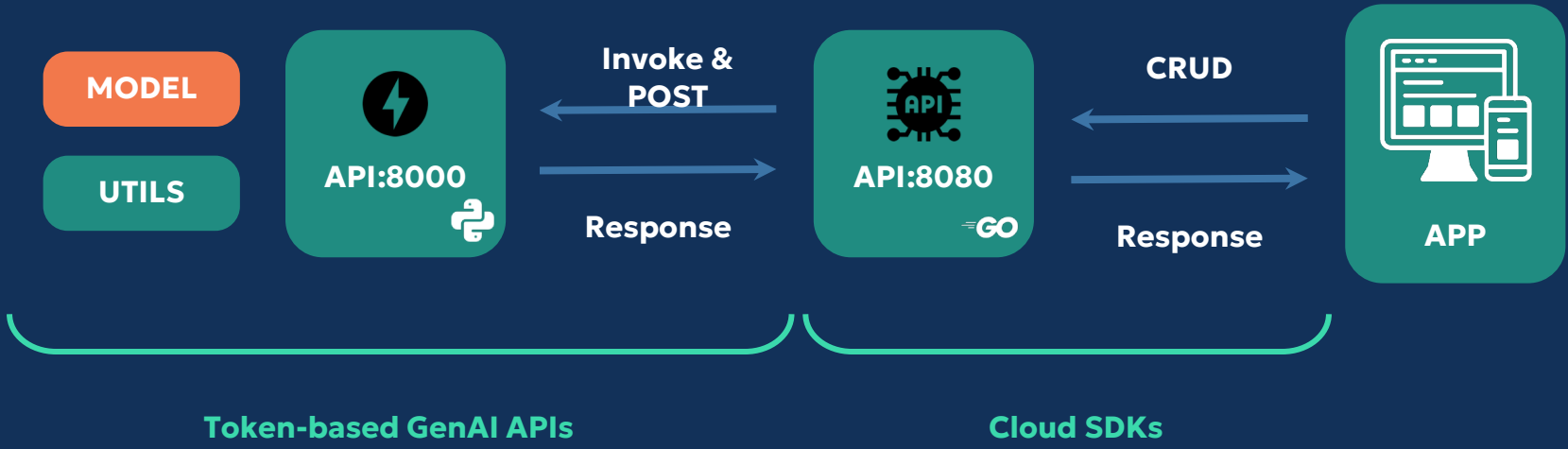
Akış

01 Neden Go & GenAI?

02 GenAI 101

03 **Use Case: Txt2Img**

TEMEL MİMARİ



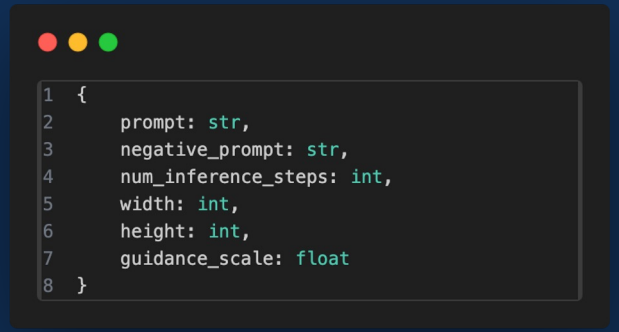
GENERAL IMAGE REQUEST STRUCT

JSON

Bir prompt ile istediğimiz ölçülerde çıktı alabileceğimiz basit bir request yapısı

Constants

- Model
- Sampler
- Scheduler
- Output Format



```
{
    prompt: str,
    negative_prompt: str,
    num_inference_steps: int,
    width: int,
    height: int,
    guidance_scale: float
}
```

17 Şubat 2024
Yağmur YILDIZ

MODEL API

Modeli barındıran repo ile konuşan ve modele requestlerin doğru atılmasını sağlayan katmandır.

- **Python ile yapılması gereken işler**
- **Device settings**
- **Sampler, Scheduler, Seed**

```python

app = FastAPI()

@app.post("/txt2img/")
async def generate_image(request: ImageRequest):
    try:
        device = get_device_info()["device"]
        dtype = get_device_info()["dtype"]
        pipeline = create_pipeline("Realistic_Vision_V6.0_B1_noVAE", device, dtype)

        generator = torch.Generator(device)
        generator = torch.Generator(device=device)
        seed = generator.seed()

        image = pipeline(prompt=request.prompt,
                         negative_prompt=request.negative_prompt,
                         generator = generator.manual_seed(seed),
                         num_inference_steps=request.num_inference_steps,
                         width = request.width,
                         height = request.height,
                         guidance_scale = request.guidance_scale

                        ).images[0]

        img_byte_arr = BytesIO()
        image.save(img_byte_arr, format='PNG')
        img_byte_arr.seek(0)
        return StreamingResponse(img_byte_arr, media_type="image/png")
    except Exception as e:
        raise HTTPException(status_code=500, detail=str(e))


```

17 Şubat 2024
Yağmur YILDIZ

CLIENT API - I

FAST API'ye gönderilecek request'e user girdisinin olmadığı senaryolar

- Styled Avatar Generation

```go
type ImageRequest struct
    Prompt            string  `json:"prompt"`
    NegativePrompt    string  `json:"negative_prompt"`
    NumInferenceSteps int     `json:"num_inference_steps"`
    Width             int     `json:"width"`
    Height            int     `json:"height"`
    GuidanceScale     float32 `json:"guidance_scale"`
}
```

```go
func main() {

    requestData := ImageRequest{
        Prompt:
"A potrait of a man , comic styled, detailed hair, smiling face",
        NegativePrompt:    "nsfw, bad eyes, bad teeth",
        NumInferenceSteps: 40,
        Width:             512,
        Height:            512,
        GuidanceScale:     7.5,
    }

    jsonData, err := json.Marshal(requestData)

    resp, err := http.Post("http://localhost:8000/txt2img/",
"application/json", bytes.NewBuffer(jsonData))

    defer resp.Body.Close()

    outFile, err := os.Create("output.png")

    _, err = io.Copy(outFile, resp.Body)

    log.Println("Image saved as output.png")
}
```

CLIENT API - II

FAST API'ye gönderilecek request'e user girdisi olan senaryolar

- **Stock Image Generation**

```go
func proxyHandler(w http.ResponseWriter, r *http.Request) {

    client := &http.Client{}
    req, err := http.NewRequest("POST",
"http://localhost:8000/txt2img/", r.Body)

    req.Header.Set("Content-Type", "application/json")

    resp, err := client.Do(req)

    io.Copy(w, resp.Body)
}
```

```go
func main() {
    http.HandleFunc("/", proxyHandler)
    log.Println("Starting server on :8080")
    log.Fatal(http.ListenAndServe(":8080", nil))
}
```

17 Şubat 2024
Yağmur YILDIZ

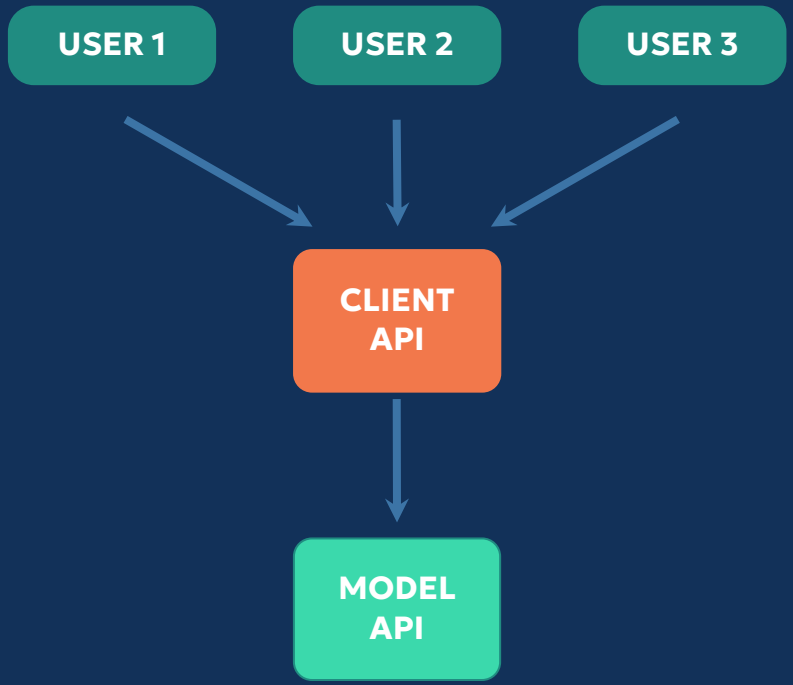
CLIENT API - III

User tarafından eş zamanlı birden fazla request geldiği senaryolar

***Goroutines**

```go
var requestQueue = make(chan []byte, 100)
```

```go
func main() {
    go processRequests()
    http.HandleFunc("/", proxyHandler)
    log.Println("Starting server on :8080")
    log.Fatal(http.ListenAndServe(":8080", nil))
}
```



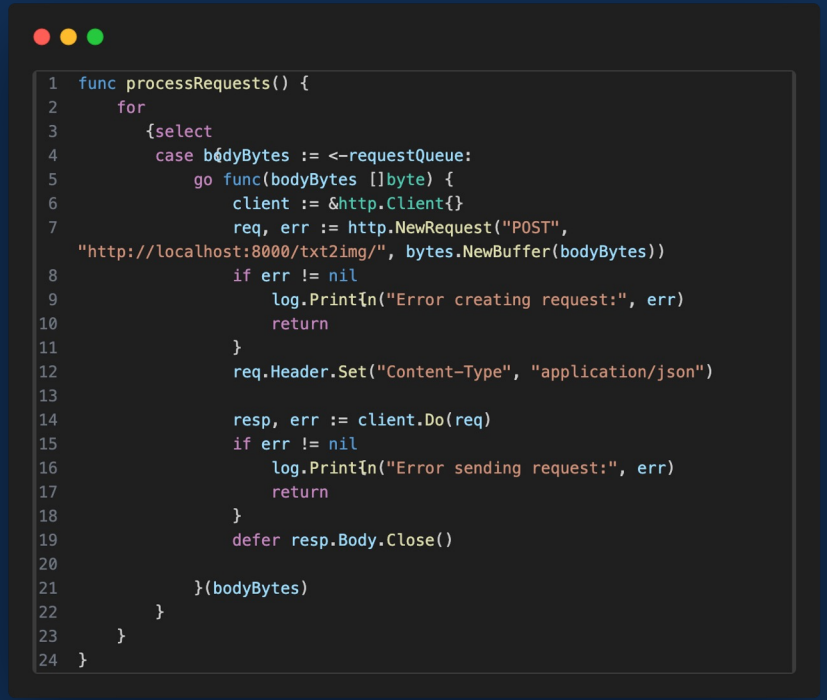
17 Şubat 2024
Yağmur YILDIZ

- Request Body'yi okuyup byte[ ] kaydeder
- Eğer queue dolmadıysa değerleri Queue'ya ekler, dolduysa http error gönderir.

```go
func proxyHandler(w http.ResponseWriter, r *http.Request) {

    bodyBytes, err := ioutil.ReadAll(r.Body)

    defer r.Body.Close()

    select
    case requestQueue <- bodyBytes:
        log.Println("Request queued")
        fmt.Fprintln(w, "Request queued")
    default:
        http.Error(w, "Request queue is full", http.
StatusServiceUnavailable)
    }
}
```

17 Şubat 2024
Yağmur YILDIZ

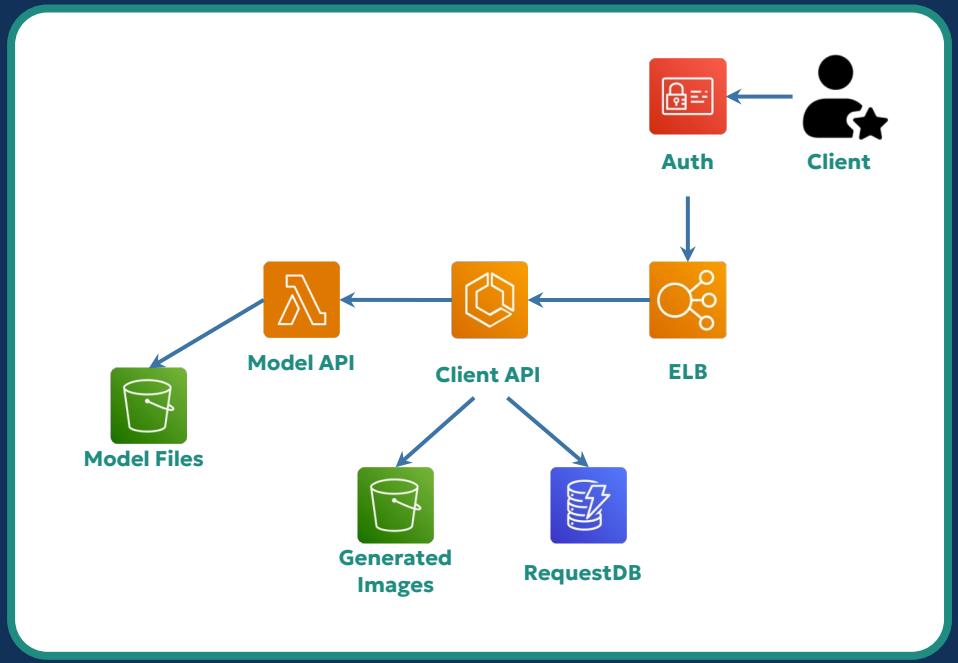
- Sonsuz döngü
- Queue’dan request bekler
- Request geldiği zaman bir goroutine başlatır
- Her request’i ayrı bir goroutine’de işler



```go
func processRequests() {
    for
       {select
        case bodyBytes := <-requestQueue:
            go func(bodyBytes []byte) {
                client := &http.Client{}
                req, err := http.NewRequest("POST",
"http://localhost:8000/txt2img/", bytes.NewBuffer(bodyBytes))
                if err != nil
                    log.Println("Error creating request:", err)
                    return
                }
                req.Header.Set("Content-Type", "application/json")

                resp, err := client.Do(req)
                if err != nil
                    log.Println("Error sending request:", err)
                    return
                }
                defer resp.Body.Close()

            }(bodyBytes)
        }
    }
}
```

17 Şubat 2024
Yağmur YILDIZ

ÖRNEK SERVERLESS AWS MİMARİSİ



17 Şubat 2024
Yağmur YILDIZ

Gokonf 2024
17 Şubat
Yağmur YILDIZ

TEŞEKKÜRLER!


/in/ygryildiz


@yagmurudurdurabilirmisin


@yagmurxyildiz


@yagmurx


yagmur.cc