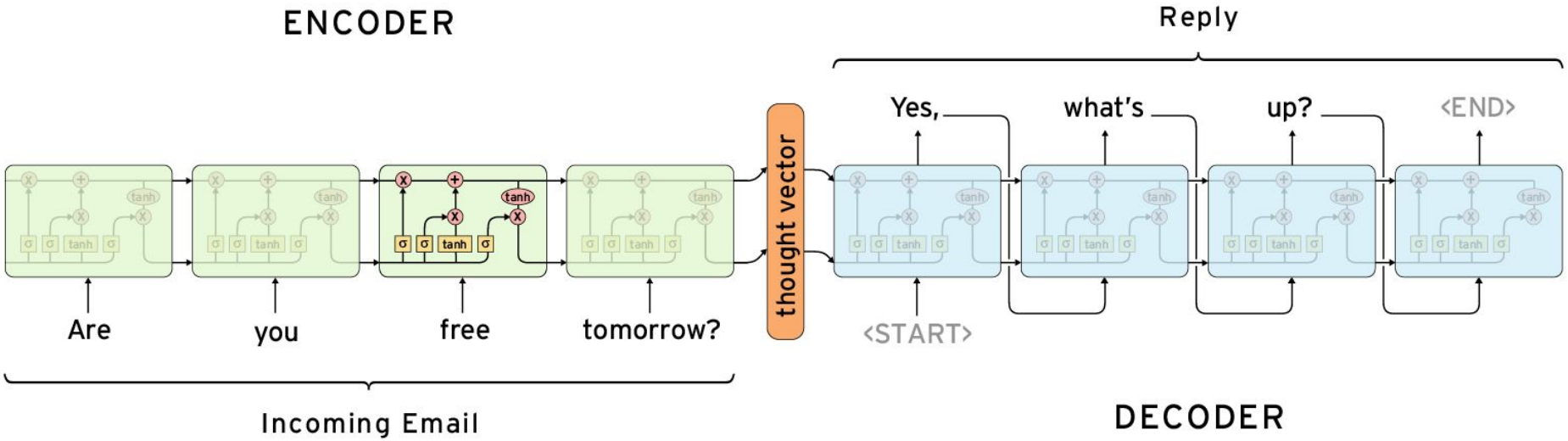
Как использовать мультимодальные трансформеры в исследованиях и не страдать?

Обо мне

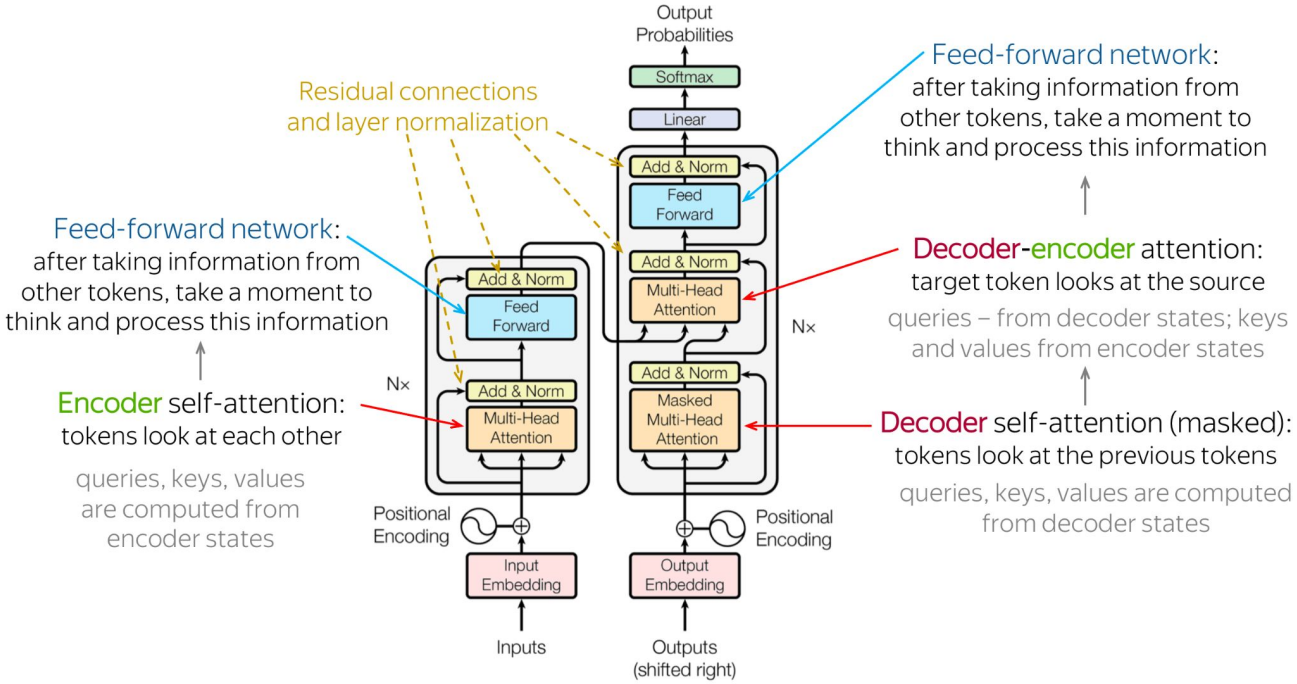
Исследую и применяю гигантские мультимодальные трансформеры, пишу статьи, имплементирую архитектуры

Коротко о трансформерах и почему их стоит полюбить вам

Помните этого парня?
(seq2seq cnn-rnn)



Теперь(с 2018) он выглядит так



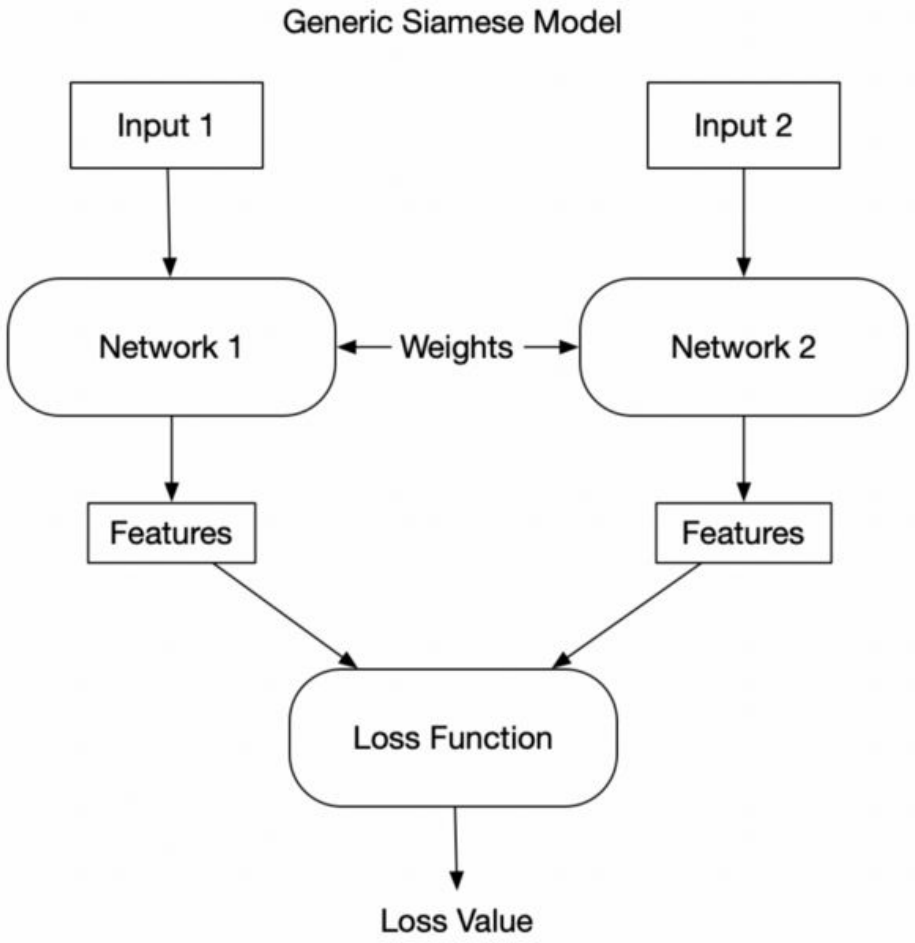
https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

- Документация
- Курсы
- Запуск в облаке(бесплатно на небольших масштабах)
- Почти все модели работают “из коробки”

Мультимодальность

Сиамские сети

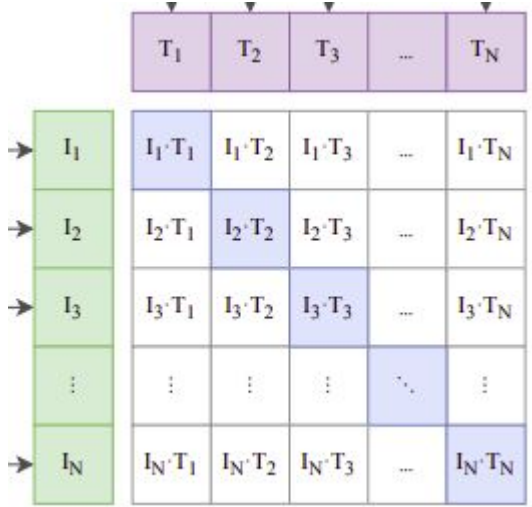
- Учим две сети с одним лоссом
- Нам не очень важно что у нас в разных энкодерах



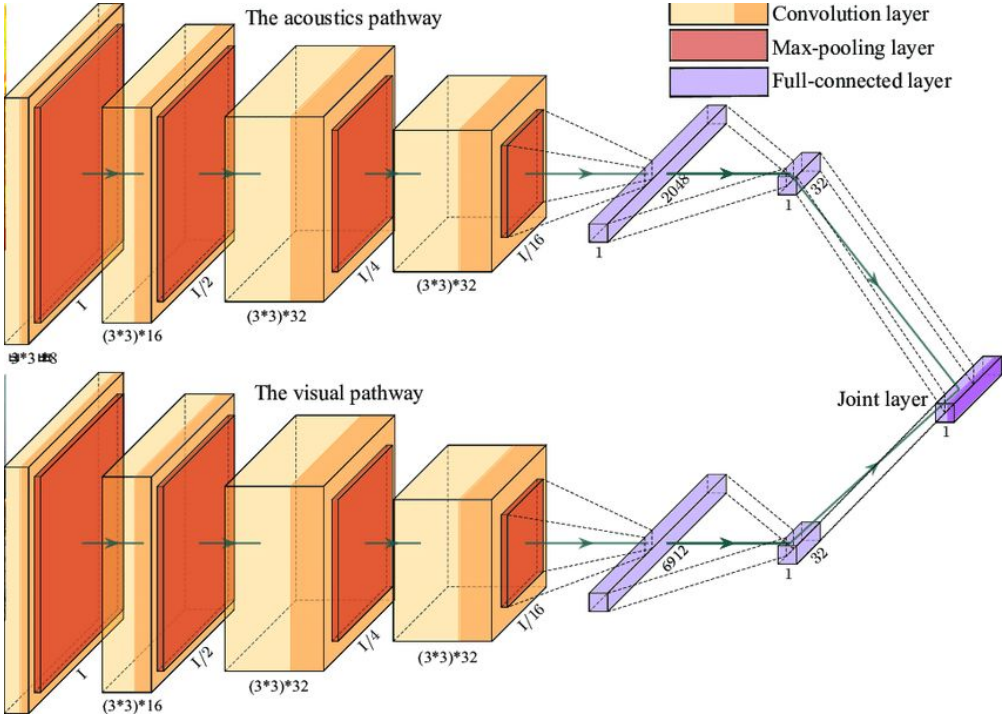
Contrastive loss

- Максимизируем главную диагональ, нужен большой batch size для хороших моделей

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} {}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$



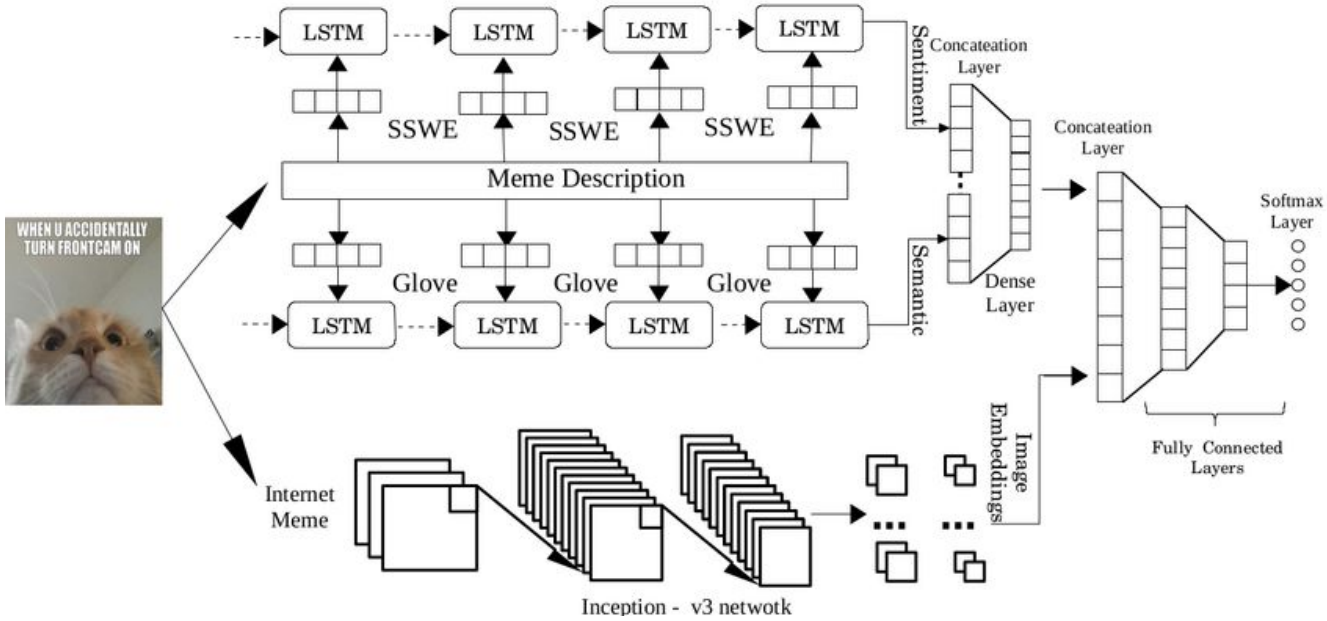
Да кто такая эта ваша мультимодальность?



Ключевые отличия

- Используем на каждую модальность свой энкодер
- Соединяем их(DSSM like) / кормим в декодер (Seq2Seq)

Немного про решаемые задачи

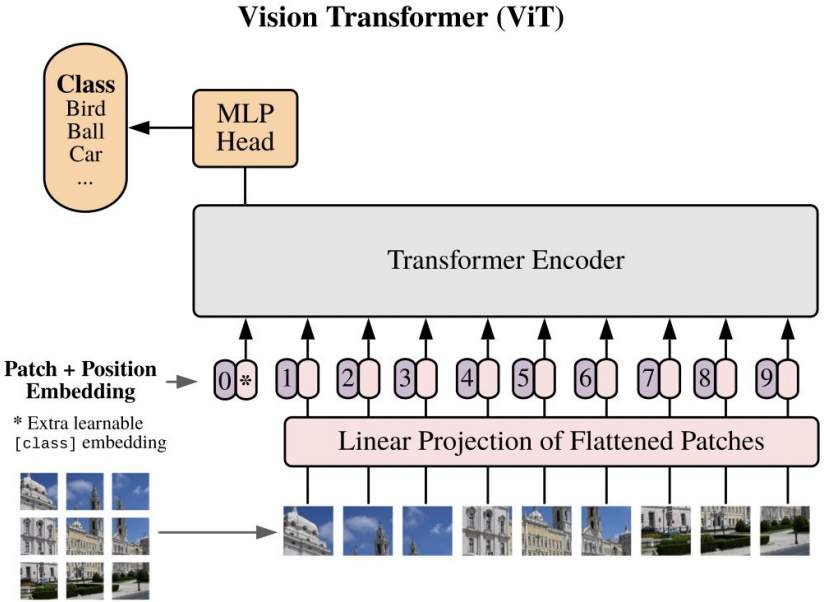


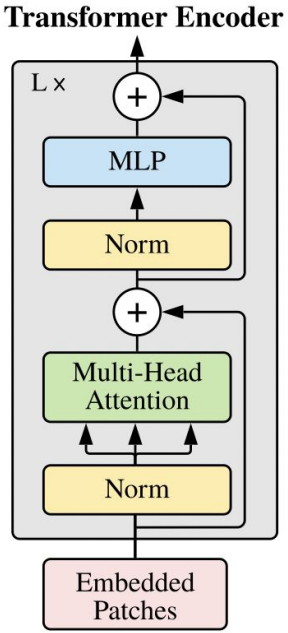
https://www.researchgate.net/figure/Architecture-of-Multimodal-Neural-Network-II_fig4_344603358

Типы входных данных

- На самом деле речь может идти и про картинки и про тексты и про видео(!)

Visual transformer





Идея

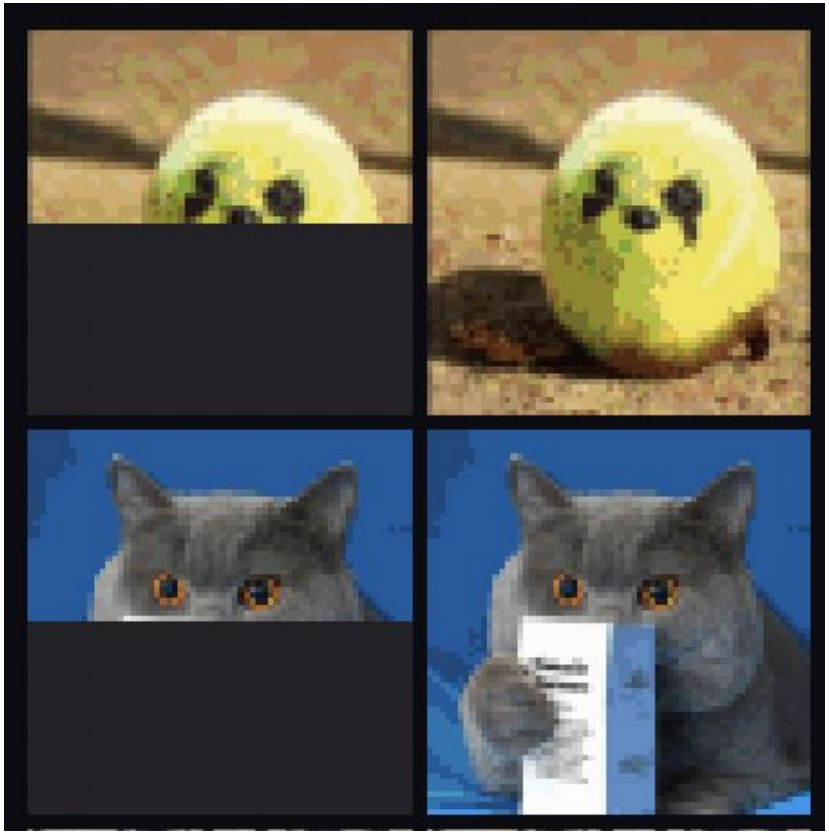
- Давайте разрежем изображение на визуальные токены (256*256->1024)
- Обучим на токенах классификатор изображений

Token based

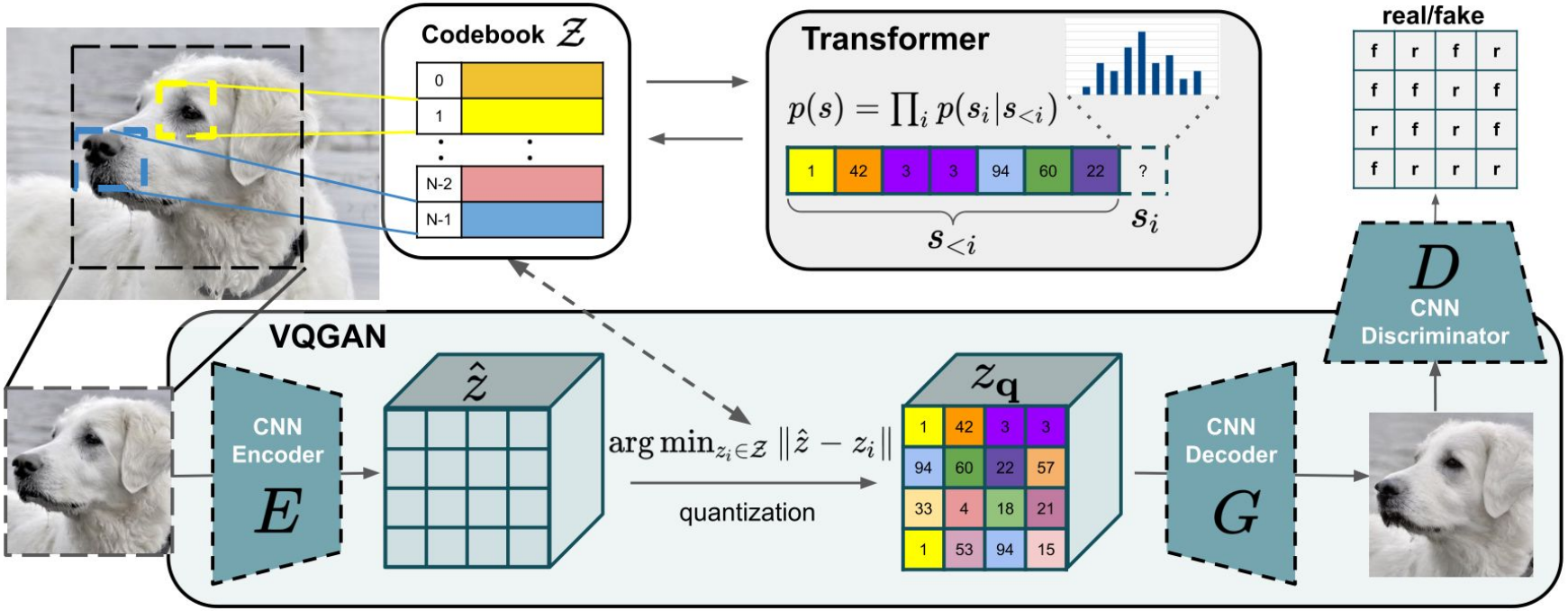
- Image GPT
- DALLE 1 like
- Rudolph
- BLIP
- Сила и слабость image tokens

Image Gpt

Предсказываем следующий пиксель картинки, каждый пиксель - токен.



Vqgan



https://github.com/CompVis/taming-transformers

DALLE 1

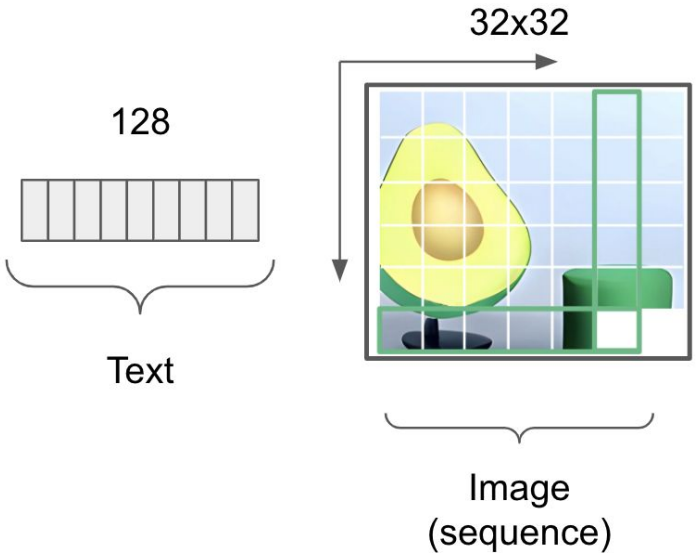
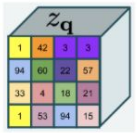


Image Encoder/Decoder Sber-VQ-GAN

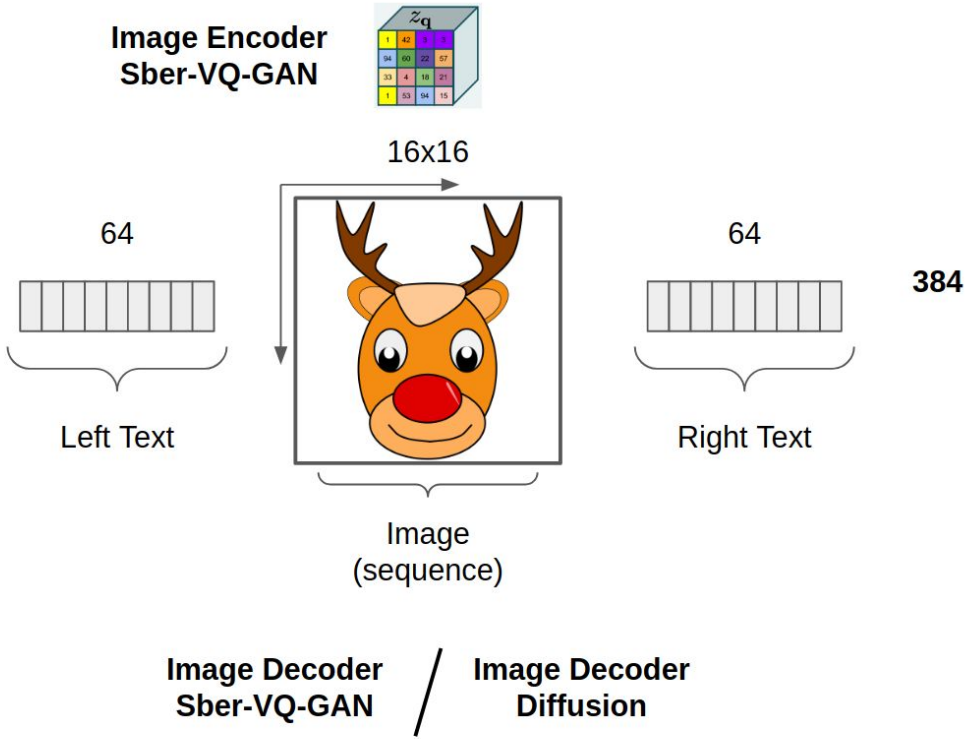


Учим GPT3 продолжать последовательность текстовых токенов, картиночными

https://github.com/ai-forever/ru-dalle

Rudolph

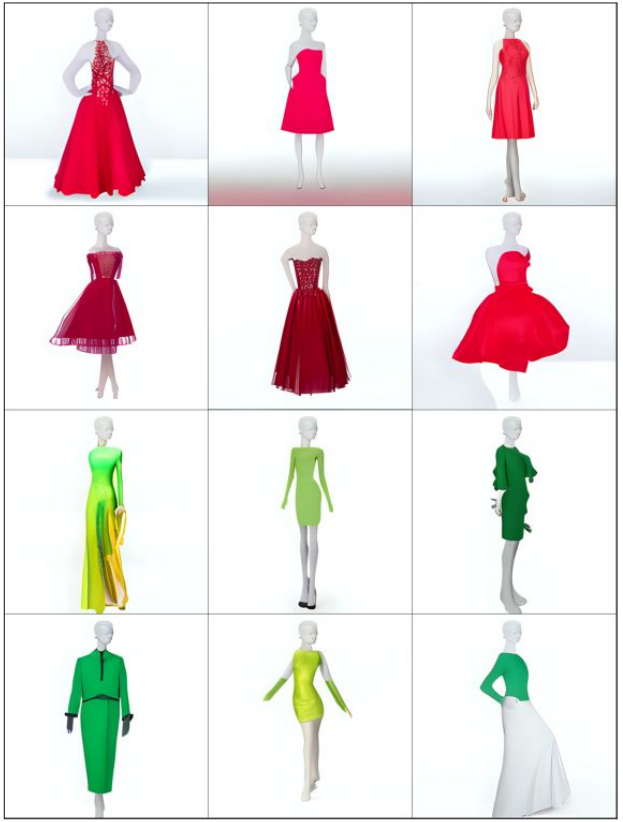
Текстовые токены слева и справа, мы можем предсказывать ответы на вопросы по картинке, генерировать картинку



https://github.com/ai-forever/ru-dolph

Сила и слабость токенов в картинках

Так как мы оперируем картинками на уровне токенов, то мы можем дать текст, начало картинки и по текстовому запросу продолжить картинку до конца



https://github.com/ai-forever/ru-dalle/blob/master/jupyters/ruDALLE-image-prompts-dress-mannequins-V100.ipynb

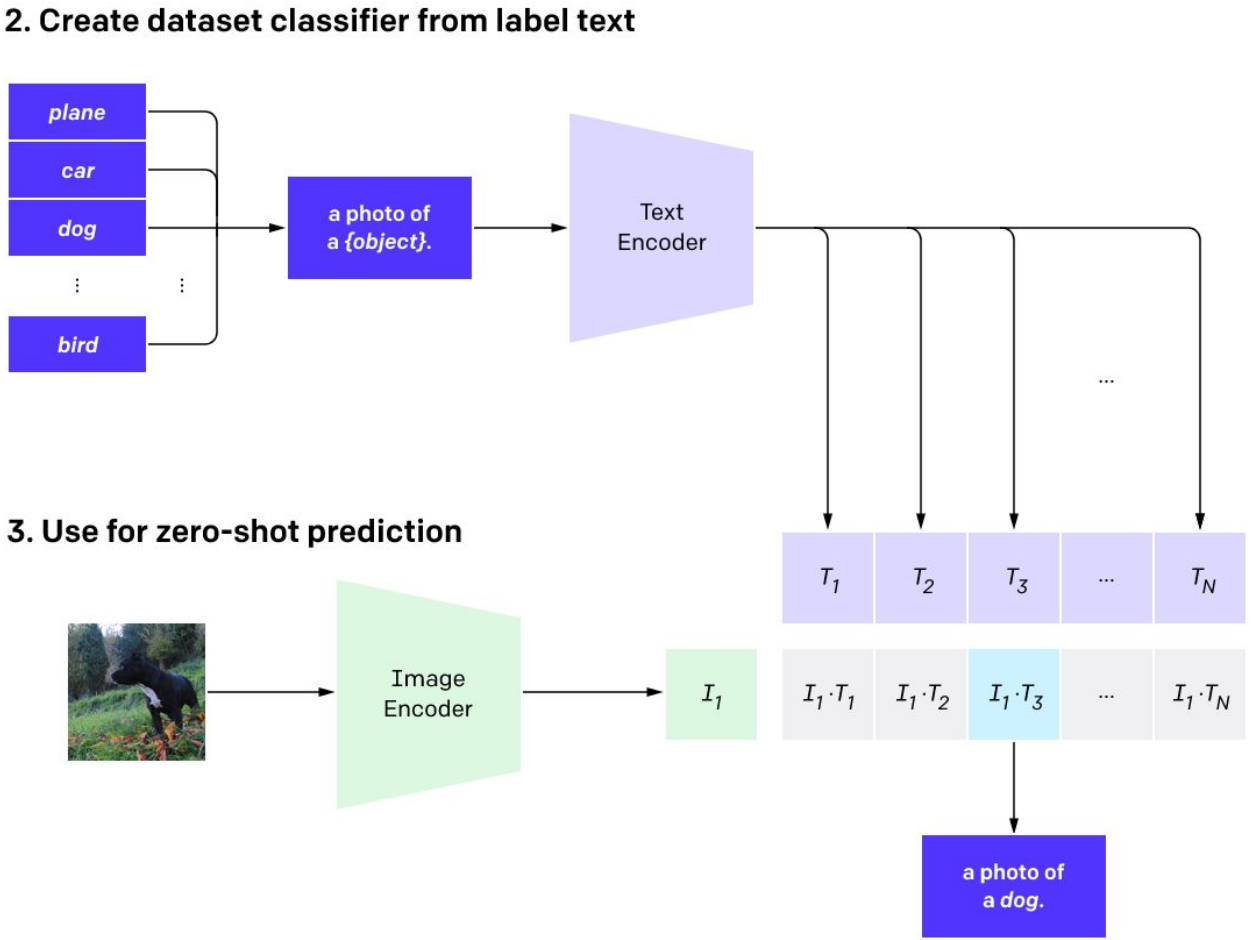
Multi Encoder

- CLIP
- KM BART
- FUSION BRAIN
- GATO
- DALLE2, ImageGen, …

CLIP

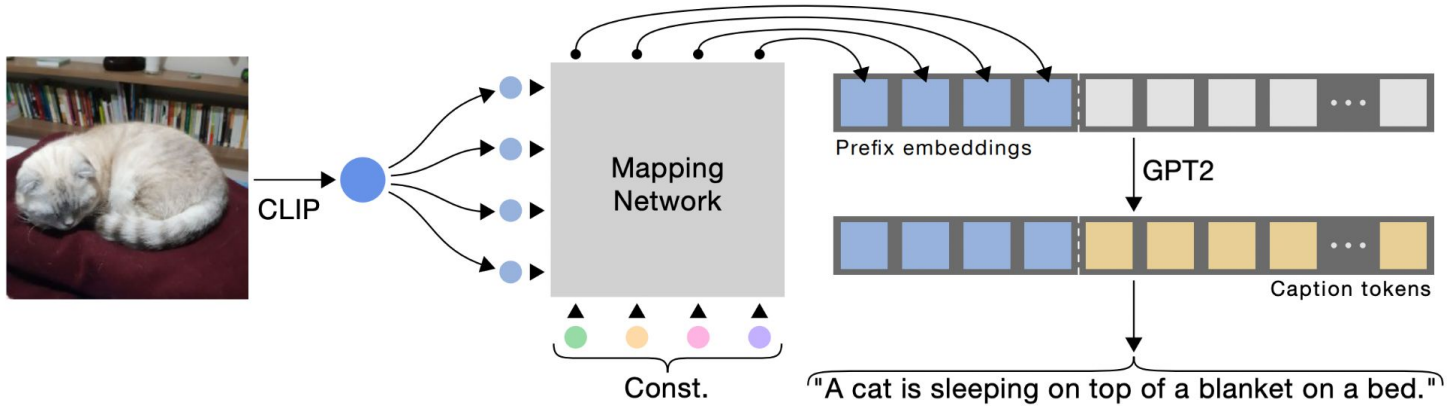
Основные особенности

- Используем BERT для текстового энкодера
- Используем ResNet50 / VIT16 в качестве картиночного энкодера



https://openai.com/blog/clip/

CLIP позволяет эффективно кодировать картинку в пространство текста

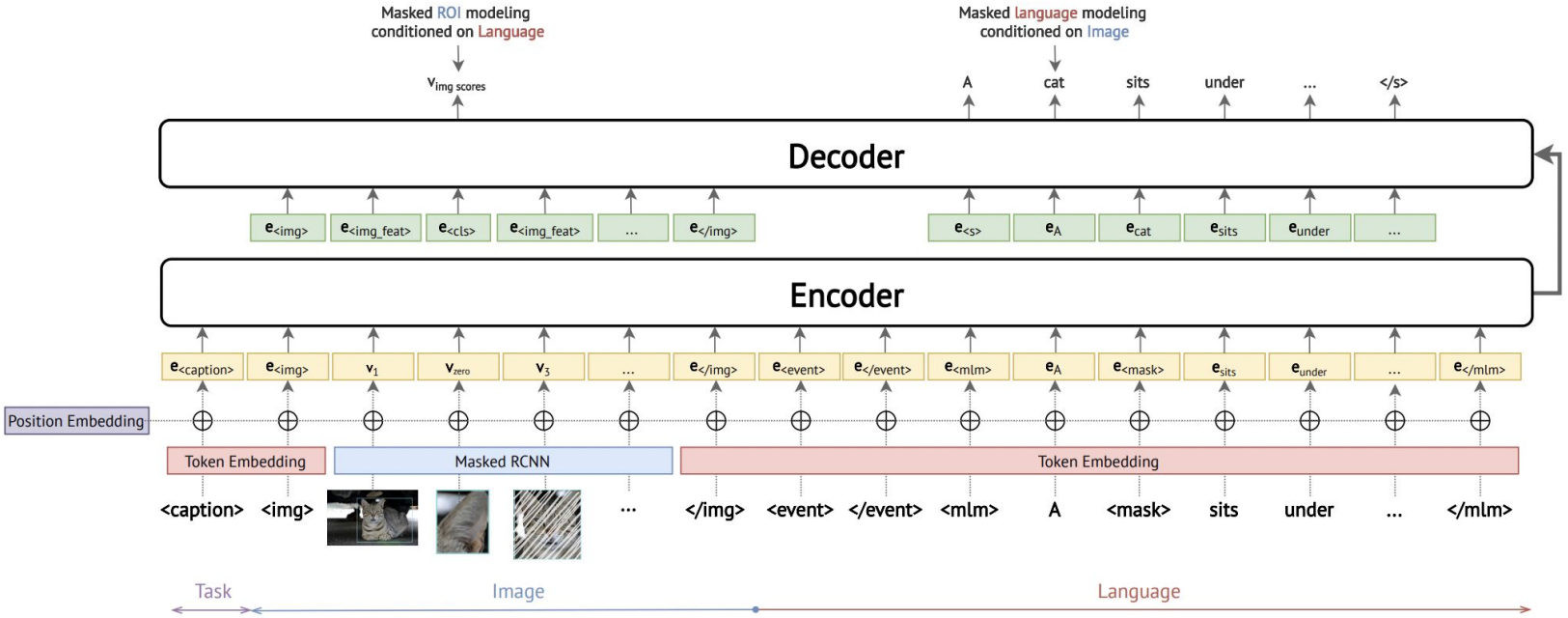


Задачи

- Если добавить декодер(GPT), то мы получим архитектуру для решения image2text, VQA,

https://github.com/rmokady/CLIP_prefix_caption
https://github.com/AlexWortega/ruImageCaptioning

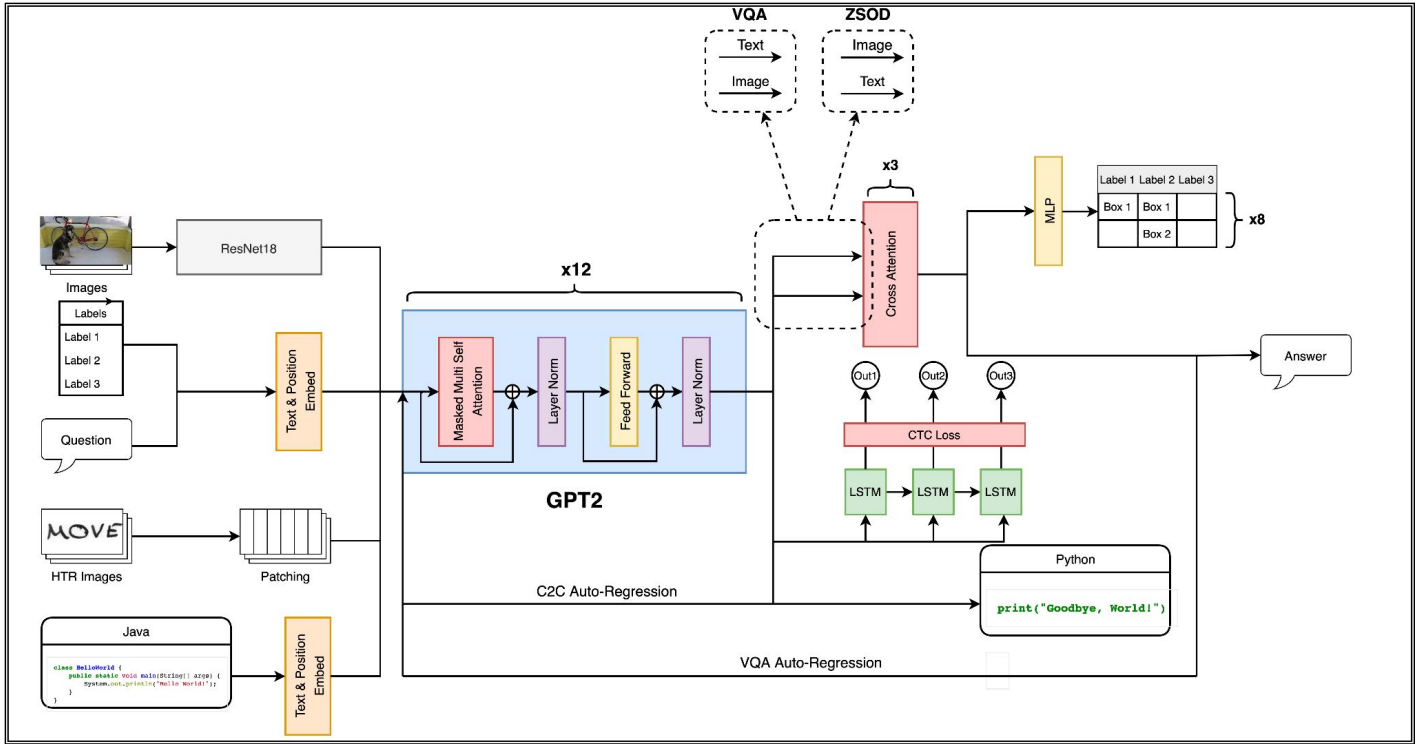
KM BART



https://arxiv.org/pdf/2101.00419.pdf

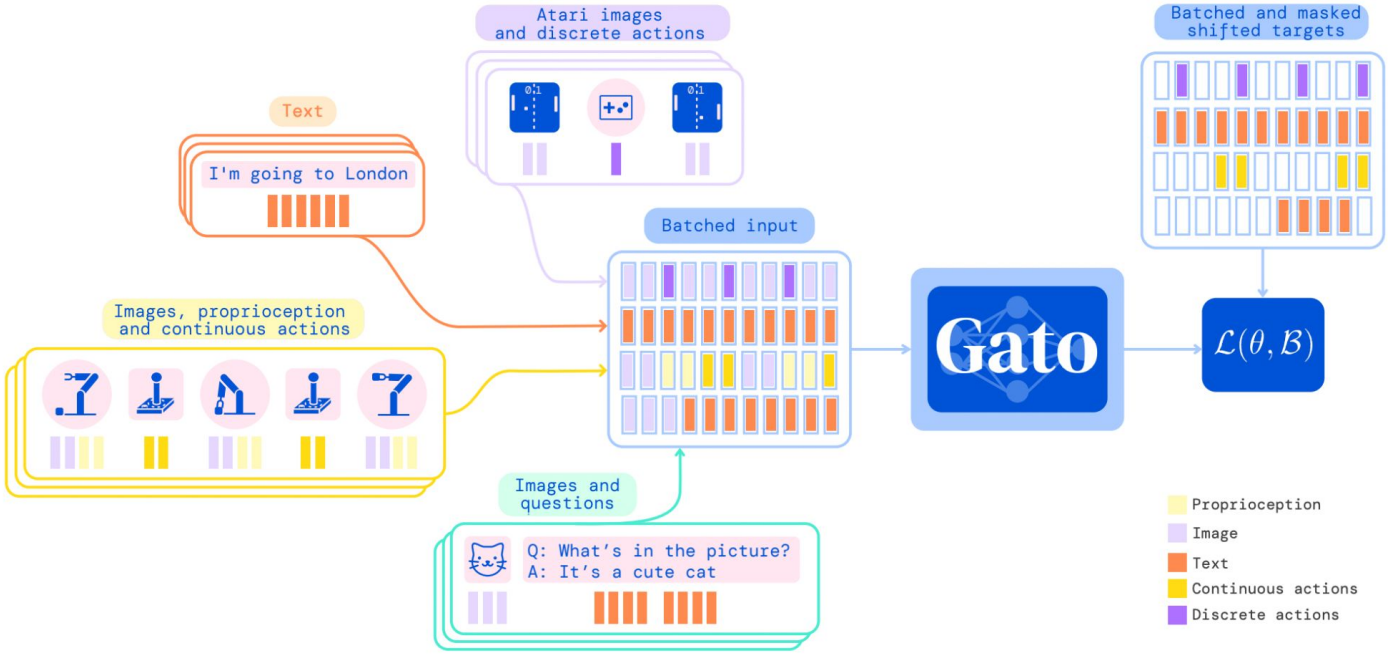
https://github.com/FomalhautB/KM-BART

Fusion Brain



https://github.com/ai-forever/fusion_brain_aij2021

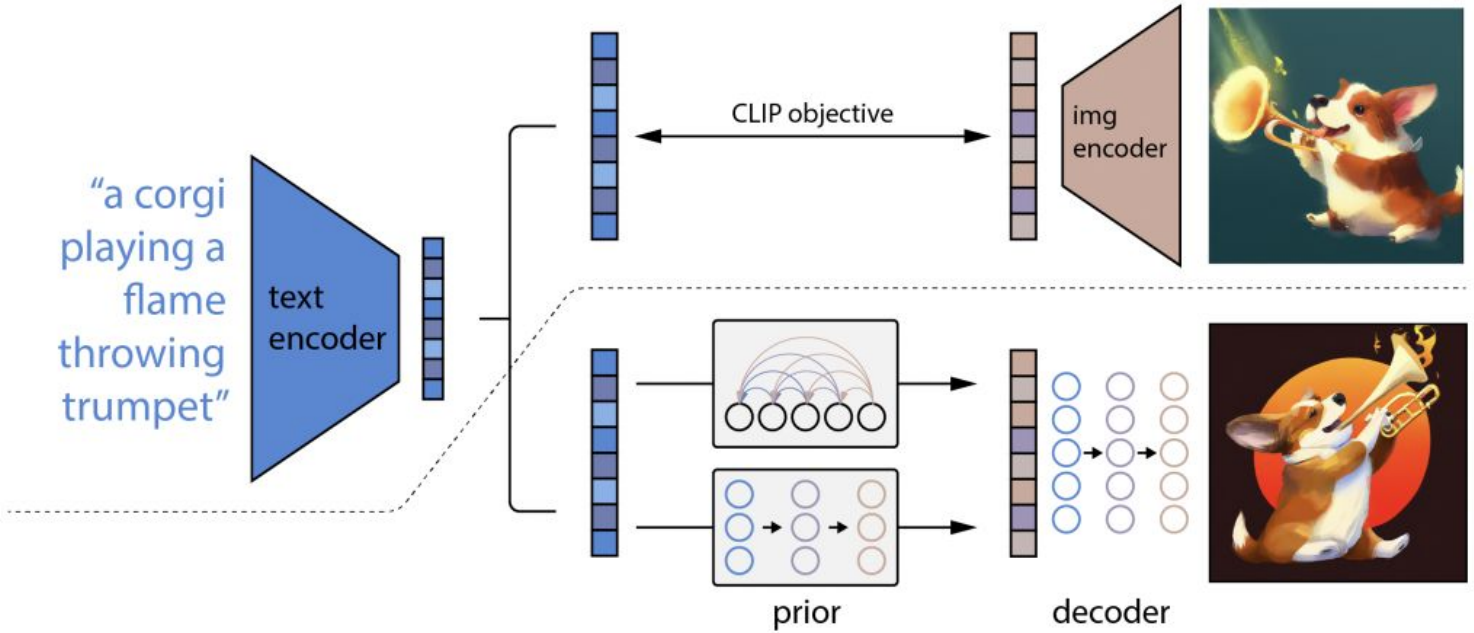
GATO



идея

- берем много задач и учим GPT like

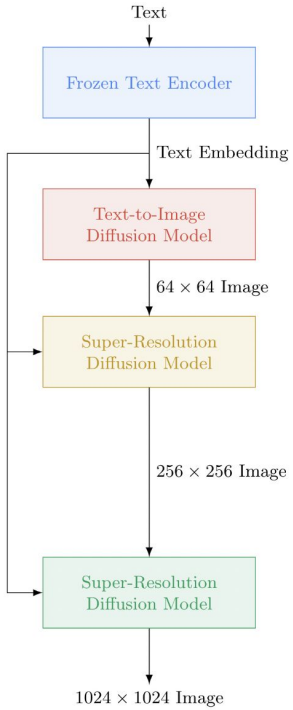
DALLE 2

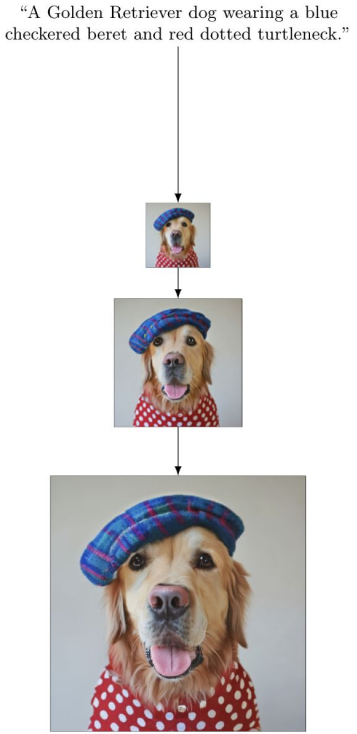


https://cdn.openai.com/papers/dall-e-2.pdf

ImageGen

Используем энкодер T5 вместо CLIP text encoder





Спасибо за внимание!

Мои контакты

https://t.me/Alexwortegae

Давайте что нибудь потыкаем

https://colab.research.google.com/drive/12YRRzhl5cHER_U2F-buQxif8GlhMPWq3?usp=sharing#scrollTo=OJTpZacBn0OI