

Text + Image. CLIP. DALL-E

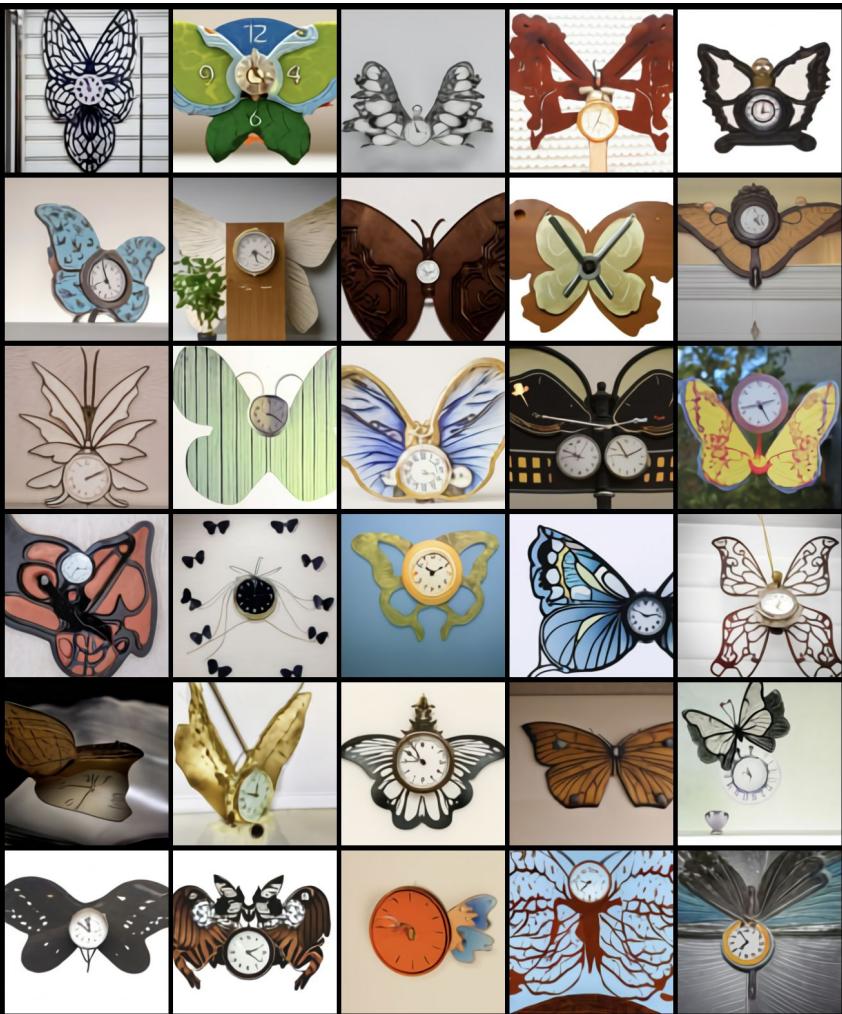
Даниил Лысухин, ML Team Lead @ Ozon





An illustration of a Pikachu in a leather jacket using a calculator

An **clock** in a **form** of a **butterfly wing**





A neon sign that reads 'gpt'. gpt neon sign. 'g p t'. gpt typography

<https://openai.com/blog/dall-e>

(кресло в форме авокадо тоже там)

Захотелось лекцию про CLIP & DALL-E.

- Zero-shot learning?
- Соединение текста и изображения в одной модели?
- Просто посмотреть картинки?

Захотелось лекцию про CLIP & DALL-E.

- Zero-shot learning?
- Соединение текста и изображения в одной модели?
- Просто посмотреть картинки?



Yes.

Supervised learning

- Классическое обучение с учителем:
 - Взяли ~ целевые данные, **размеченные**
 - Взяли модель
 - Обучили её
 - ???
 - PROFIT!

Supervised learning vs мало данных

- ~ Целевых данных много, но размечена малая часть?

Supervised learning vs мало данных

- ~ Целевых данных много, но размечена малая часть?
 - Выход 1: **Self-supervised learning**
 - Обучим модель для другой задачи, не требующей разметки
 - Пример: классификация смежных регионов ↗
 - Другой пример: автоэнкодеры (wait for it)
 - Получим pretrained-модель, которая (как-то) "понимает" данные
 - Далее - finetuning на размеченных данных
 - Еще

Supervised learning vs мало данных

- ~ Целевых данных много, но размечена малая часть?
 - Выход 2: **Semi-supervised learning**
 - Например, использовать псевдо-разметку
 - Обучим модель на размеченной части данных
 - Предскажем метки для неразмеченной части
 - (Как-то) отфильтруем, добавим в размеченную часть данных
 - Повторим N раз
 - Еще

Supervised learning vs мало данных

- ~ Целевых данных много, но размечена малая часть?
 - Выход 3: Active Learning
 - Выход 4: Генерация данных
 - ...

Supervised learning vs мало данных

- А если непосредственно целевых данных нет вообще?

Supervised learning vs мало данных

- А если непосредственно целевых данных нет вообще?
 - Пусть надо определять собак в шляпах

Supervised learning vs мало данных

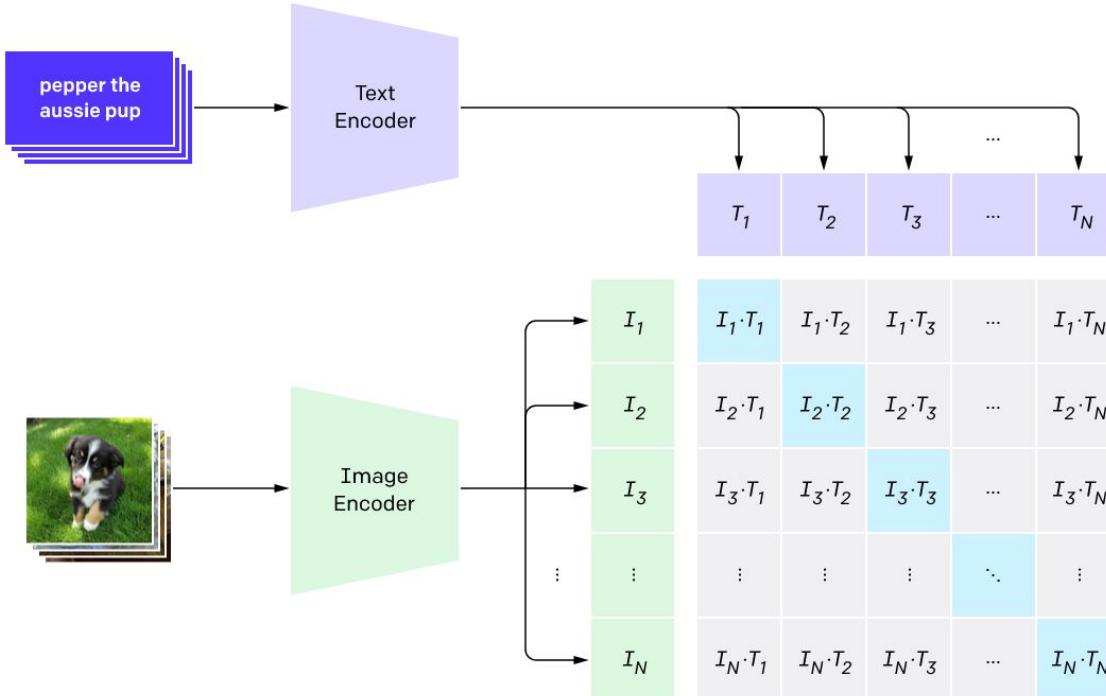
- А если непосредственно целевых данных нет вообще?
 - Пусть надо определять собак в шляпах
 - В ImageNet есть классы "шляпа" и класс "собака"
 - Сможет ли предобученная на ImageNet модель классификации определять собак в шляпах?
 - *Может быть*

OpenAI CLIP

- [Learning Transferable Visual Models From Natural Language Supervision](#) (2021)
- 1 Энкодер для изображения
- 1 Энкодер для текстового описания
- 400 000 000 пар "изображение + текст"
- На выходе модель с *хорошой* способностью к Zero-Shot-классификации

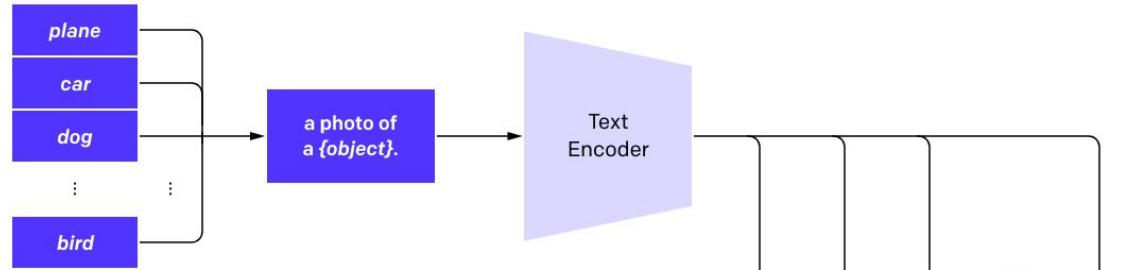
OpenAI CLIP: pretraining

1. Contrastive pre-training

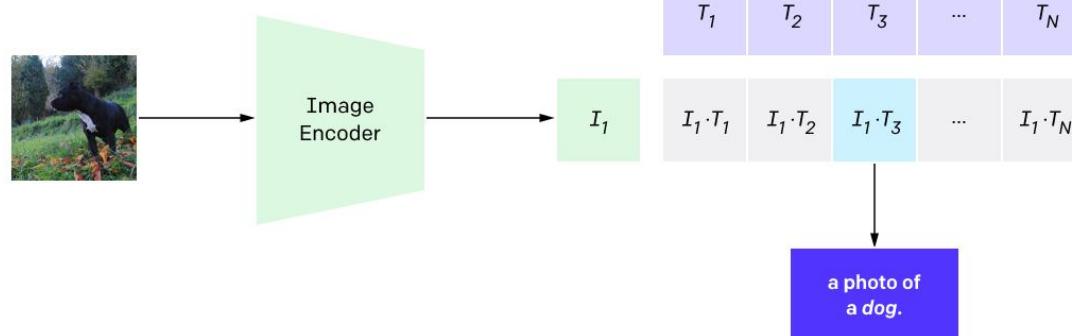


OpenAI CLIP: zero-shot inference

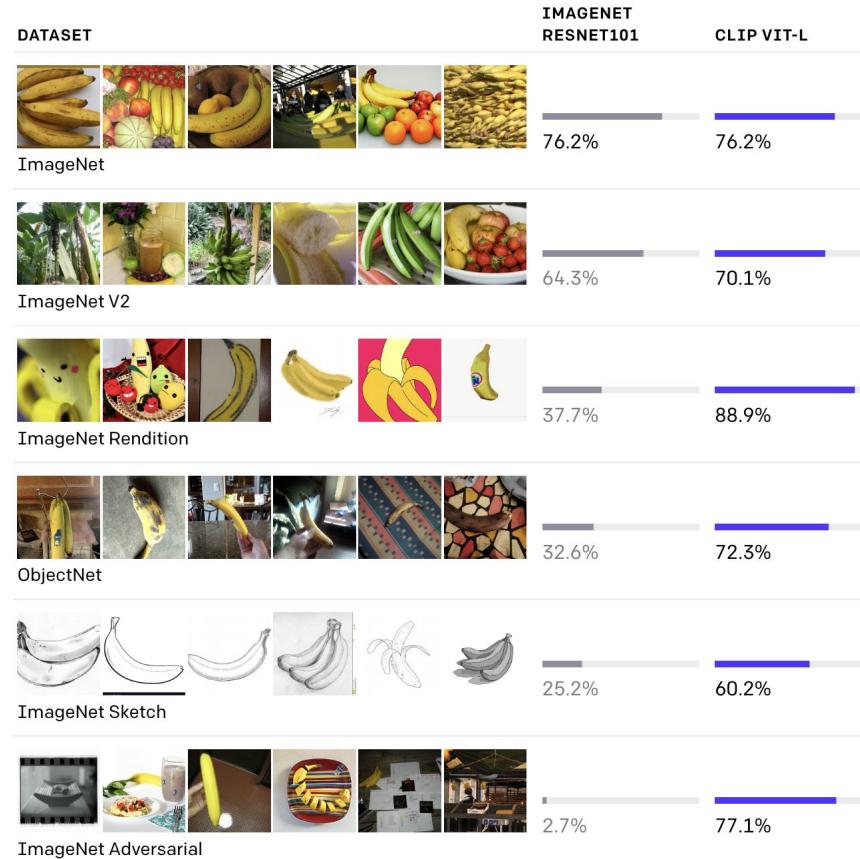
2. Create dataset classifier from label text



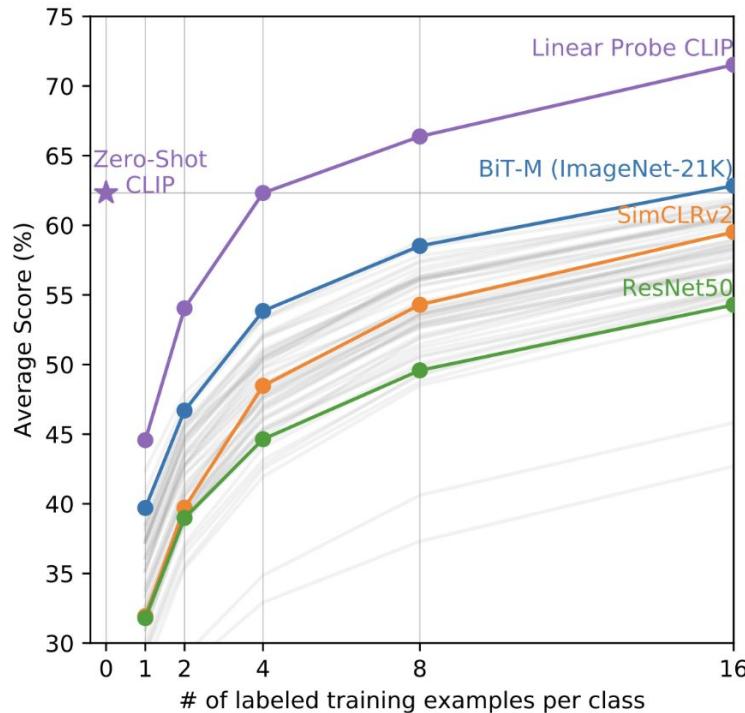
3. Use for zero-shot prediction



OpenAI CLIP: кругозор модели

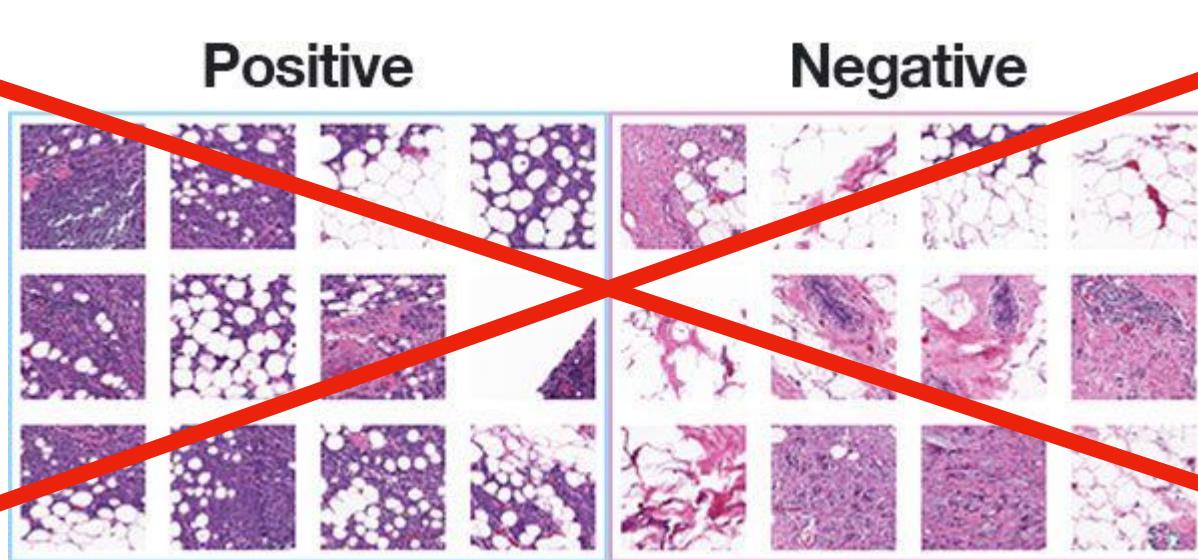


OpenAI CLIP: zero-shot vs linear-probes



OpenAI CLIP: не по назначению

- CLIP здорово работает на "бытовых" сущностях
- На специфических данных (медицины, спутники, ...) пользоваться им не стоит



OpenAI CLIP: побочные эффекты

- Пред-обучение и датасет так хороши, что модель научилась читать

Attack text label iPod ▾



Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%



Granny Smith	0.1%
iPod	99.7%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.0%

When we put a label saying “iPod” on this Granny Smith apple, the model erroneously classifies it as an iPod in the zero-shot setting.

<https://openai.com/blog/multimodal-neurons>

OpenAI CLIP: итого

- "Обучение без обучения"
 - Взяли готовый CLIP (ResNet50x, ViT-*)
 - Сформировали описание класса
 - "a photo of dog wearing hat"
 - Вычислили его эмбеддинг (`clip.text_encoder(text)`)
 - Вычислили эмбеддинги изображений
(`clip.image_encoder(image)`)
 - Скор модели = `similarity(text_emb, image_emb)`

OpenAI CLIP: итого

- CLIP соединяет данные двух различных модальностей, текста и изображений
 - О "мультимодальных" нейронах
 - В теории, можно расширить список модальностей (например, аудио)
 - ...был бы датасет на сотни миллионов примеров

OpenAI CLIP: код

- [github](#)

Zero-shot classification

- Классификатор без обучения делать научились
- Как насчет генерации?

OpenAI DALL-E

- Zero-Shot Text-to-Image Generation (2021)
- На вход - текстовый запрос (+ маскированное изображение)
- На выходе - изображение
- Под капотом: ↗
 - GPT-3
 - dVAE
 - CLIP
 - ... 250 000 000 пар изображение+текст (куда делись 150М?)

OpenAI DALL-E

- Zero-Shot Text-to-Image Generation (2021)
- На вход - текстовый запрос (+ маскированное изображение)
- На выходе - изображение
- Под капотом:
 - **GPT-3**
 - dVAE
 - CLIP
 - ... 250 000 000 пар изображение+текст (куда делись 150М?)

OpenAI GPTs

- (GPT-1) Improving Language Understanding by Generative Pre-Training (2018)
- (GPT-2) Language Models are Unsupervised Multitask Learners (2019)
- (GPT-3) Language Models are Few-Shot Learners (2020)
- Всё это - **языковые модели**

Language Models



Hey, Jude, don't ...

A: make it bad

B: be afraid

C: let me down

D: stop me now

Language Models

- Смоделировать язык ~= оценить вероятность появления в языке конкретного элемента (слова, фразы, ...)
 - $p(y) = ?$
- Тот же пример (условно):
 - $p(\text{Hey, Jude. Don't make it bad}) \sim 0.33(3)\dots$
 - $p(\text{Hey, Jude. Don't be afraid}) \sim 0.33(3)\dots$
 - $p(\text{Hey, Jude. Don't let me down}) \sim 0.33(3)\dots$
 - $p(\text{Hey, Jude. Don't stop me now}) \sim 0\dots?$
 - ...

Language Models

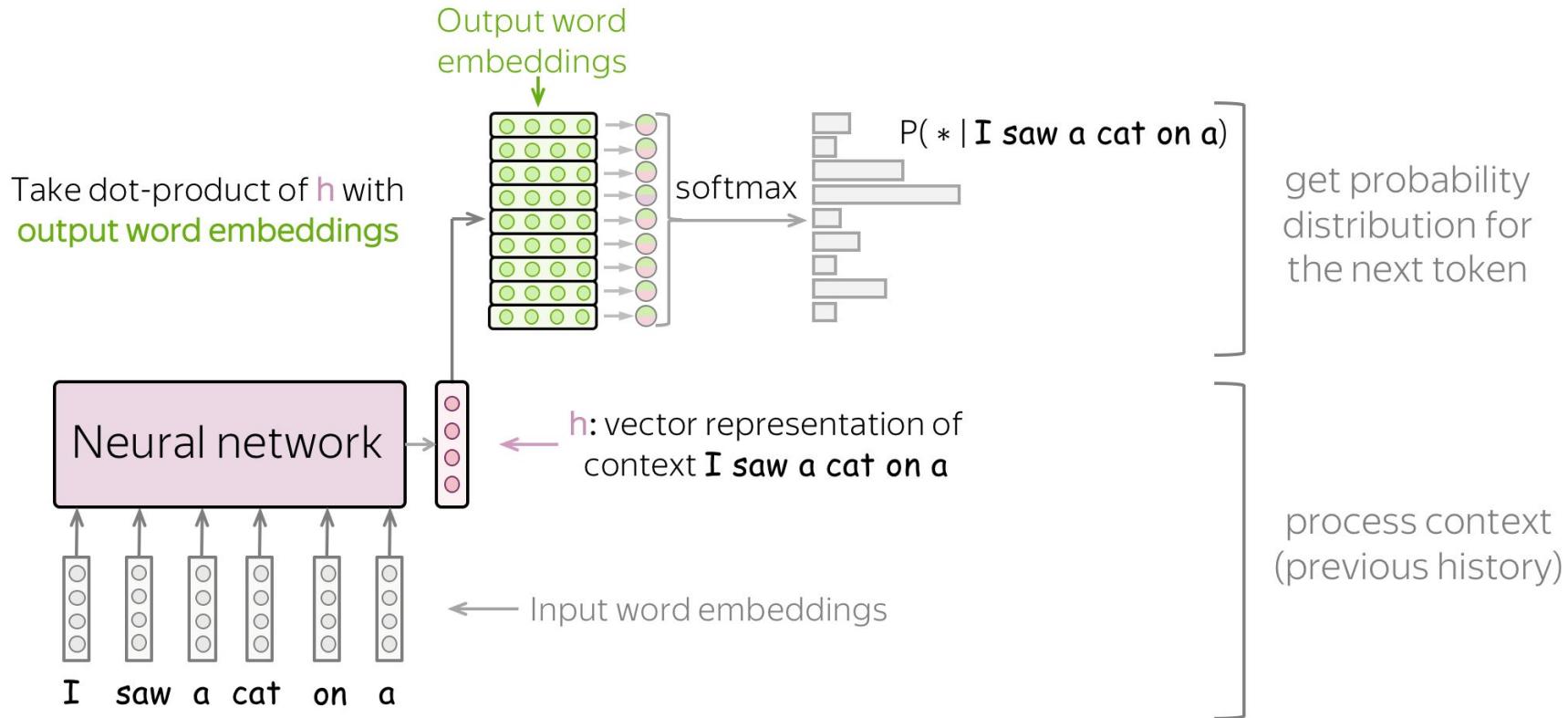
- Оценить напрямую вероятность появления целого предложения y (т.е. $p(y)$) крайне трудно
 - Рабочий подход - разложить предложение на отдельные элементы - токены (слова или суб-слова) (left-to-right LMs)
 - $y = [\text{hey}] [\text{jude}] [\text{don't}] [\text{be}] [\text{afraid}]$
 - Тогда $p(y) =$
 - $p(\text{hey}) *$
 - $p(\text{jude} \mid \text{hey}) *$
 - $p(\text{don't} \mid \text{hey jude}) *$
 - $p(\text{be} \mid \text{hey jude don't})$
 - $p(\text{afraid} \mid \text{hey jude don't be})$

Language Models

$$P(y_1, y_2, \dots, y_n) = P(y_1) \cdot P(y_2|y_1) \cdot P(y_3|y_1, y_2) \cdot \dots \cdot P(y_n|y_1, \dots, y_{n-1}) = \prod_{t=1}^n P(y_t|y_{\leq t}).$$

- Как получить эти условные вероятности?
 - Посчитать по датасету частоты
 - Использовать n-grams: $p(y_t|y_{\leq t}) = p(y_t|y_{t-1}, y_{t-2}, \dots, y_{t-n+1})$
 - Использовать обучаемые модели для последовательностей
 - Задача = предсказание (классификация по словарю) следующего элемента последовательности
 - RNNs ↗
 - Transformers

Language Models - Training



Language Models - Generation

- Обученную языковую модель можно использовать для генерации текста
- Авторегрессионная генерация:
 - Модель получает на вход "затравку" (prompt)
 - Из затравки получается контекст
 - По контексту предсказывается распределение по словарю
 - Сэмплируется новый токен
 - Новая затравка = старая затравка + предсказанный токен

Language Models keks @ 2022

- This is the worst AI ever
 - Сказ о том, как GPT на постах с 4chan обучили
- Did Google's LaMDA chatbot just become sentient?
 - Сказ о том, как инженер Google нанял адвоката для разумной языковой модели

~~Language~~ Sequence Models

- "Языковые" модели - частный случай
- Логику обучения LM можно обобщить на многие другие последовательности
 - Предсказание временных рядов
 - Генерация музыки (опять OpenAI!)
 - ... и изображений?

OpenAI GPTs

- (GPT-1) Improving Language Understanding by Generative Pre-Training (2018)
- (GPT-2) Language Models are Unsupervised Multitask Learners (2019)
- (GPT-3) Language Models are Few-Shot Learners (2020)
- Всё это - **языковые модели**

OpenAI GPTs

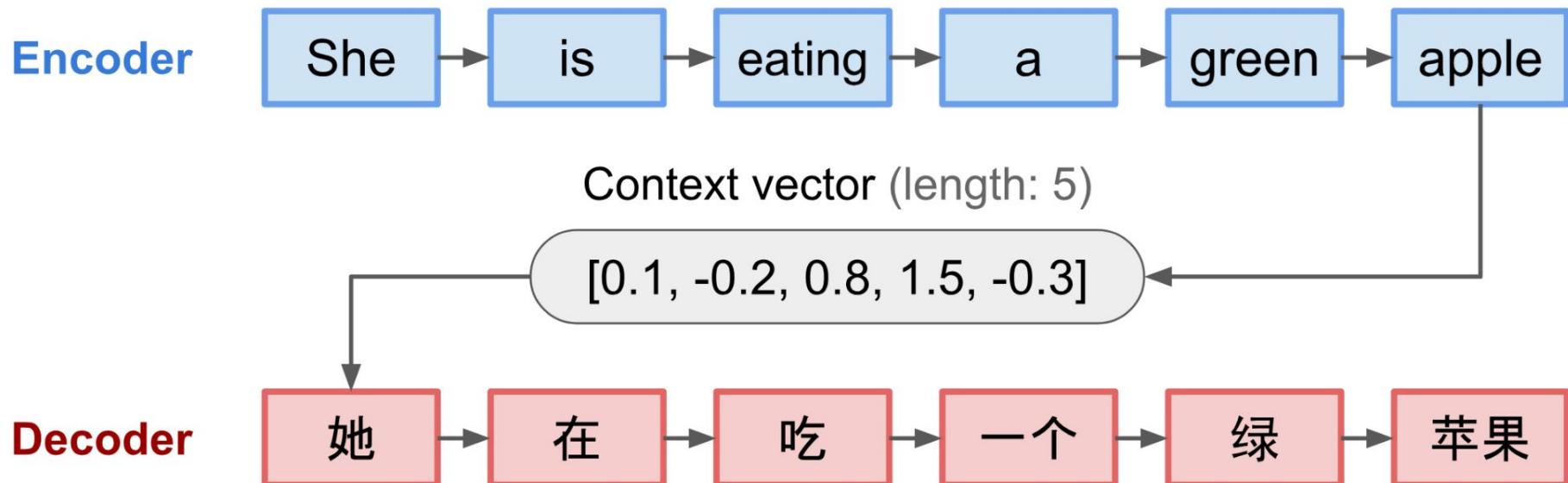
- (GPT-1) [Improving Language Understanding by Generative Pre-Training](#) (2018)
- (GPT-2) [Language Models are Unsupervised Multitask Learners](#) (2019)
- (GPT-3) [Language Models are Few-Shot Learners](#) (2020)
- Всё это - **языковые модели**
 - **Большие языковые модели**
 - GPT-3: max 175B params
 - Датасет: > 500B токенов
 - На основе механизма (self-)**Attention**

Attention

- [Neural Machine Translation by Jointly Learning to Align and Translate](#) (2016) (Attention)
- Был предложен как улучшение подхода seq2seq в задаче машинного перевода

Seq2seq

- Seq2seq: RNN-энкодер -> вектор контекста -> RNN-декодер

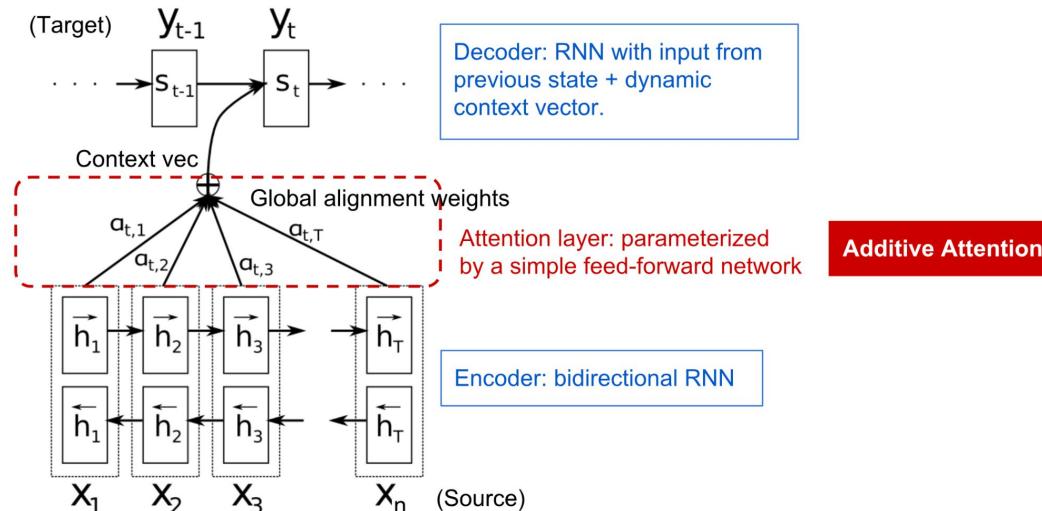


Seq2seq

- Seq2seq: RNN-энкодер -> вектор контекста -> RNN-декодер
- Проблемы RNN:
 - К концу последовательности забывают, что было в начале
 - Сложно сжать большой контекст в единственный вектор
 - Есть трудности с обучением ("длинный" backprop)

Seq2seq + Attention

- Разрешим декодеру смотреть не на один вектор контекста, а на всю входную последовательность токенов ↗
- Важность токенов ~ обучаемые веса



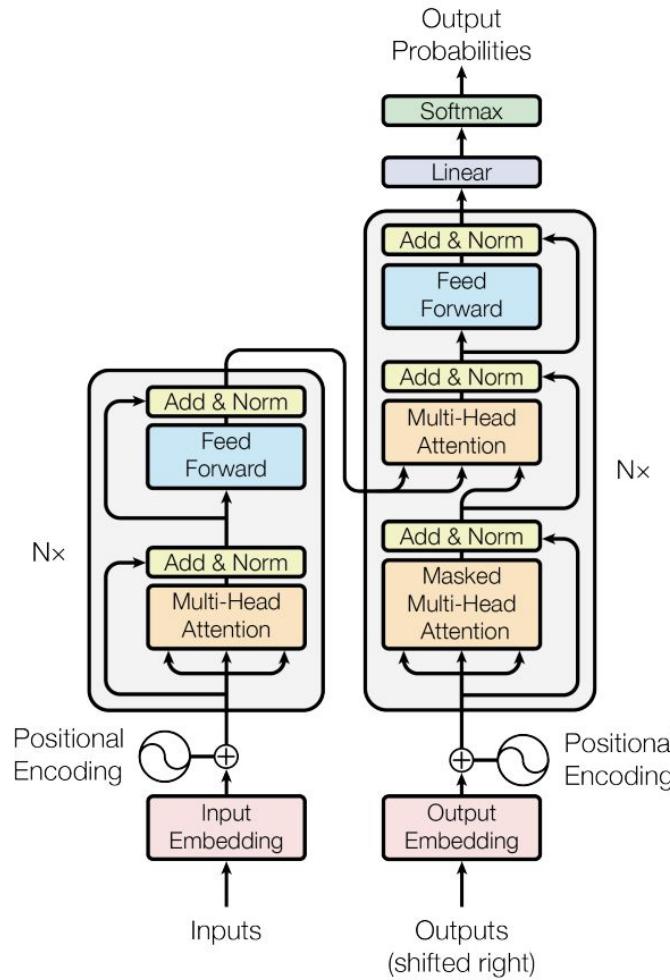
Attention vs RNN

- В оригинальной статье Attention связывал блоки RNN-декодера и RNN-энкодера
 - А что, если вообще убрать RNN, и оставить только Attention?

Transformer

- [Attention is All You Need](#) (2017) (Transformer)
- По аналогии с seq2seq в модели Transformer есть две части - энкодер и декодер
 - Но все взаимодействие токенов (между собой + между блоками) - через Attention
 - Информация о взаимном положении токенов в последовательностях кодируется отдельно (Positional Encoding)

Transformer



OpenAI GPTs

- (GPT-1) [Improving Language Understanding by Generative Pre-Training](#) (2018)
- (GPT-2) [Language Models are Unsupervised Multitask Learners](#) (2019)
- (GPT-3) [Language Models are Few-Shot Learners](#) (2020)
- Всё это - **языковые модели**
 - **Большие языковые модели**
 - GPT-3: max 175B params
 - Датасет: > 500B токенов
 - На основе механизма (self-)**Attention**

OpenAI GPTs

- (GPT-1) [Improving Language Understanding by Generative Pre-Training](#) (2018)
- (GPT-2) [Language Models are Unsupervised Multitask Learners](#) (2019)
- (GPT-3) [Language Models are Few-Shot Learners](#) (2020)
- Всё это - **языковые модели**
 - **Большие языковые модели**
 - GPT-3: max 175B params
 - Датасет: > 500B токенов
 - На основе механизма (self-)**Attention**
 - **GPT-* ~= Transformer Decoder**

OpenAI GPT-3

- GPT-3 очень большая и очень "начитанная"
- Может решать *некоторые* языковые задачи в режимах Few-shot и даже Zero-shot
- Ну и генерировать текст по затравке, разумеется

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1 Translate English to French:      ← task description  
2 cheese => .....                ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1 Translate English to French:      ← task description  
2 sea otter => loutre de mer    ← example  
3 cheese => .....                ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1 Translate English to French:      ← task description  
2 sea otter => loutre de mer    ← examples  
3 peppermint => menthe poivrée  ← examples  
4 plush girafe => girafe peluche ← examples  
5 cheese => .....                ← prompt
```

OpenAI GPT-3

- Свободного доступа к GPT-3 от OpenAI нет
- Но есть альтернативы
 - GPT-NeoX 20B / GPT-J 6B: <https://20b.eleuther.ai/>
 - Sber ruGPT-3-XL: <https://russiannlp.github.io/rugpt-demo/>

OpenAI DALL-E

- Zero-Shot Text-to-Image Generation (2021)
- На вход - текстовый запрос (+ маскированное изображение)
- На выходе - изображение
- Под капотом:
 - **GPT-3**
 - dVAE
 - CLIP
 - ... 250 000 000 пар изображение+текст (куда делись 150М?)

OpenAI DALL-E

- Zero-Shot Text-to-Image Generation (2021)
- На вход - текстовый запрос (+ маскированное изображение)
- На выходе - изображение
- Под капотом:
 - GPT-3
 - **dVAE**
 - CLIP
 - ... 250 000 000 пар изображение+текст (куда делись 150М?)

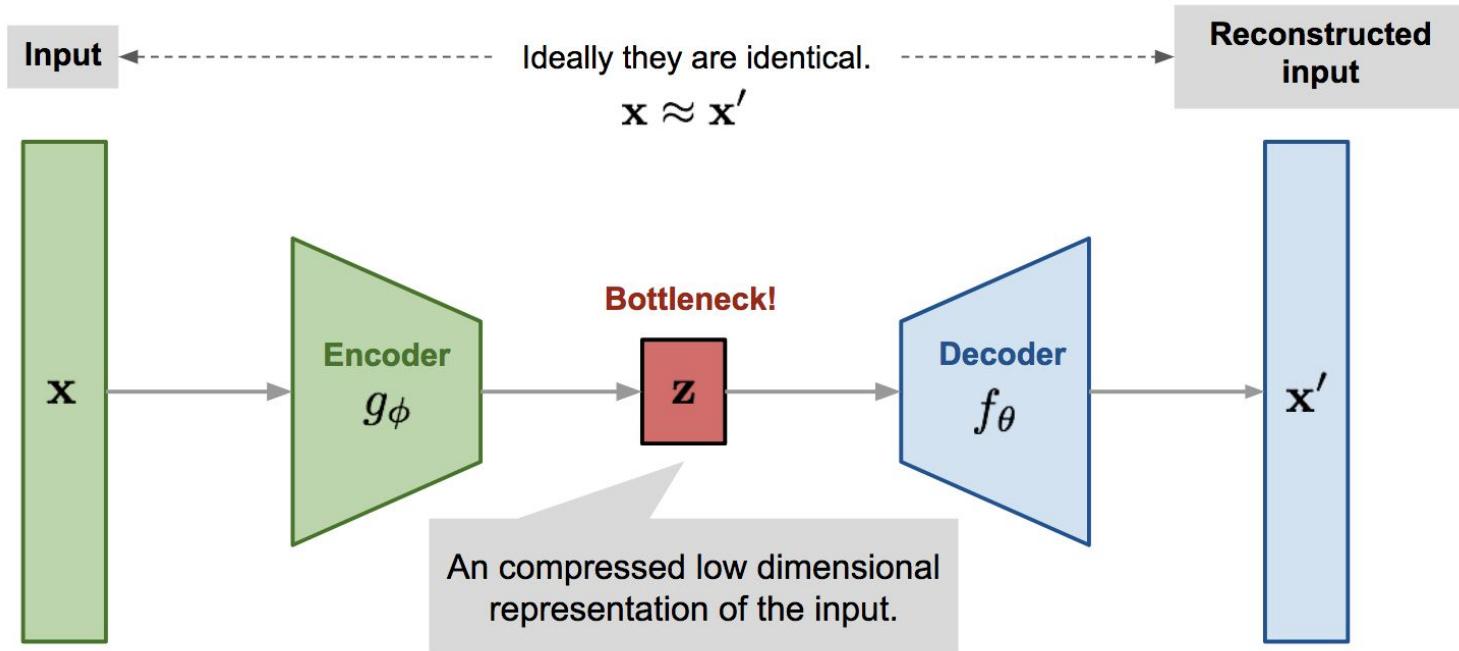
dVAE

- За "картиночную" часть DALL-E отвечает **dVAE**
- **dVAE** = **discrete Variational AutoEncoder**

Autoencoders

- Автокодировщик - это архитектура для сжатия многомерных объектов до меньших размерностей
- Кодирование выполняется так, чтобы по сжатому представлению можно было восстановить исходное

Autoencoders



4

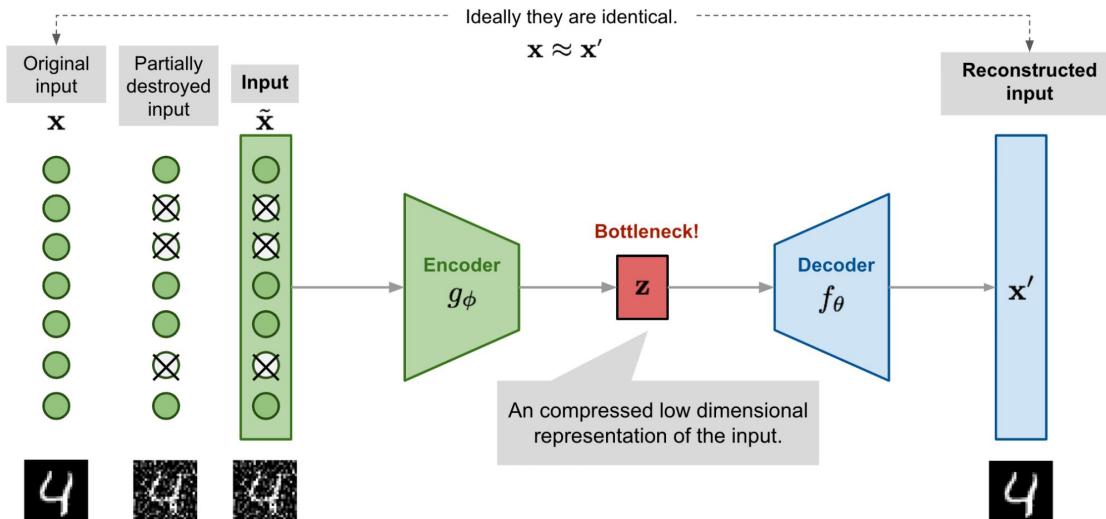
4

Autoencoders

- Автокодировщик - это архитектура для сжатия многомерных объектов до меньших размерностей
- Кодирование выполняется так, чтобы по сжатому представлению можно было восстановить исходное
 - Почему это может работать? ↗

Autoencoders

- **Denoising Autoencoder:** на вход подаются зашумленные объекты, восстанавливать нужно исходные



Autoencoders

- **Sparse Autoencoder:** принуждаем вектор скрытого состояния к разреженности
 - Накладываем регуляризацию на сжатое представление \mathbf{z}
 - Чтобы (в среднем) только часть активаций была ненулевой

Autoencoders as generators

- Если из обученного автокодировщика убрать энкодер, то получится генератор? ↗

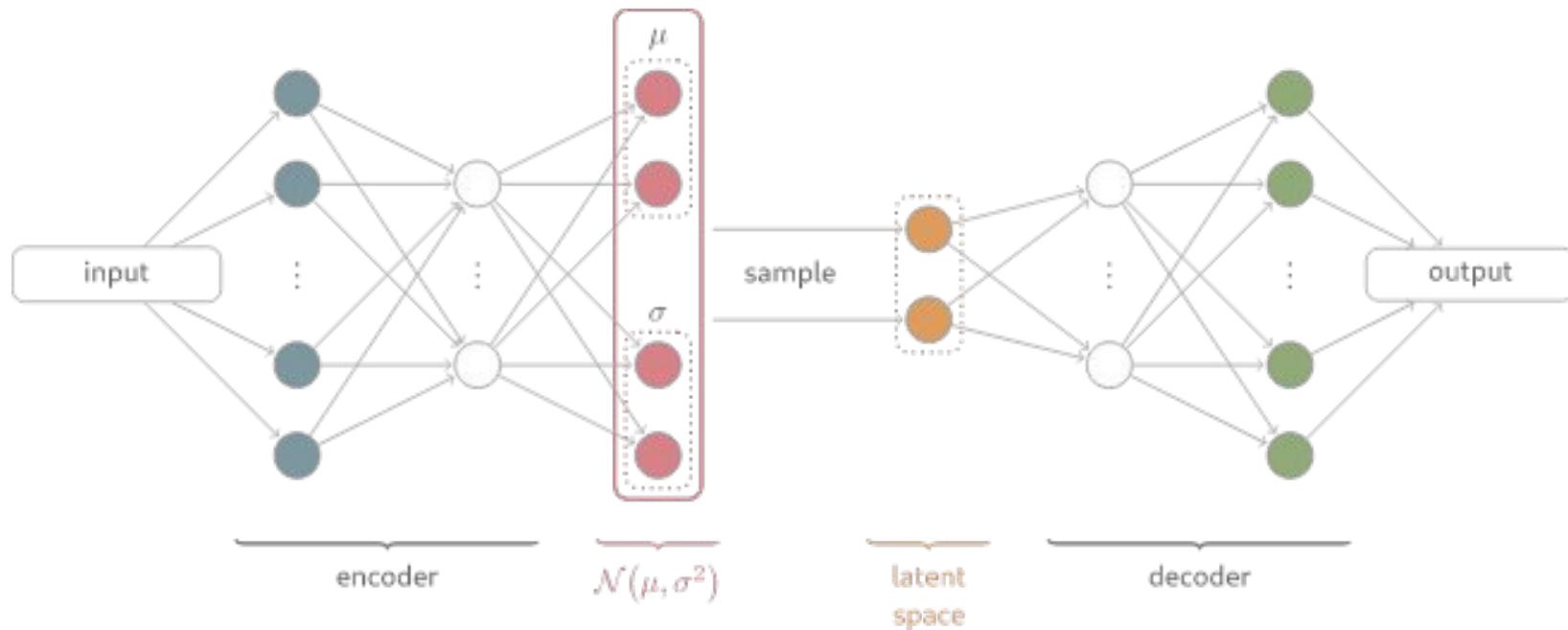
Autoencoders as generators

- Для хорошей генерации с помощью декодера АЕ от скрытого пространства требуется:
 - "Непрерывность": интерполяция между двумя точками должна быть гладкой
 - "Полнота": из случайного вектора z хочется иметь правдоподобно восстановленный x
- Ванильные автокодировщики такими свойствами обладают редко

Variational Autoencoders (VAE)

- В вариационном автокодировщике скрытое пространство стремятся сделать "непрерывным" и "полным"
 - Вектор x отображается не в единственную точку z , а в случайный вектор из нормального распределения
 - Параметры нормального распределения выдаются энкодером \Rightarrow

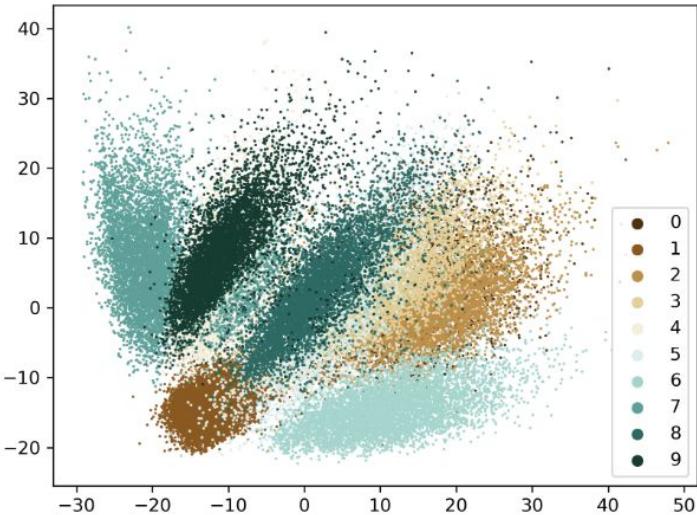
Variational Autoencoders (VAE)



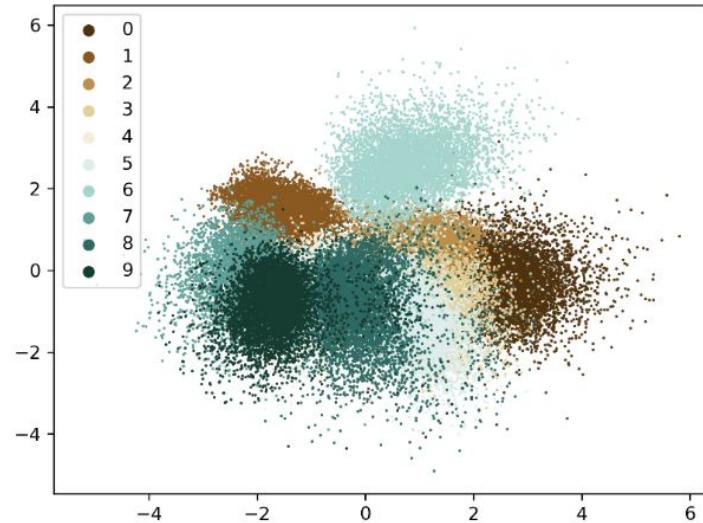
Variational Autoencoders (VAE)

- Теперь оптимизируется сумма лоссов:
 - Лосс восстановления \mathbf{x} (как в обычном AE)
 - KL-дивергенция между предсказанными параметрами нормального распределения и $N(0, I)$
- Лоссы друг другу "противоречат"?

Variational Autoencoders (VAE)



(a) Latent Distribution by Label for AE

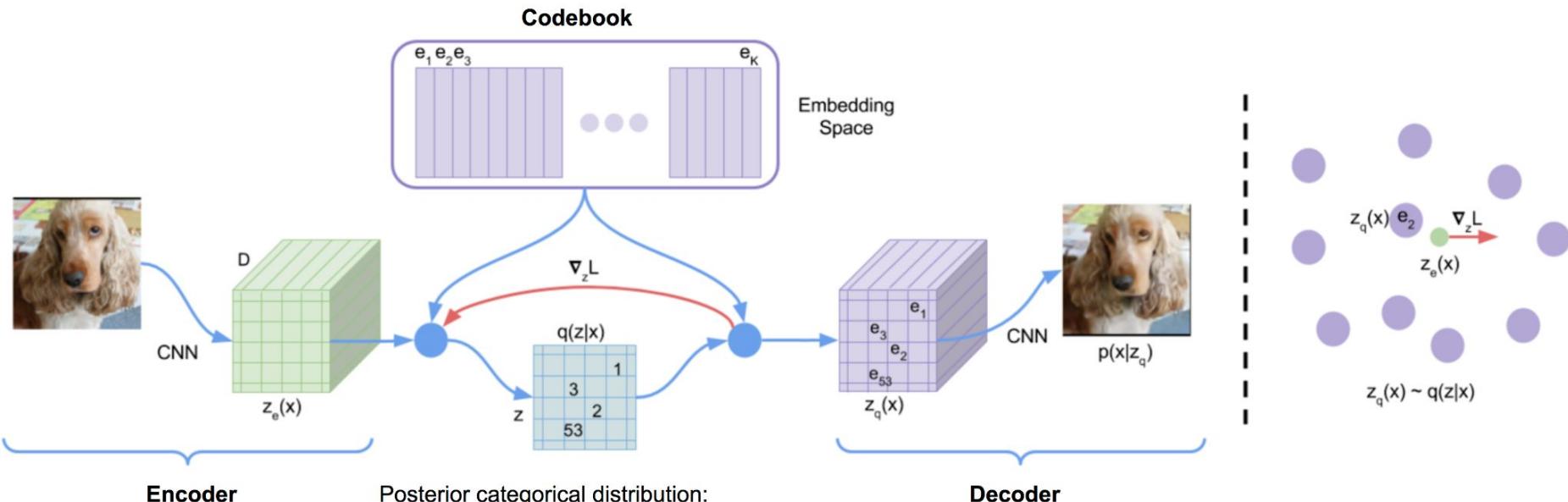


(b) Latent Distribution by Label for VAE

Vector-Quantized Variational Autoencoders (VQ-VAE)

- В VAE заставляли скрытое пространство вести себя, как смесь гауссиан
- В VQ-VAE вообще представляем его как дискретный набор обучаемых векторов (т.е. категориальное распределение)

Vector-Quantized Variational Autoencoders (VQ-VAE)



$$q(\mathbf{z} = \mathbf{e}_k | \mathbf{x}) = \begin{cases} 1 & \text{if } k = \arg \min_i \|\mathbf{z}_e(\mathbf{x}) - \mathbf{e}_i\|_2 \\ 0 & \text{otherwise.} \end{cases}$$

<http://papers.nips.cc/paper/7210-neural-discrete-representation-learning.pdf>

Vector-Quantized Variational Autoencoders (VQ-VAE)

- Изображение подается в энкодер
- Предсказывается набор (например, сетка 32x32) из векторов $\mathbf{z}_1, \dots, \mathbf{z}_n$
- Для каждого из векторов находится ближайший из обучаемого "словаря" (codebook)
- Полученные векторы из словаря подаются в декодер

Vector-Quantized Variational Autoencoders (VQ-VAE)

- Как можно генерировать разные изображения из ограниченного числа векторов?

Discrete VAE (dVAE)

- Идея такая же, как в VQ-VAE (словарь эмбеддингов)
- Свой способ решения проблемы протекания градиентов через bottleneck
 - [The Gumbel Softmax Relaxation](#)

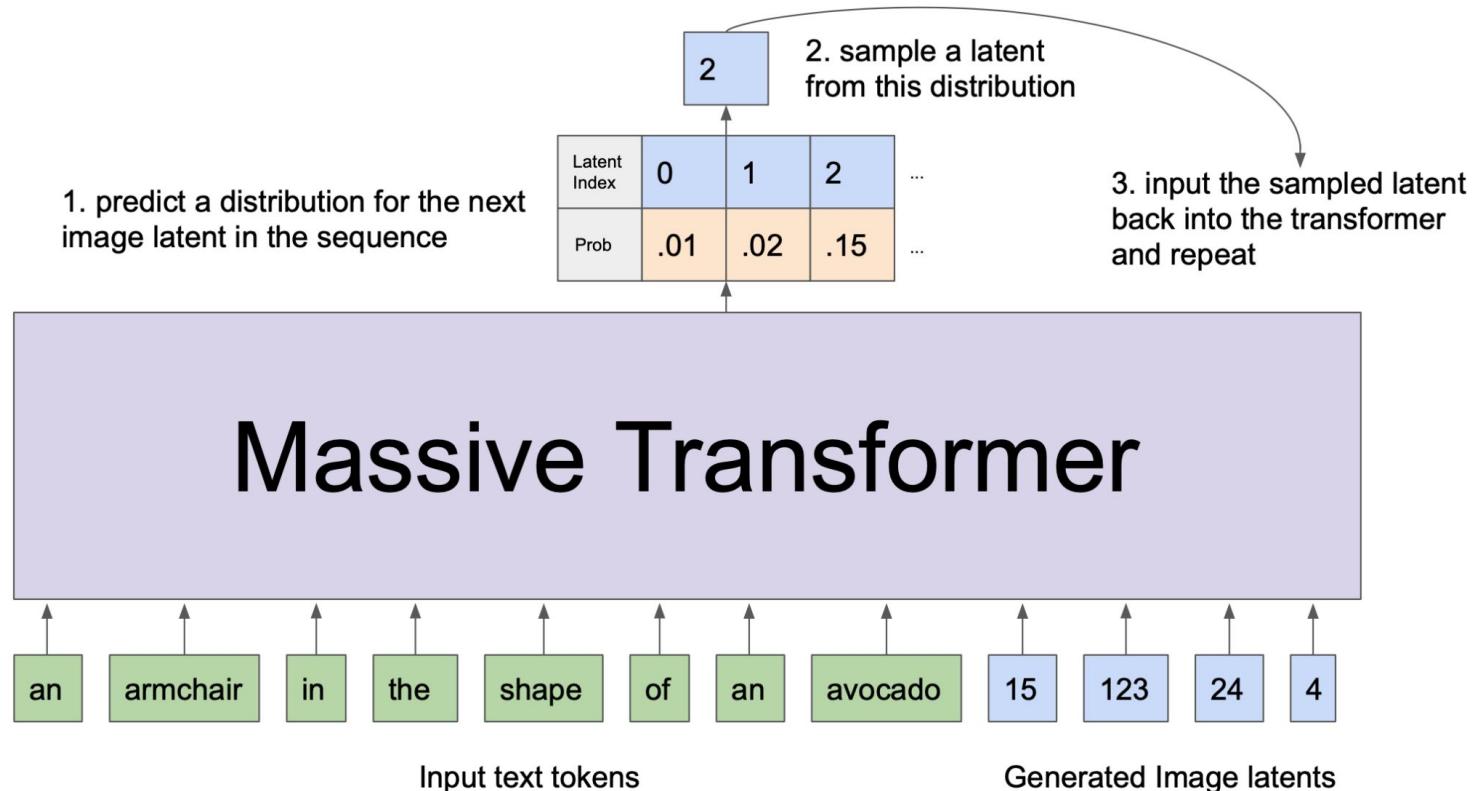
OpenAI DALL-E

- Zero-Shot Text-to-Image Generation (2021)
- На вход - текстовый запрос (+ маскированное изображение)
- На выходе - изображение
- Под капотом:
 - GPT-3
 - **dVAE**
 - CLIP
 - ... 250 000 000 пар изображение+текст (куда делись 150М?)

OpenAI DALL-E

- Zero-Shot Text-to-Image Generation (2021)
- На вход - текстовый запрос (+ маскированное изображение)
- На выходе - изображение
- Под капотом:
 - GPT-3
 - dVAE
 - CLIP
 - ... 250 000 000 пар изображение+текст (куда делись 150М?)

OpenAI DALL-E



OpenAI DALL-E

- Обучение - как у языковых моделей
 - В качестве последовательностей - 256 токенов для текста и 1024 токена для изображения
 - Предсказание токенов для изображения - авторегрессионное

x_1				x_n
		x_i		
				x_{n^2}

OpenAI DALL-E

- В качестве "затравки" можно передавать не только текст, но и часть изображения (верхние строки)
 - Изображение будет пропущено через энкодер dVAE для получения токенов
 - Эти токены добавятся к токенам текста затравки

OpenAI DALL-E

- Zero-Shot Text-to-Image Generation (2021)
- На вход - текстовый запрос (+ маскированное изображение)
- На выходе - изображение
- Под капотом:
 - GPT-3
 - dVAE
 - CLIP???
 - ... 250 000 000 пар изображение+текст (куда делись 150М?)

OpenAI DALL-E

- Вспомним, что задача моделирования последовательностей - это классификация
- На каждом шаге модель возвращает распределение вероятностей по словарю для следующего токена
 - Можно всегда брать argmax (скучно)
 - Можно сэмплировать из этого распределения

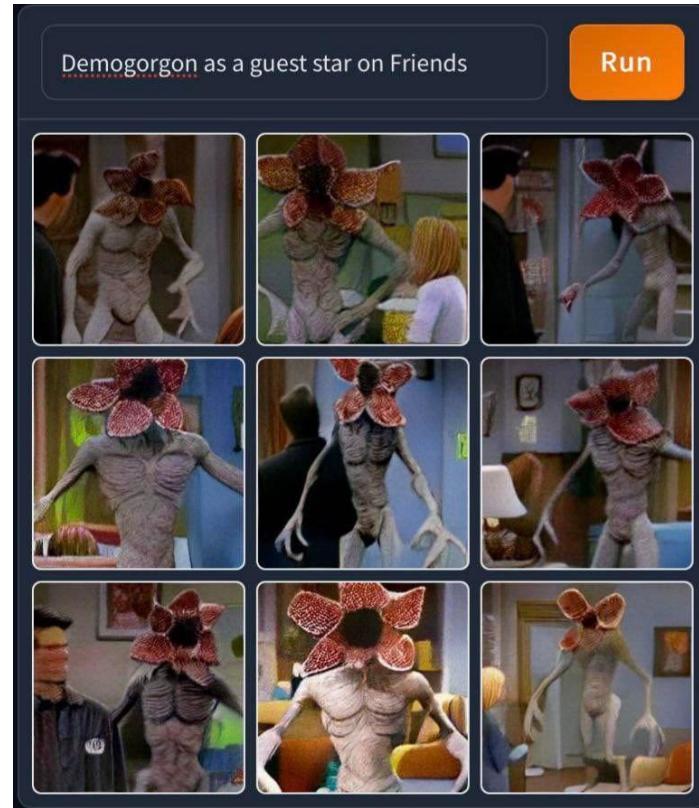
OpenAI DALL-E

- В DALL-E генерируется много (256) различных наборов токенов
- Каждый набор = 1 сгенерированное изображение
- Из полученных 256 изображений с помощью CLIP отбираются топ-сколько-то
 - Был запрос "собака в шляпе"
 - DALL-E сгенерировал 256 вариантов
 - CLIP ранжирует сгенерированные изображения по близости к запросу "собака в шляпе" и возвращает лучшие из них

OpenAI DALL-E

- Открытого доступа к оригинальной DALL-E нет
- Но есть альтернативы
 - dalle-mini: <https://www.craiyon.com/>
 - must-see: <https://twitter.com/weirddalle>
 - Sber ruDALLE-XL: <https://rudalle.ru/demo>

@weirddallegenerations (dalle-mini)



@weirddallegenerations (dalle-mini)



OpenAI DALL-E - итого

- Авторегрессионная генеративная языковая модель + декодер для отрисовки + мультимодальная модель для ранжирования
- Работает с текстовыми и "визуальными" токенами
- Много весов и много данных

COSMOPOLITAN

the A.I. issue

Meet the
World's
First
Artificially
Intelligent
Magazine
Cover



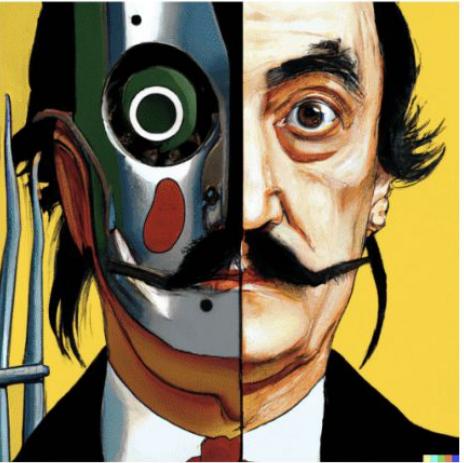
An astronaut in a white spacesuit stands on the surface of the moon, looking down at the ground. The background is the dark void of space with distant stars.

And it
only took 20
seconds to make.



OpenAI DALL-E 2

- [Hierarchical Text-Conditional Image Generation with CLIP Latents \(2022\)](#)
 - [Описание попроще от одного из авторов](#)
- CLIP для получения эмбеддингов
- Диффузионная модель для генерации изображений
 - [О диффузионных моделях в целом](#)
 - [Подборка статей по ним](#)
- 1024x1024 (было 256x256)



vibrant portrait painting of Salvador Dalí with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



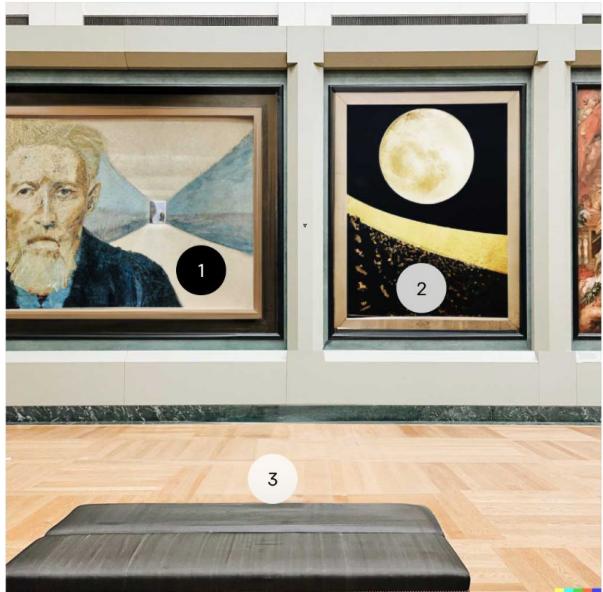
a corgi's head depicted as an explosion of a nebula

OpenAI DALL-E 2

- Помимо text2image генерации, умеет
 - Делать Inpainting: заменять отдельные области изображения в соответствии с текстовым запросом

OpenAI DALL-E 2 - Inpainting

ORIGINAL IMAGE



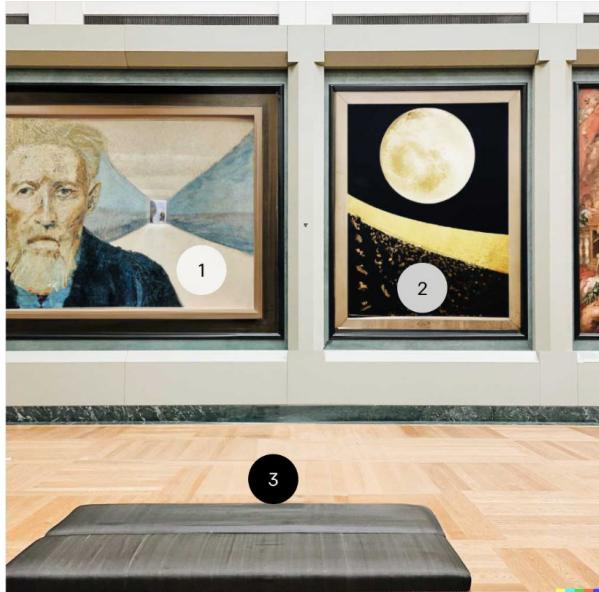
DALL-E 2 EDITS



SELECT LOCATION TO ADD A CORGI

OpenAI DALL-E 2 - Inpainting

ORIGINAL IMAGE



DALL-E 2 EDITS

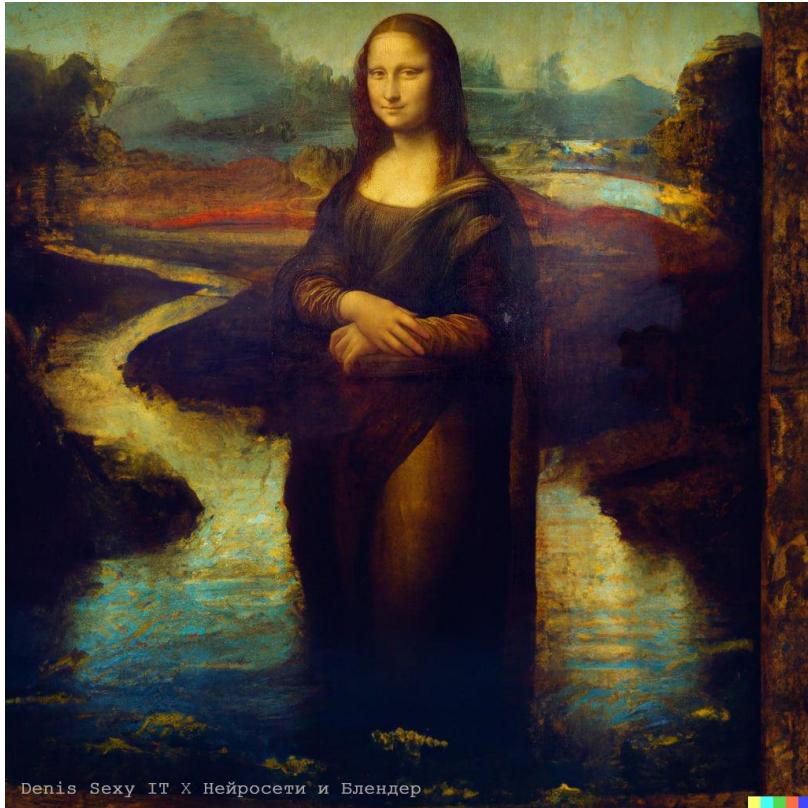


SELECT LOCATION TO ADD A CORGI

OpenAI DALL-E 2 - Inpainting

- Проект по "дорисовке" шедевров мировой живописи
- <https://t.me/denissexxy/5824>

OpenAI DALL-E 2 - Inpainting



Denis Sexy IT X Нейросети и Блендер



Denis Sexy IT X Нейросети и Блендер

Google IMAGEN

- Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding (2022)
- Text2Image от Google на основе диффузионных моделей
- Говорят, что лучше DALL-E 2 на "человеческих" бенчмарках

AI & Art 2022 - кого читать



- [Мишин Лернинг](#)
- [Derp Learning](#)
- [Нейросети и блендер](#)
- [\(бесценный\) Yannick Kilcher](#)



Захотелось лекцию про CLIP & DALL-E.

- Zero-shot learning?
- Соединение текста и изображения в одной модели?
- Просто посмотреть картинки?

Захотелось лекцию про CLIP & DALL-E.

- Zero-shot learning?
- Соединение текста и изображения в одной модели?
- Просто посмотреть картинки?

dog wearing hat says thank you

DRAW

