cv and ml

multimodal

Владимир Глазачев
cv в rosebud.ai

модальности

- Умеем работать с структурированными данными
  - таблички = матрицы

- Немного умеем работать с текстами переводя их в структурированный вид
  - мешки слов?

- Умеем работать с картинками и dl подходом
  - весь дискриминативный dl про feature learning
  - умеем решать базовые задачи, классификацию и сегментацию
  - немного посмотрели img2img и генерацию

модальности

- Умеем работать с структурированными данными
  - таблички = матрицы

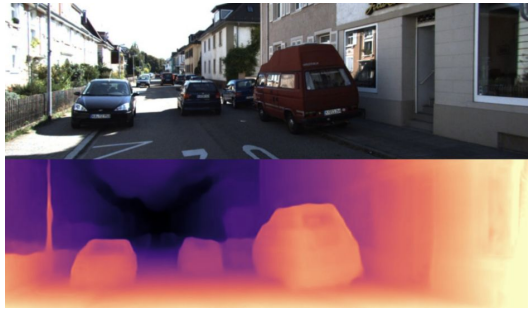
- Немного умеем работать с текстами переводя их в структурированный вид
  - мешки слов?

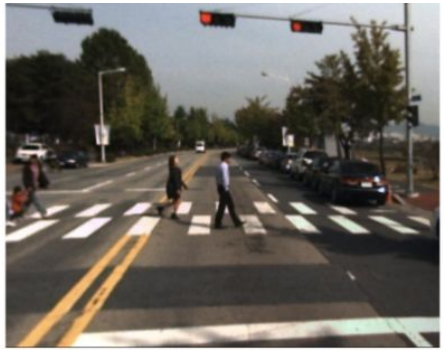
- Умеем работать с картинками и dl подходом
  - весь дискриминативный dl про feature learning
  - умеем решать базовые задачи, классификацию и сегментацию
  - немного посмотрели img2img и генерацию

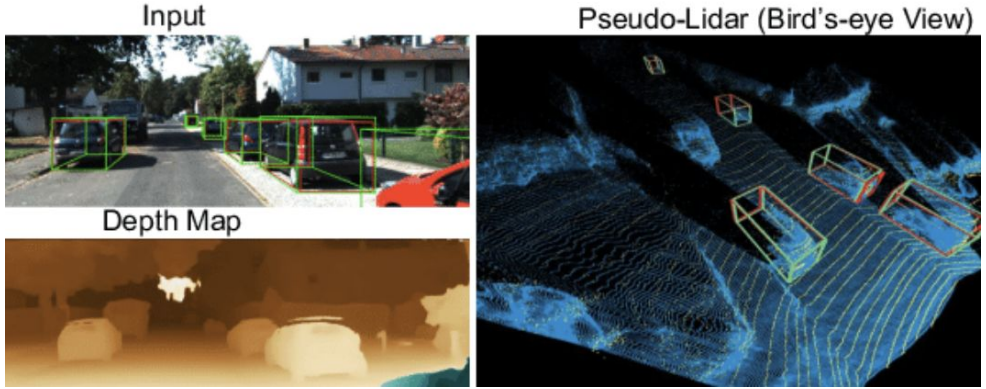
- звук?
- видео?
- 3d?
- …

модальности

- Формально для изображений - разные источники = разные модальности

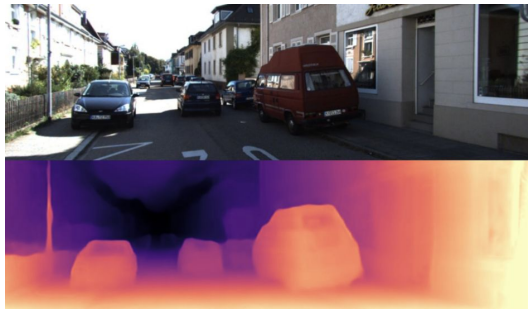


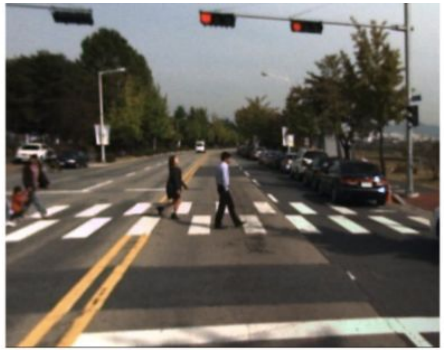


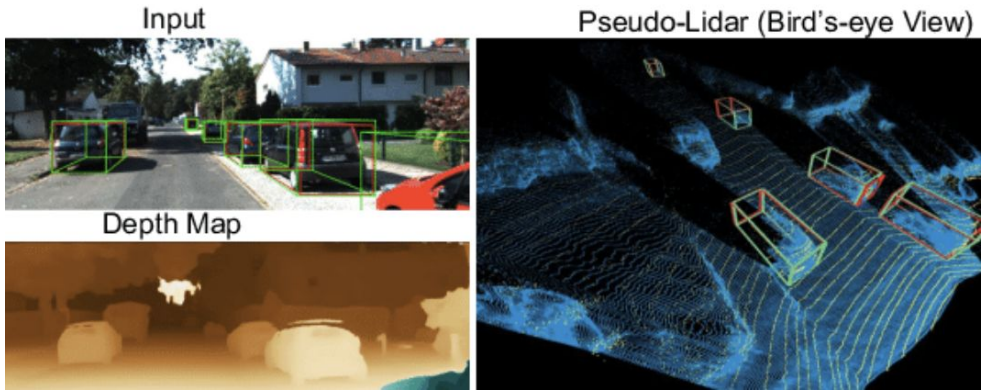


модальности

- Классическое применение - в self driving :)

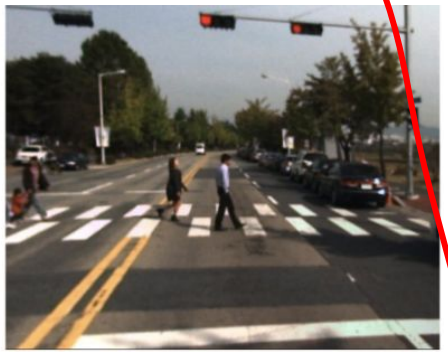


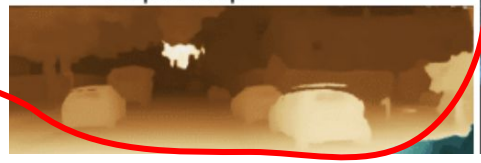


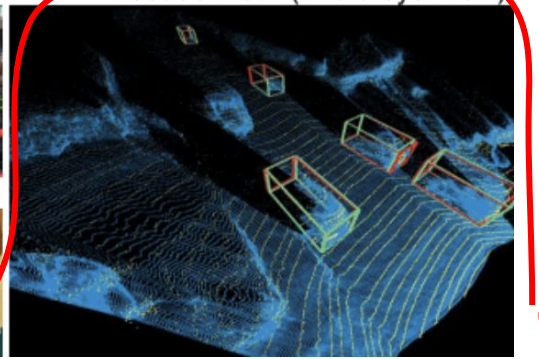


модальности

- Классическое применение - в self driving :)

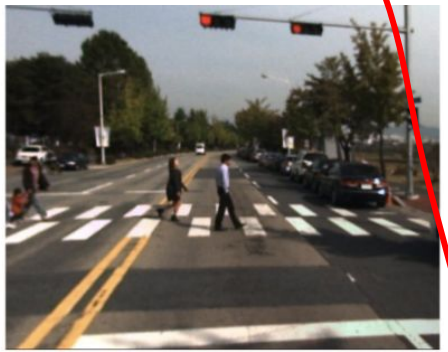


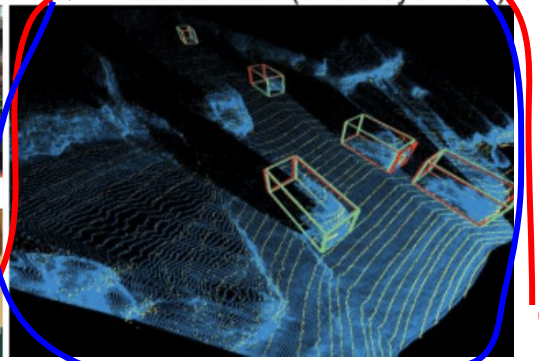


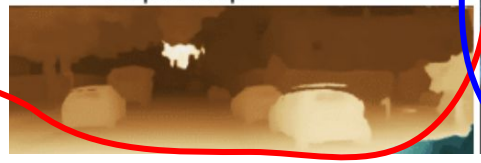


модальности

- Классическое применение - в self driving :)

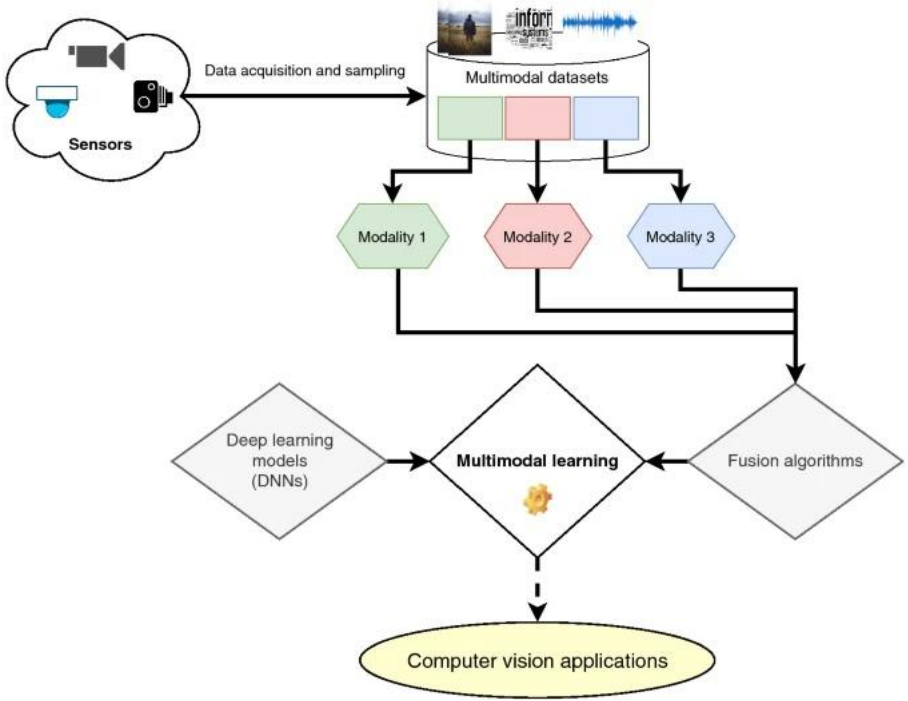






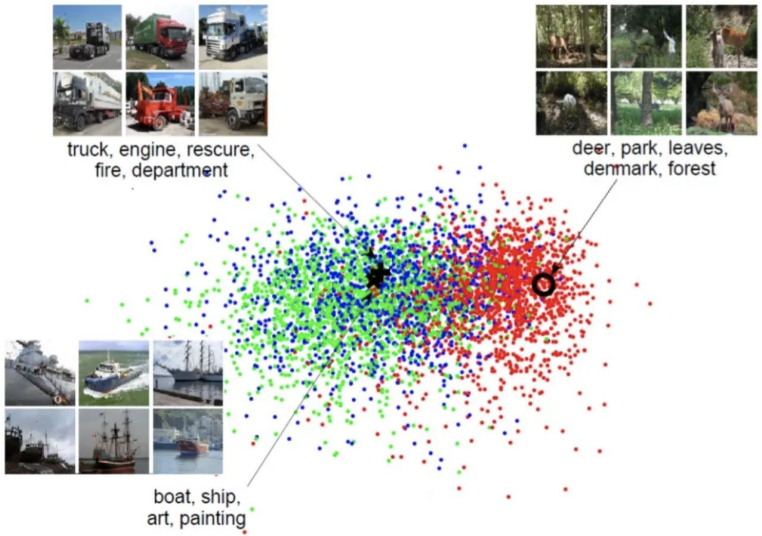
модальности

- Мы в картинках, так что нас интересует
  - image + tabular
  - image + text
  - image + image из другого домена (глубина например) не очень интересно - это все еще картинки так что хоть у них и разная природа, ничего не мешает их склеить в один тензор
  - image + speech / sounds
  - image + 3d?
  - …



типы задач

- representation
  - хотим извлечь разумные фичи (encode) из данных - могут быть в одном (как на примере) или в несвязных пространствах



типы задач

- translation
  - хотим изменить модальность
  - типичные задачи - image captioning

text: totally fine hand

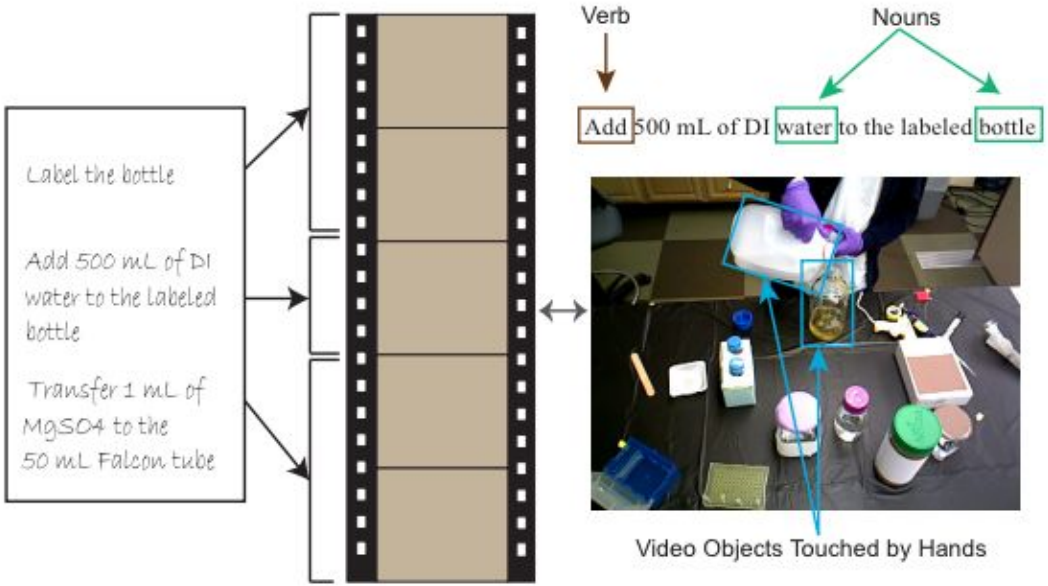
типы задач

- translation
  - хотим изменить модальность
  - типичные задачи - image generation from text

text: totally fine hand

типы задач

- alignment
  - выравниваем домены



примеры

- выравнивание субтитров
- выравнивание аудиокниг к тексту
- lip sync
- разметка временных данных

типы задач

- fusion
  - смешиваем домены чтобы решать задачу

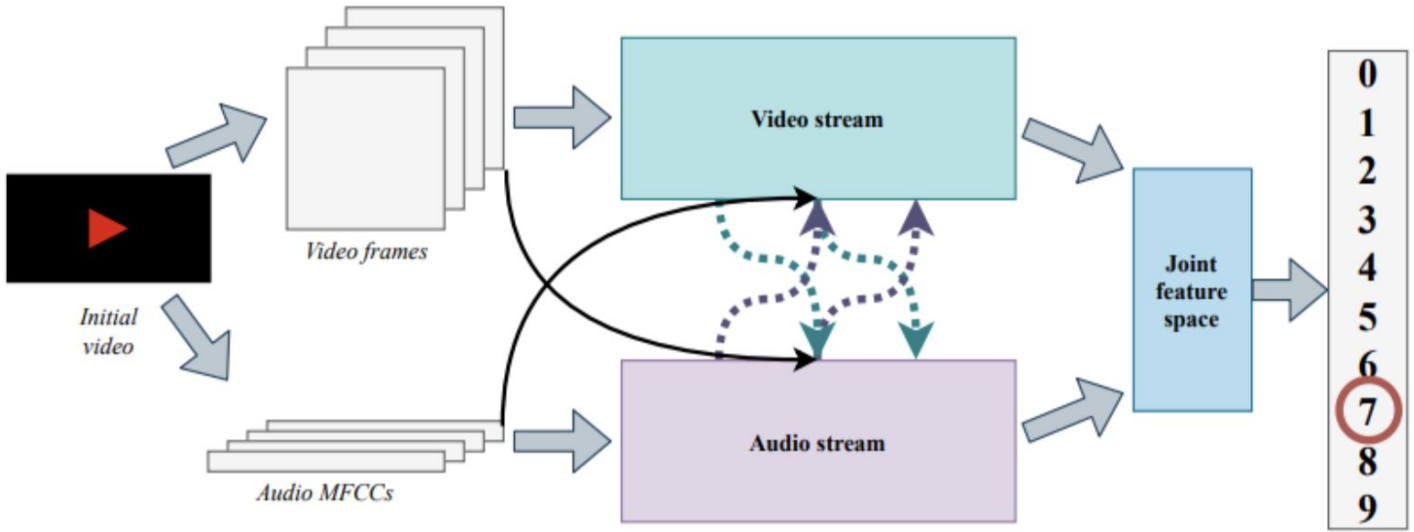


Fig: Multimodal classification system to identify which letter/digit a person is saying

типы задач

- co-learning
  - хотим использовать знания в одном домене для улучшения качества в другом

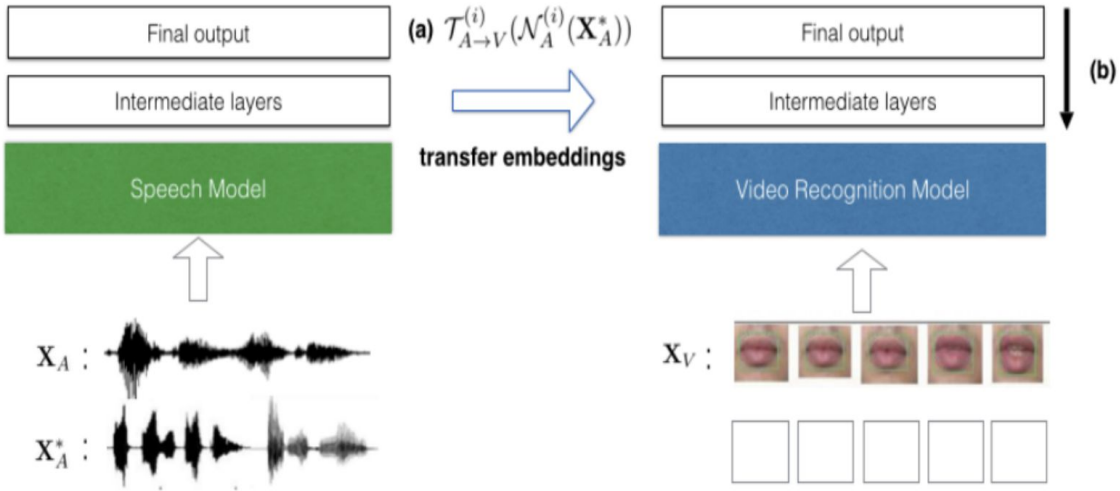


Fig: Transferring knowledge from speech recognition model to visual recognition model

Хорошо сочетается с синтетическими данными

image + tabular

- решаем классификацию или регрессию = representation + fusion

Два пути:

- свести задачу к табличной и решать как умеем через традиционный ml (бустингом например)
- использовать табличные данные в dl модели вместе с изображением

image + tabular

- свести задачу к табличной и решать как умеем через традиционный ml (бустингом например)

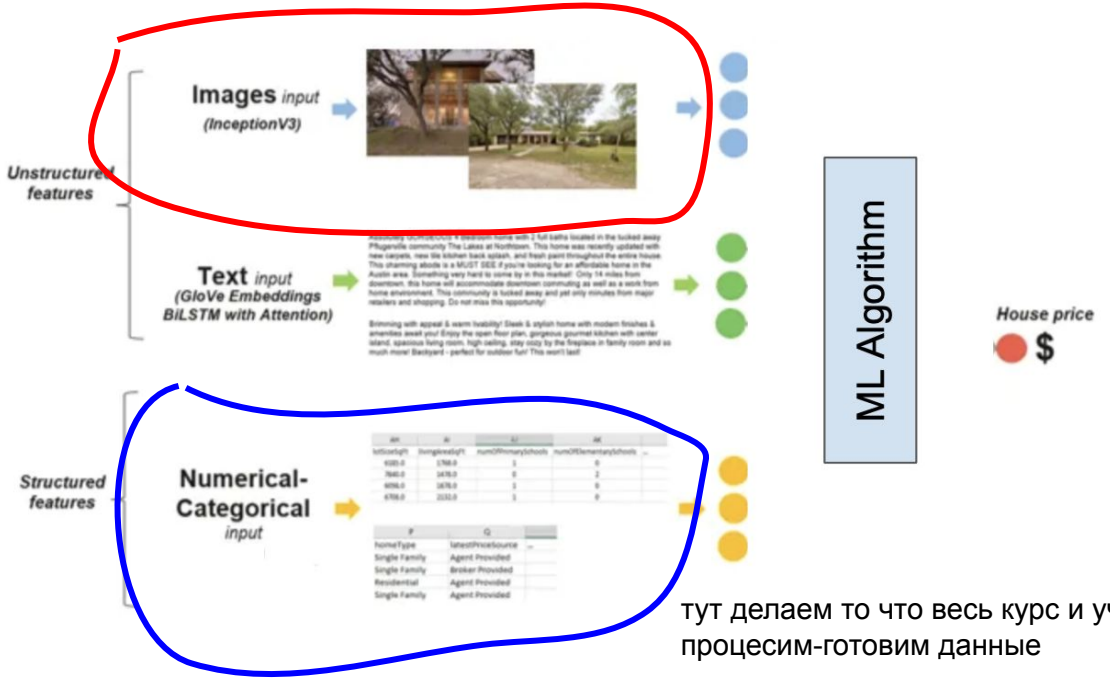
image + tabular

- свести задачу к табличной и решать как умеем через традиционный ml (бустингом например)

как мы помним dl во многом про feature extraction, так что давайте эти фичи и использовать

image + tabular

- свести задачу к табличной и решать как умеем через традиционный ml (бустингом например)



тут делаем то что весь курс и учились,
процесим-готовим данные

image + tabular

- свести задачу к табличной и решать как умеем через традиционный ml (бустингом например)

может быть любой ml метод

!!! наши картиночные/текстовые feature extractor тут не обучаются, а используются уже обученные !!!

image + tabular

- свести задачу к табличной и решать как умеем через традиционный ml (бустингом например)

плюсы:

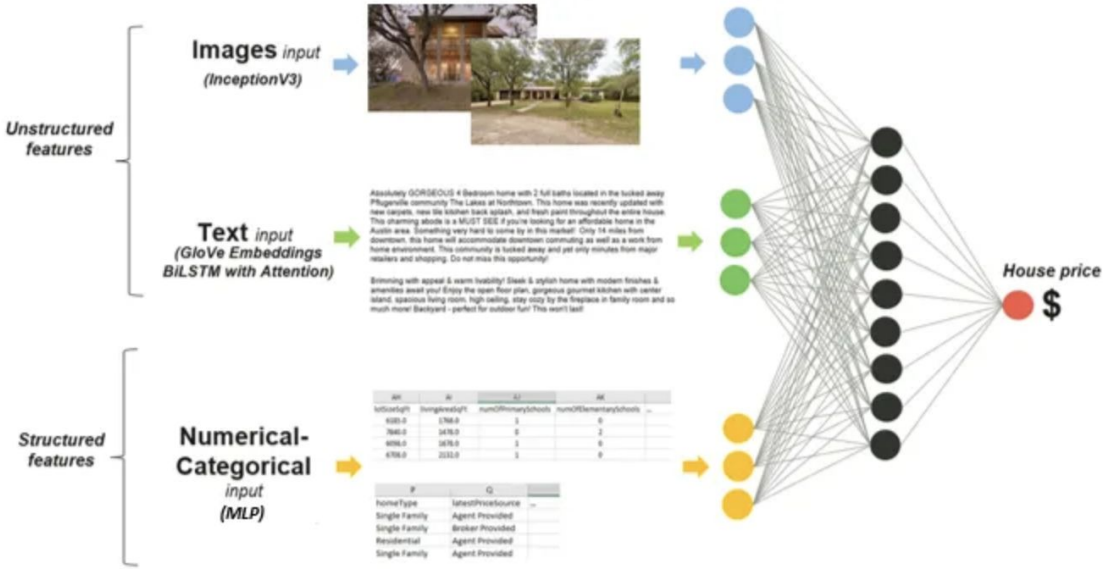
- почти бесплатно добавили модальность
- можно легко синтегрировать в ваш готовый табличный пайплайн

минусы:

- фич будет очень много (500+ минимум от картиночного энкодера, для текстов такой же порядок ; могут быть и тысячи)
- если ваши данные out of distribution готовых энкодеров - работать будет плохо
- енкодеры не обучаются в процессе
- надо зоопарк моделей поддерживать

image + tabular

- использовать табличные данные в dl модели вместе с изображением



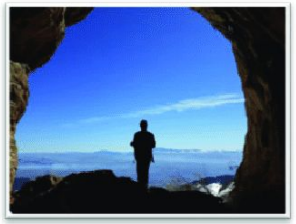
ничего не мешает нам поставить в качестве ml алгоритма какую нибудь dl модель

все становится дифференцируемо и можно обучать все вместе

- из коробки будет работать хуже чем первый метод :)
традиционные ml модели (бустинги) все еще сложно побить dl методами в табличных данных - но можно

img2text

- хотим описать что происходит на изображении



It's a man standing on a rocky hill.

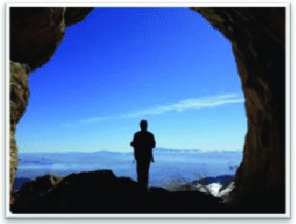
It's a large white building with Azadi Tower in the background.

It's a bowl of food on a table.

It's a wooden statue in a park.

img2text

- хотим описать что происходит на изображении
  - representation + translation



It's a man standing on a rocky hill.

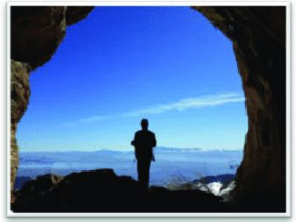
It's a large white building with Azadi Tower in the background.

It's a bowl of food on a table.

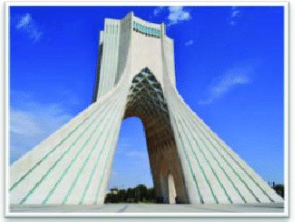
It's a wooden statue in a park.

img2text

- хотим описать что происходит на изображении
  - representation + translation



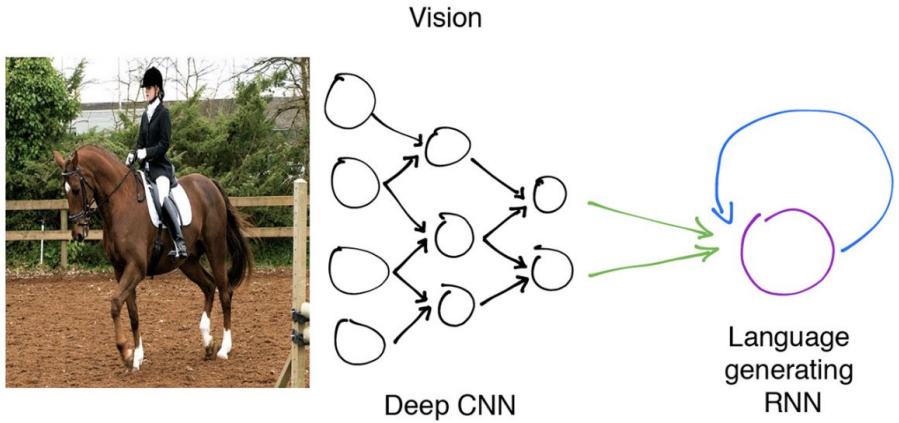
It's a man standing on a rocky hill.



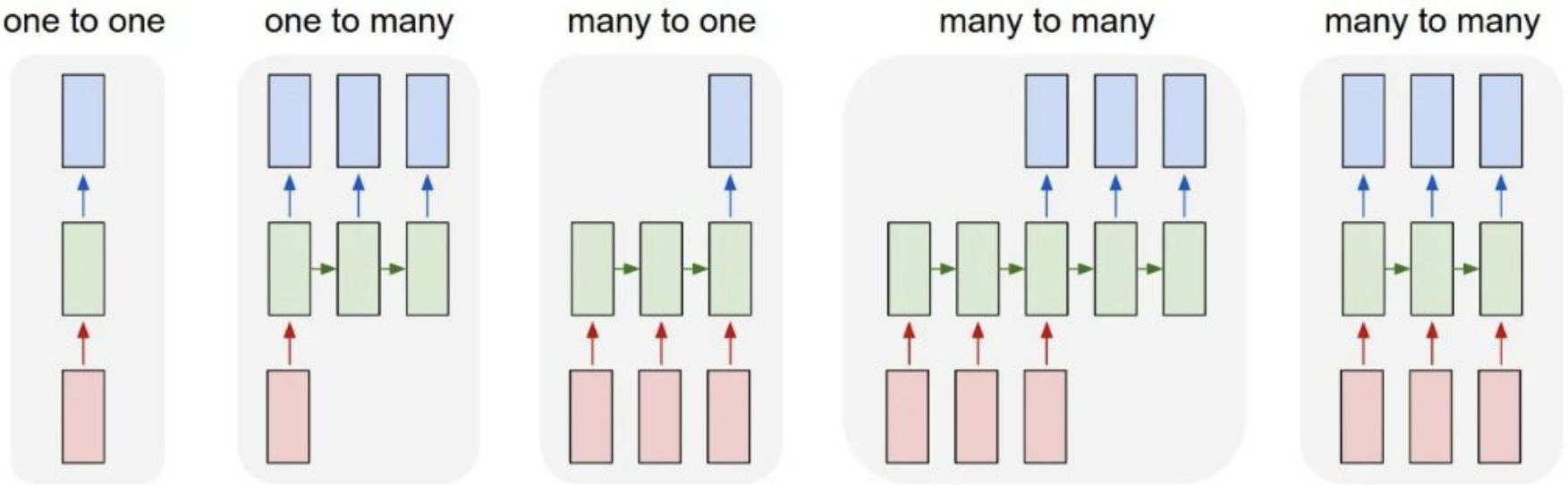
It's a large white building with Azadi Tower in the background.

It's a bowl of food on a table.

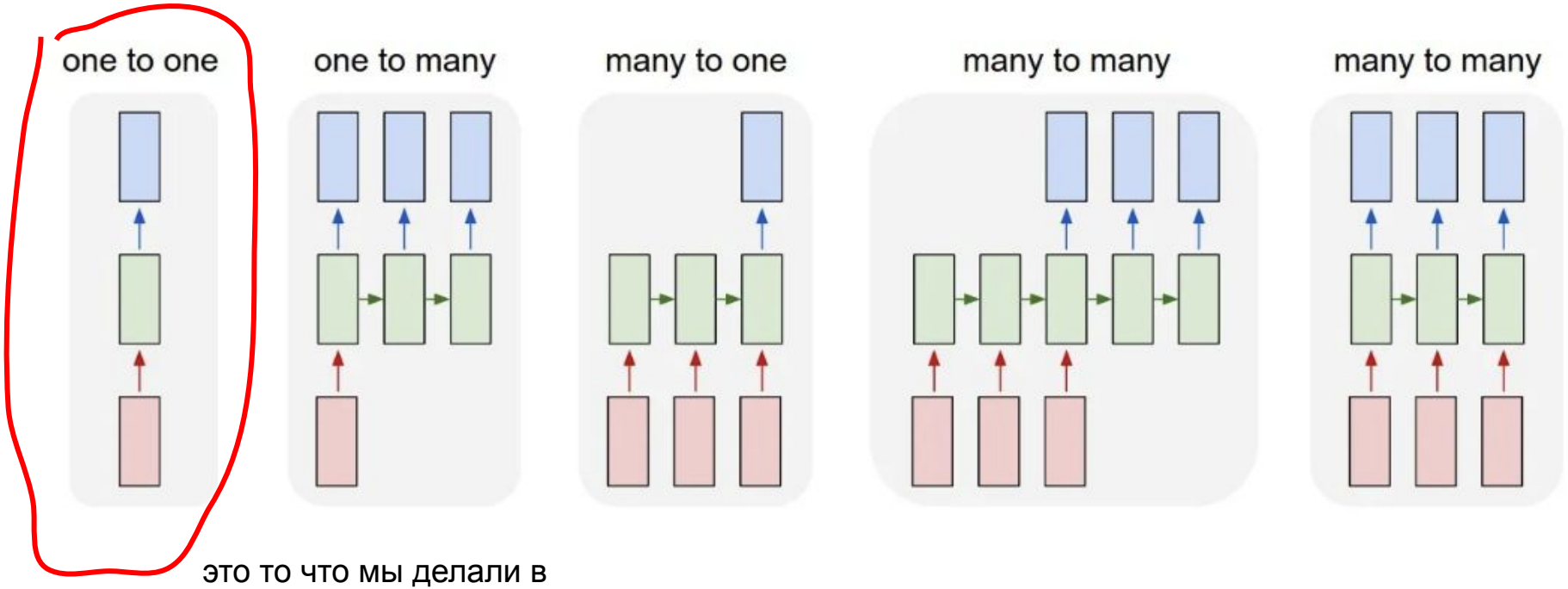
It's a wooden statue in a park.



sequence modeling за 5 минут

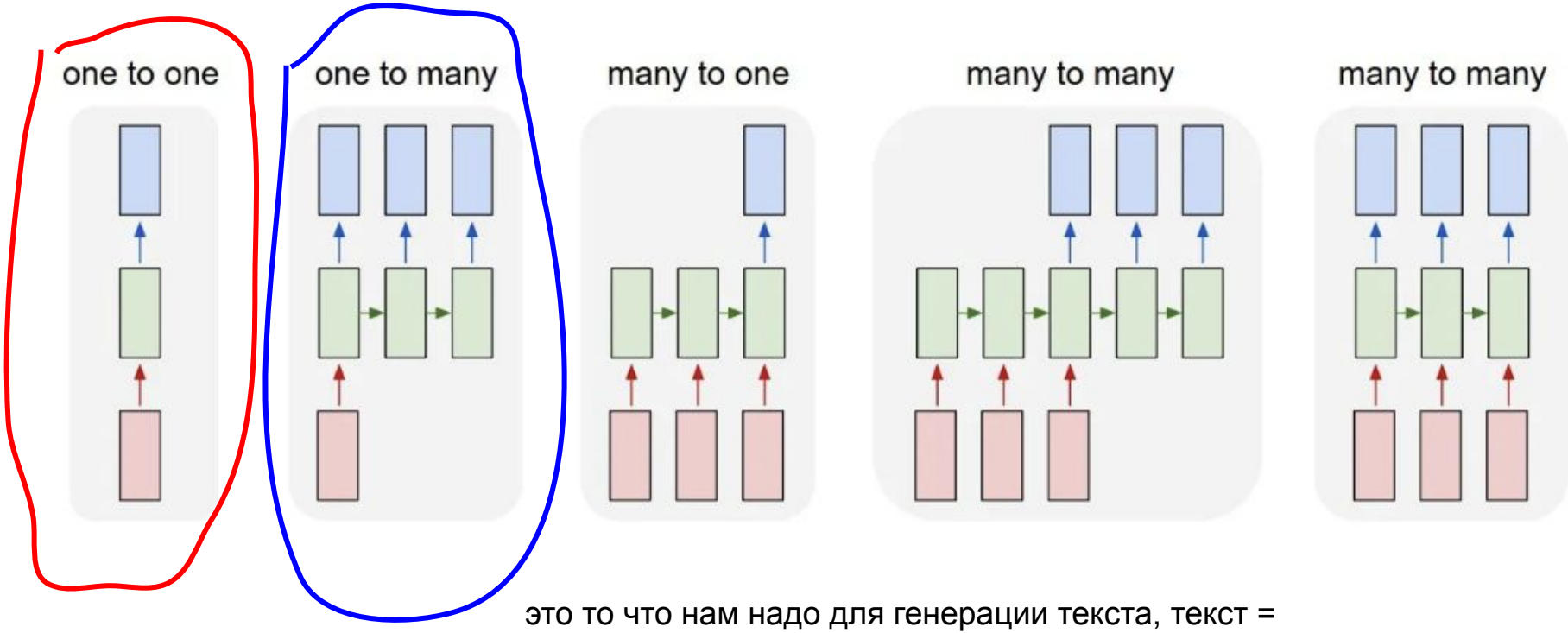


sequence modeling за 5 минут



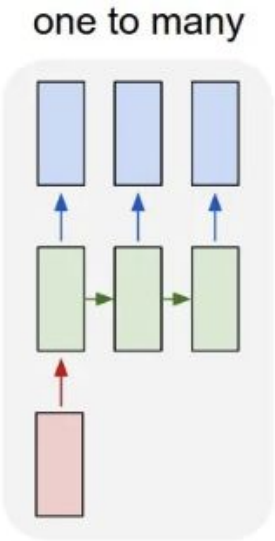
это то что мы делали в предыдущих задачах

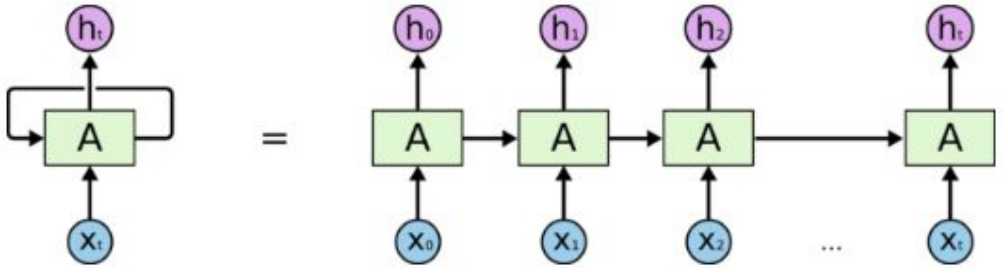
sequence modeling за 5 минут



это то что нам надо для генерации текста, текст = последовательность рандомной длины

sequence modeling за 5 минут





An unrolled recurrent neural network.

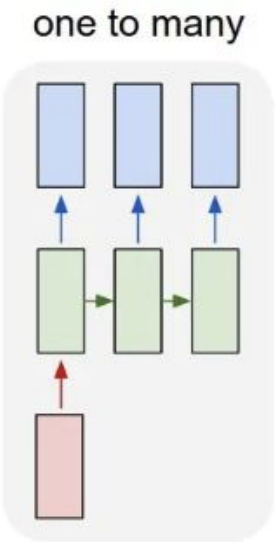
A - модель

- принимает на вход состояние z_i-1 и x_i

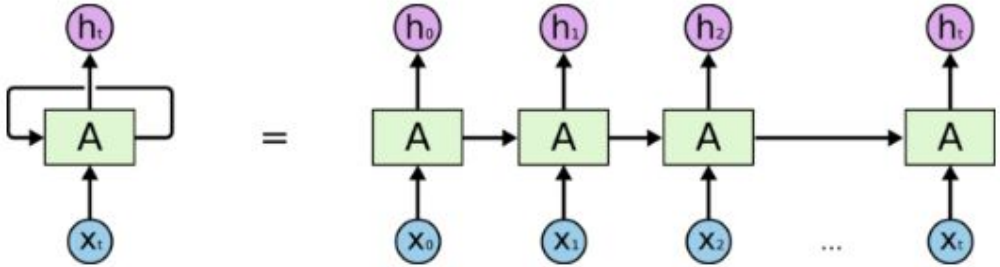
возвращает внутреннее состояние z_i и предсказание h_i

- x_i = x_0
- A - одна и та же на каждом шаге
- нам нужны все h_i

sequence modeling за 5 минут



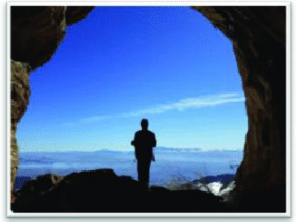
- Предсказываем вероятность следующего токена (например буквы), loss - класификационный
- Добавляем спец символ = end token, x_0 = рандом или какой то condition
- while h != end token предсказываем следующий токен



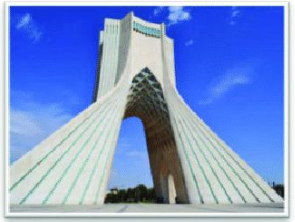
An unrolled recurrent neural network.

img2text

- хотим описать что происходит на изображении
  - representation + translation



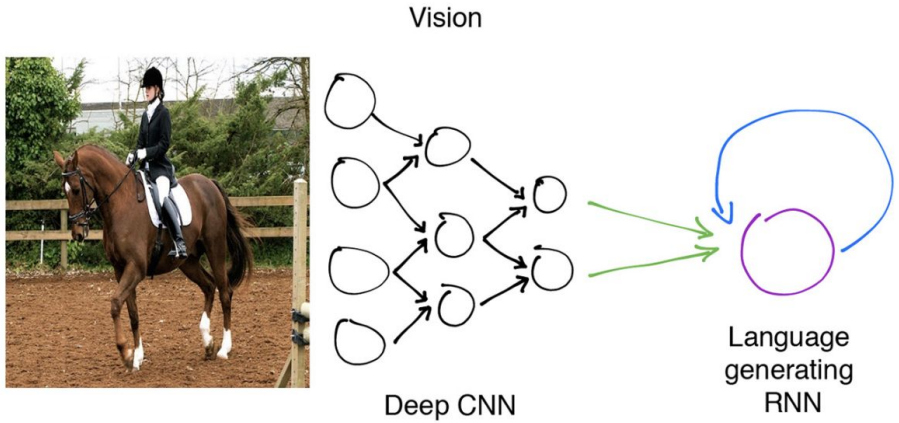
It's a man standing on a rocky hill.



It's a large white building with Azadi Tower in the background.

It's a bowl of food on a table.

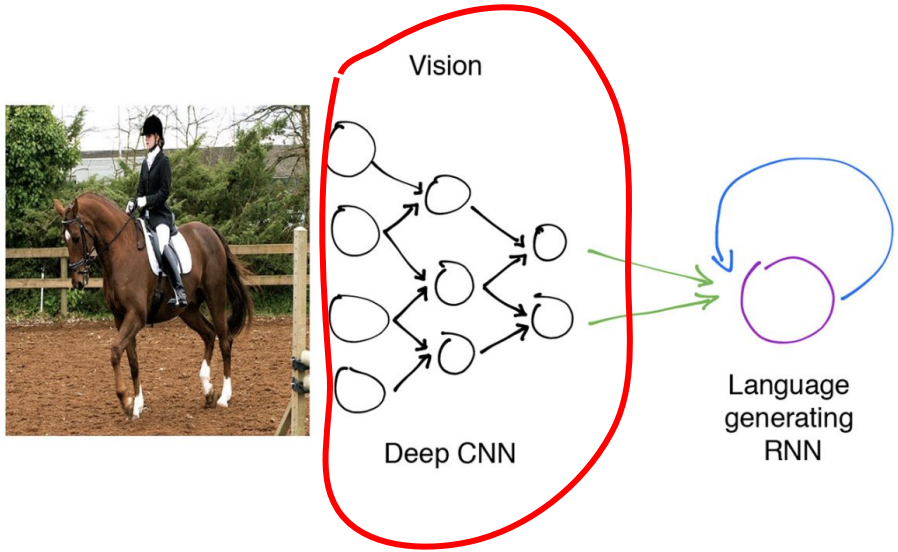
It's a wooden statue in a park.



img2text

- хотим описать что происходит на изображении
  - representation + translation

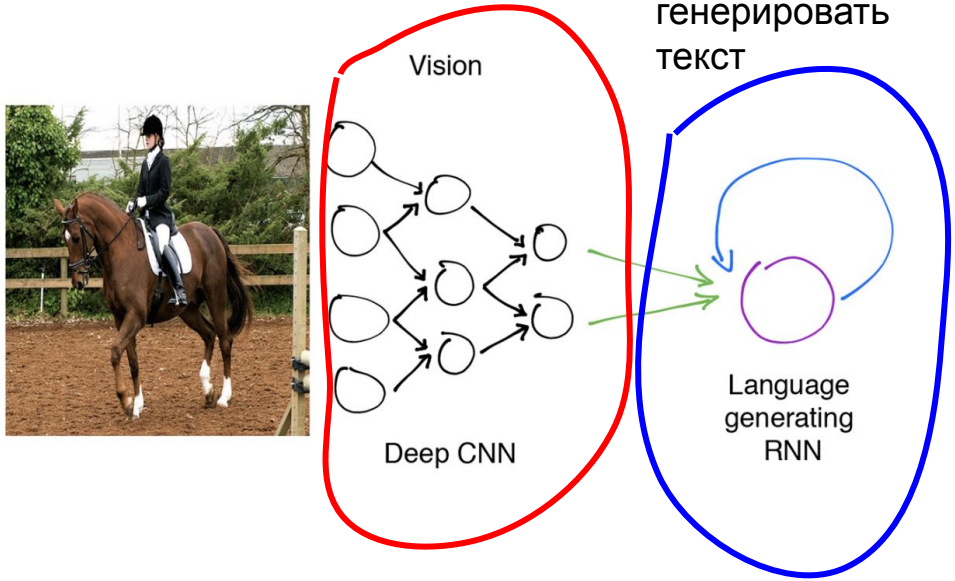
тут берем какой нибудь предобученный vgg / resnet / что угодно



img2text

- хотим описать что происходит на изображении
  - representation + translation

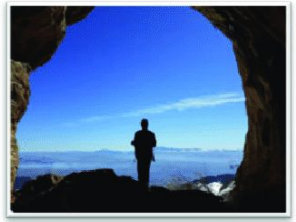
обучаем модель генерировать текст



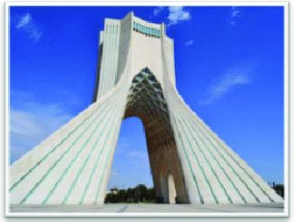
img2text

Воспринимать как псевдо-описание алгоритма, в актуальных архитектурах у вас не будет RNN а будет language model на трансформерах

- хотим описать что происходит на изображении
  - representation + translation



It's a man standing on a rocky hill.

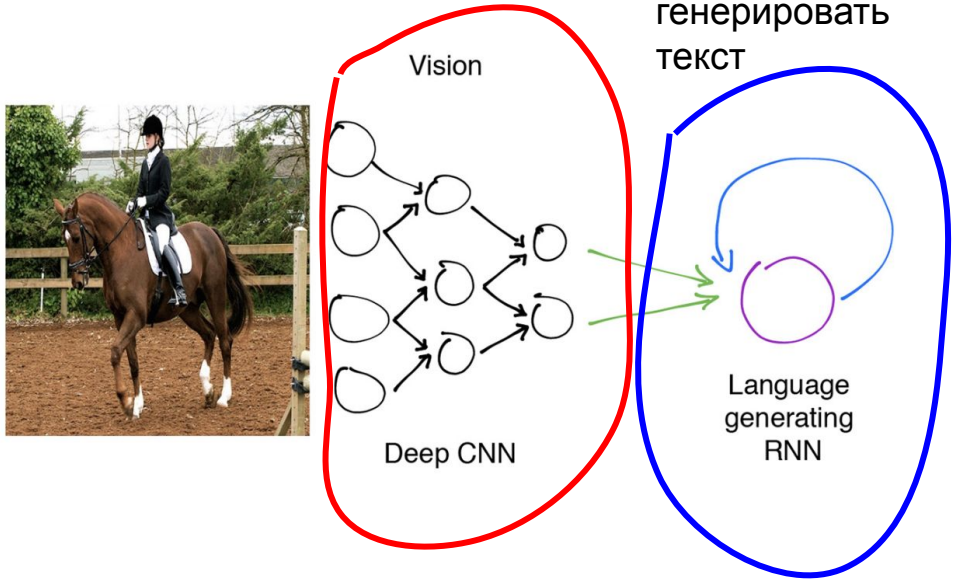


It's a large white building with Azadi Tower in the background.

It's a bowl of food on a table.

It's a wooden statue in a park.

обучаем модель генерировать текст

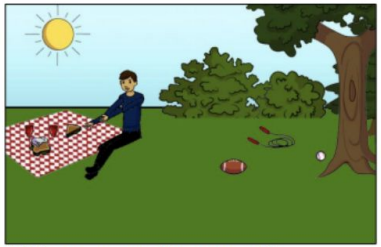


img+text2text

- хотим задавать вопросы по изображению
  - representation + translation + maybe alignment

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



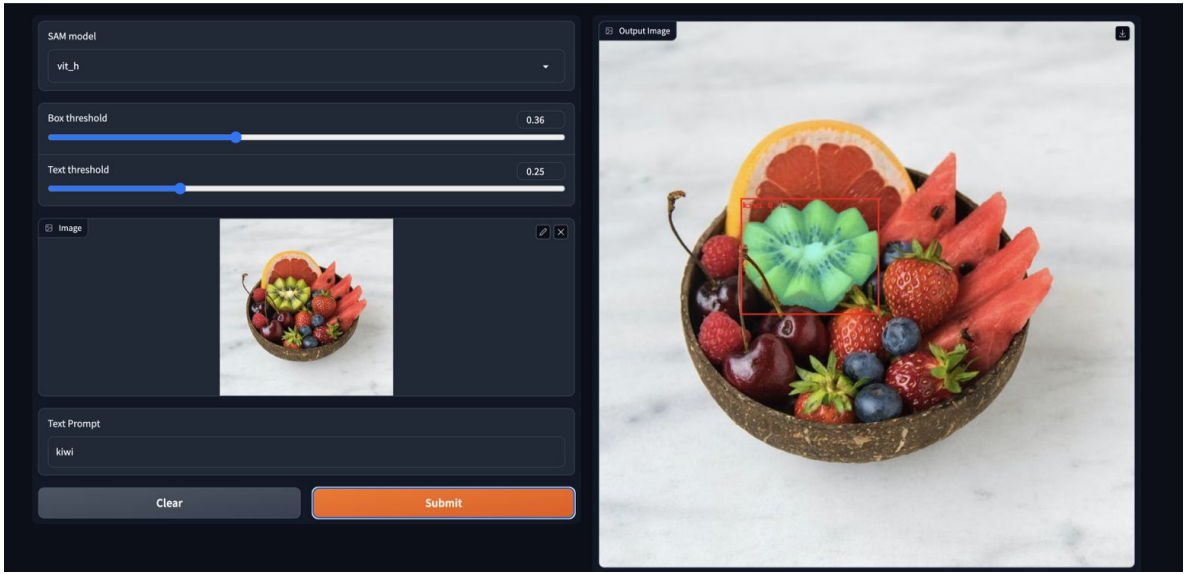
Is this person expecting company?
What is just under the tree?

question: how many dogs are in the picture?
answer: 1

- Задача называется visual question answering (VQA)
  - https://huggingface.co/spaces/OFA-Sys/OFA-vqa - не SOTA но можно потыкать тут

img+text2detection

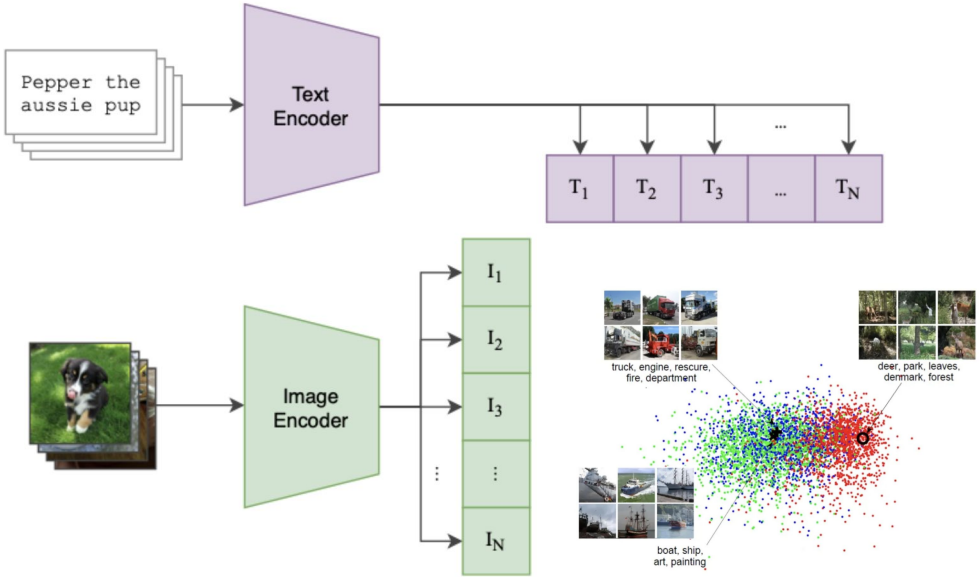
- хотим находить объекты, не обучали классификацию в явном виде
  - representation + translation + alignment



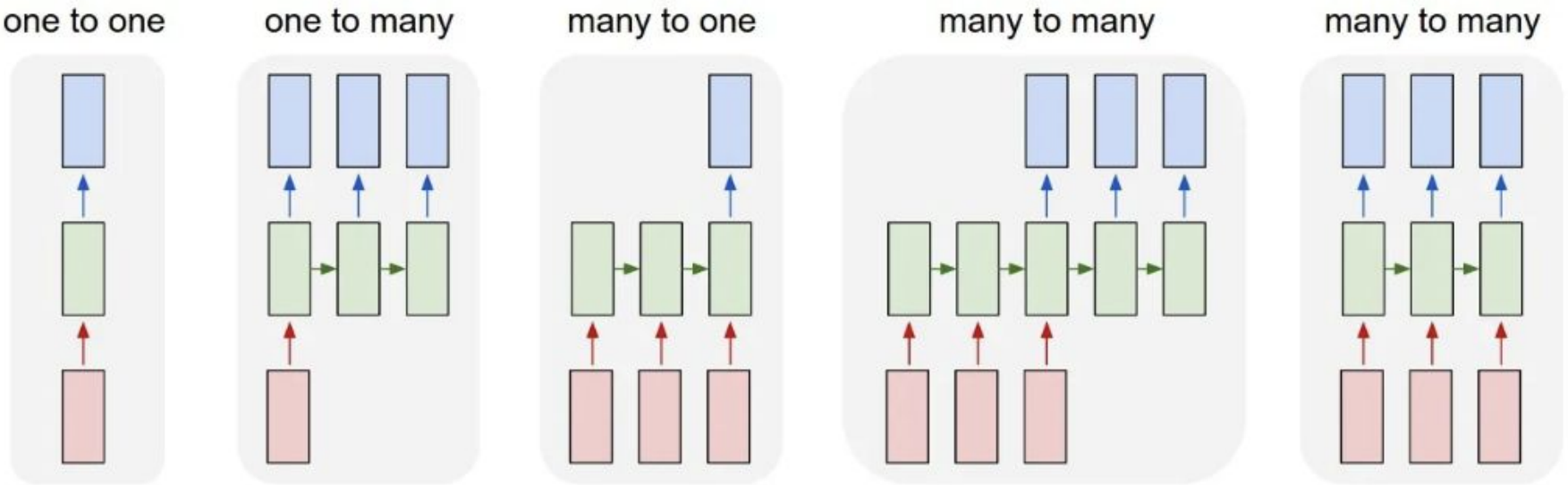
- https://huggingface.co/spaces/jbrinkma/segment-anything - SOTA, можно потыкать тут, но текст prompt кажется отключен
- https://github.com/facebookresearch/segment-anything
- https://github.com/luca-medeiros/lang-segment-anything

img and text2shared space

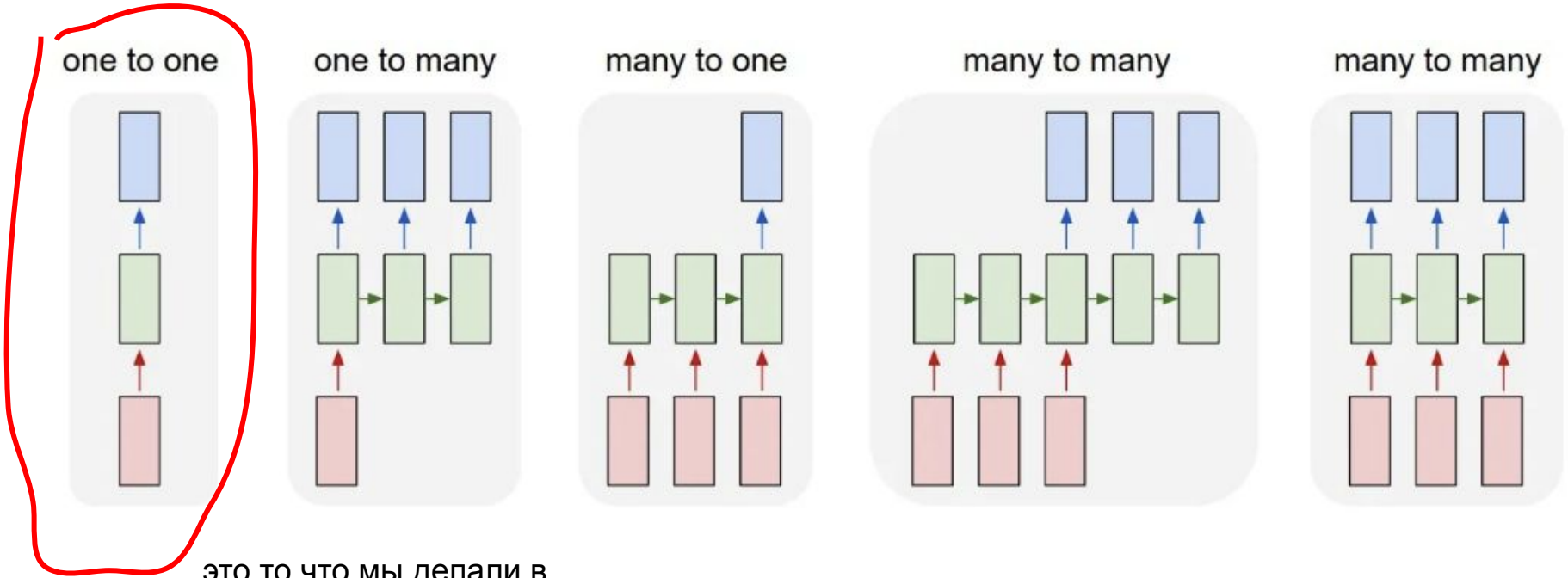
- хотим уметь сжимать текст и изображения в одно и тоже пространство
  - representation + co-learning



sequence encoding за 5 минут

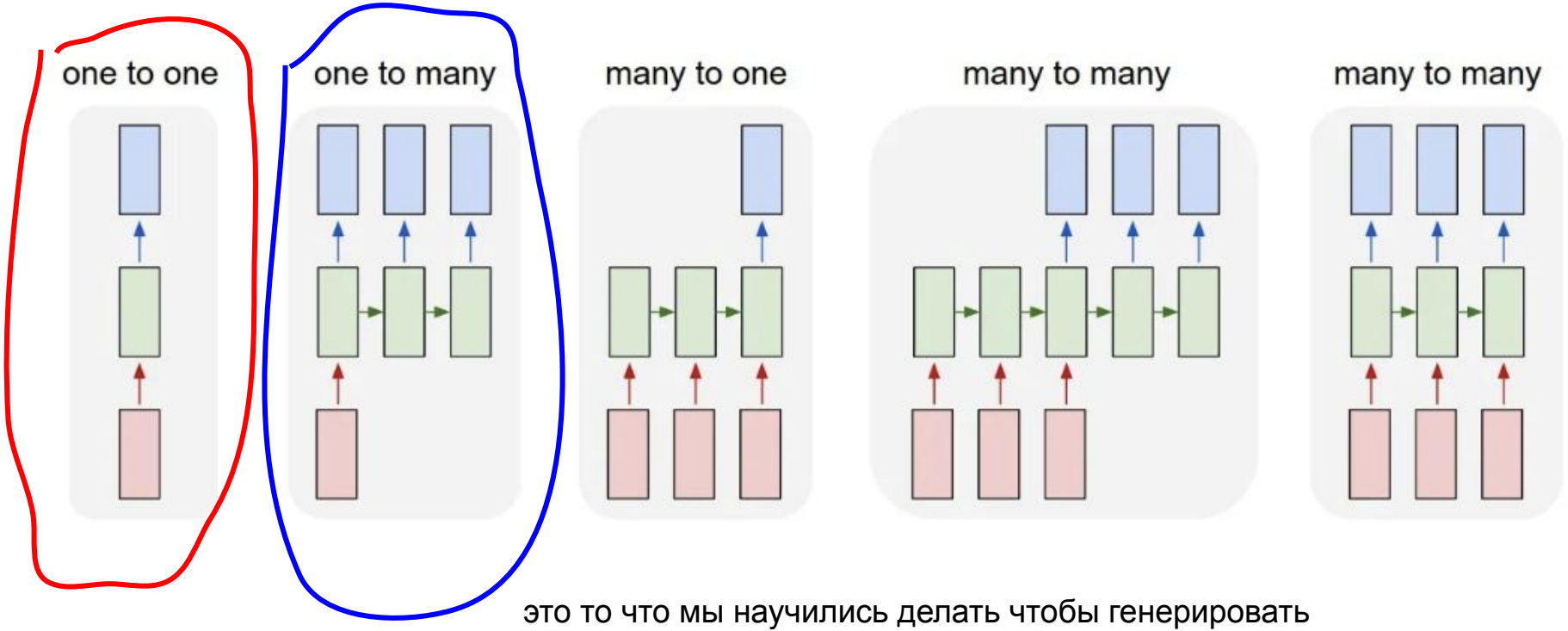


sequence encoding за 5 минут



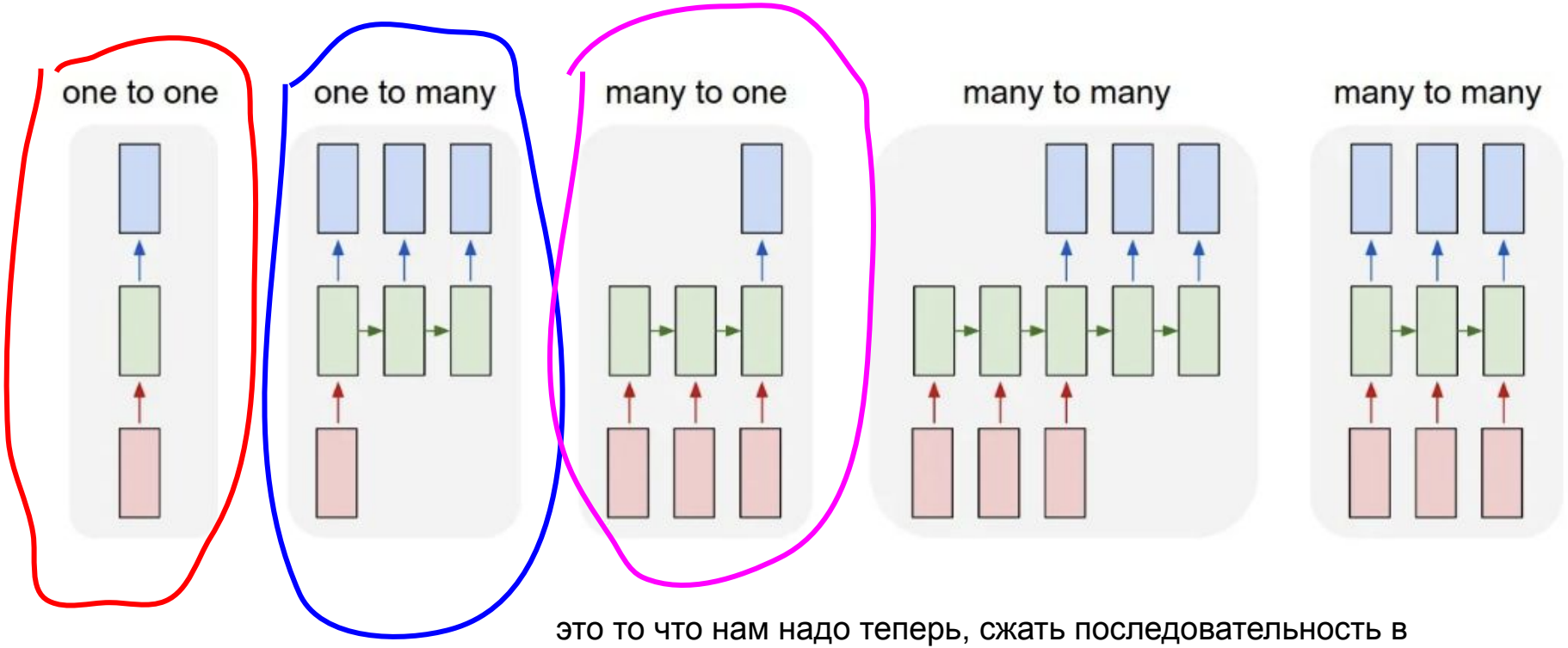
это то что мы делали в предыдущих задачах

sequence encoding за 5 минут



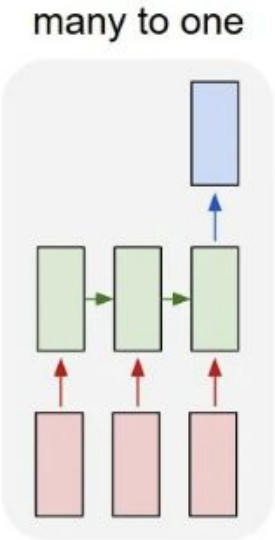
это то что мы научились делать чтобы генерировать последовательности

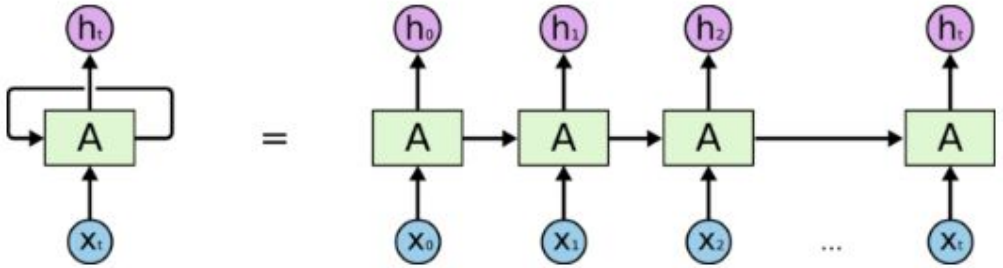
sequence encoding за 5 минут



это то что нам надо теперь, сжать последовательность в компактное представление фиксированной размерности

sequence encoding за 5 минут





An unrolled recurrent neural network.

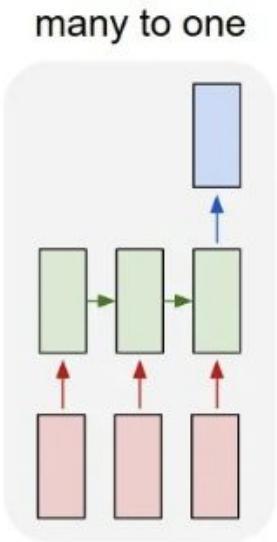
A - модель

- принимает на вход состояние z_i-1

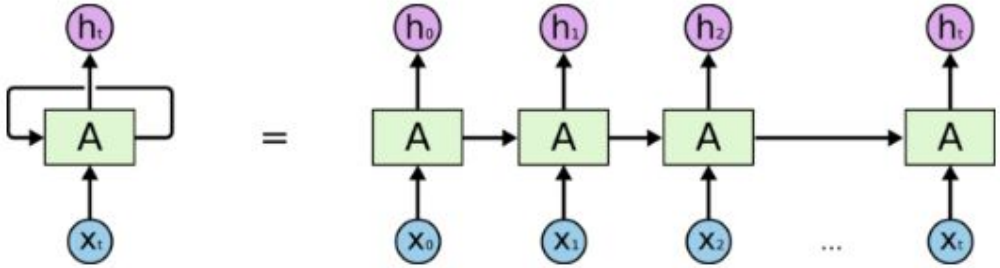
возвращает внутреннее состояние z_i и предсказание h_i

- z_0 = обучается
- A - одна и та же на каждом шаге
- нам нужен только h_t - самое последнее предсказание

sequence encoding за 5 минут



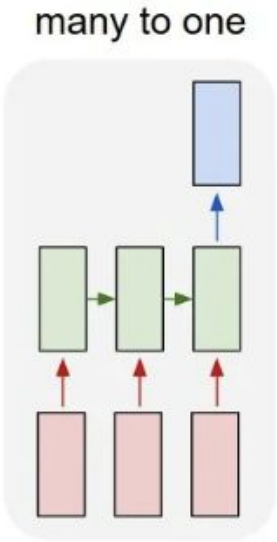
- Предсказываем какое то encoded состояние (например 512 float чисел)
- while есть x_i на входе делаем предсказание, когда закончились - возвращаем последнее h



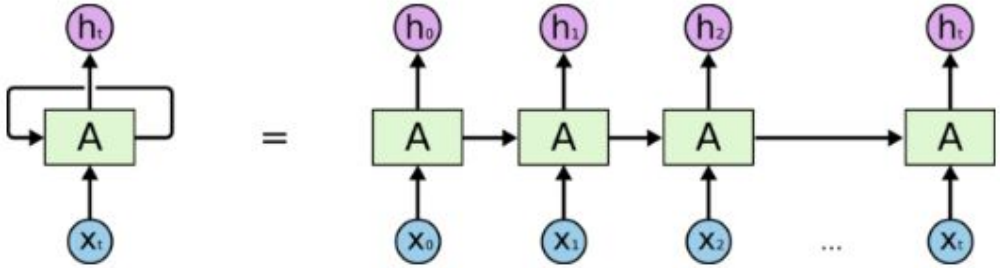
An unrolled recurrent neural network.

sequence encoding за 5 минут

Воспринимать как псевдо-описание алгоритма, в актуальных архитектурах у вас не будет RNN а будет language model на трансформерах



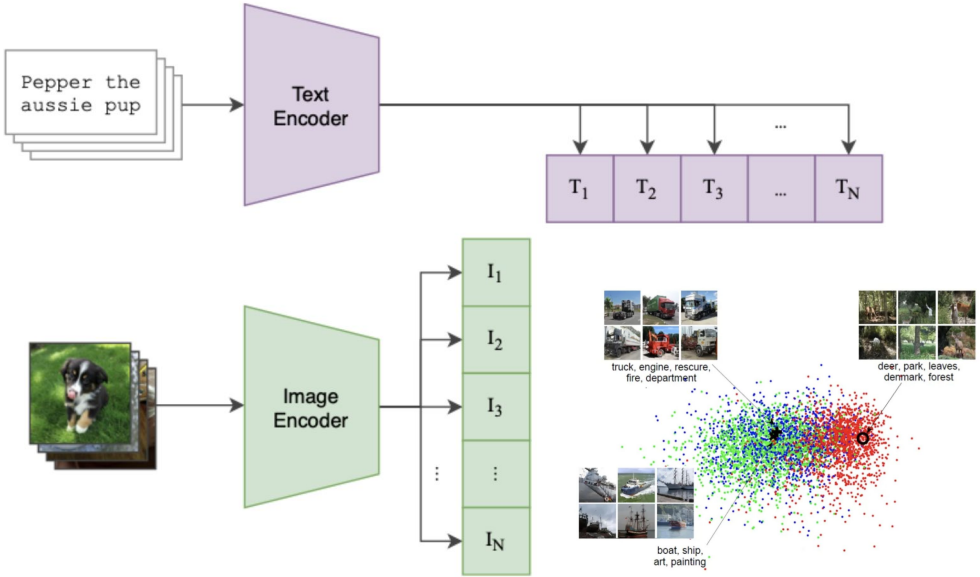
- Предсказываем какое то encoded состояние (например 512 float чисел)
- while есть x_i на входе делаем предсказание, когда закончились - возвращаем последнее h



An unrolled recurrent neural network.

img and text2shared space

- хотим уметь сжимать текст и изображения в одно и тоже пространство
  - representation + co-learning

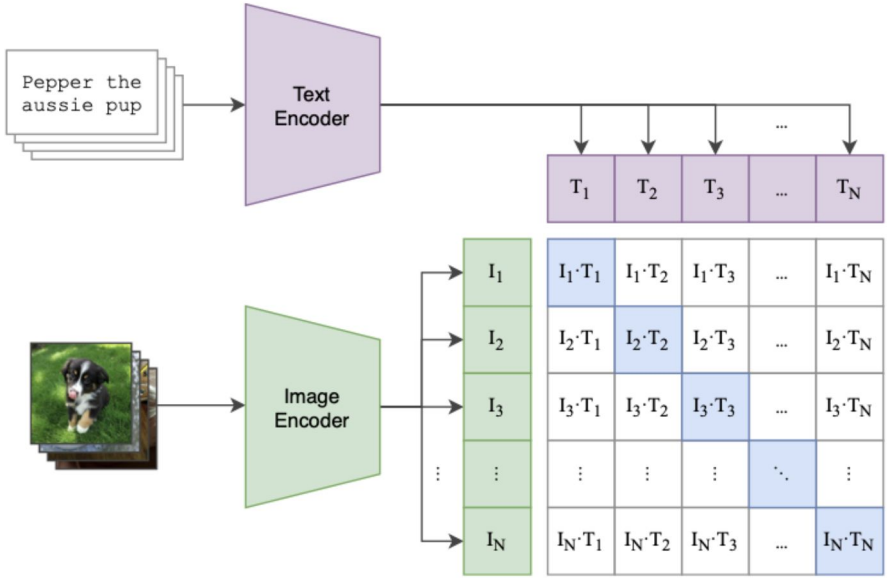


img and text2shared space

- хотим уметь сжимать текст и изображения в одно и тоже пространство
  - representation + co-learning

img and text2shared space

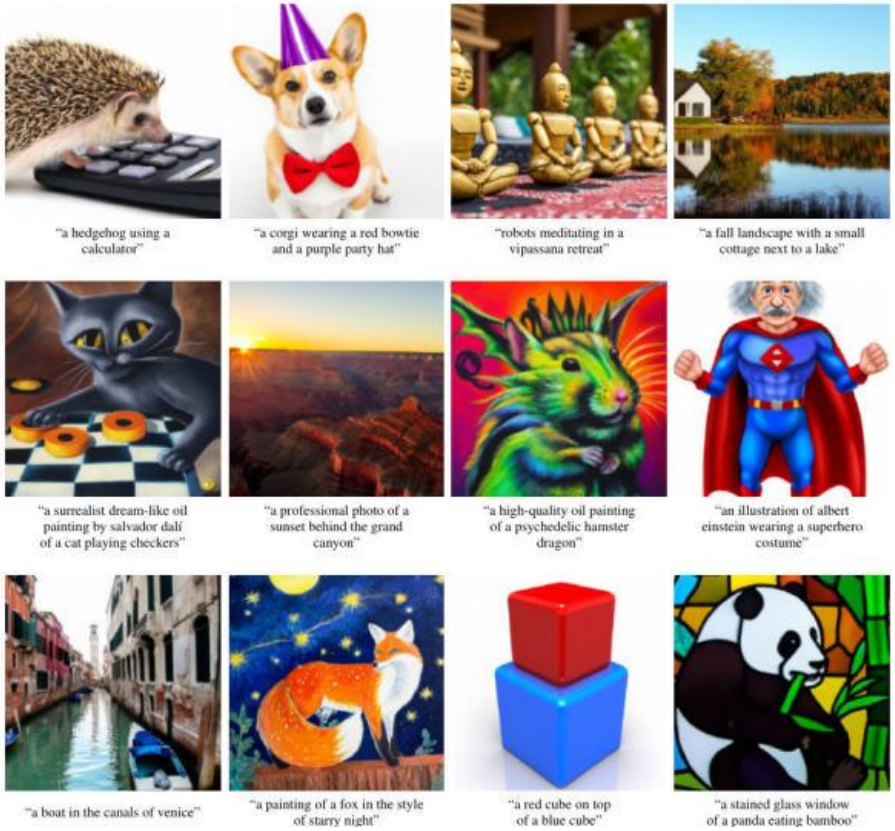
- хотим уметь сжимать текст и изображения в одно и тоже пространство
  - representation + co-learning



- sota - CLIP, https://openai.com/research/clip
- можно делать image search - https://rom1504.github.io/clip-retrieval
- можно использовать для few shot задач - knn или линейная модель поверх ембединга хорошо работает
- можно использовать как универсальный pre trained condition для задач где надо склеить текст и изображения

text2img

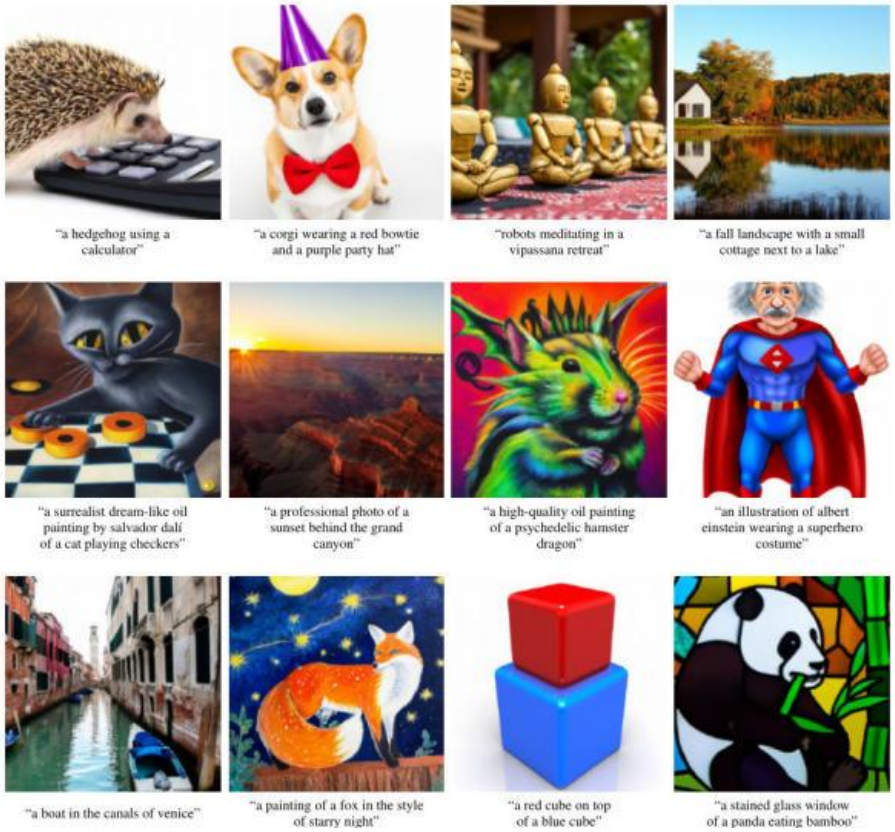
- finally - хотим по тексту генерить изображения
  - representation + translation + co-learning



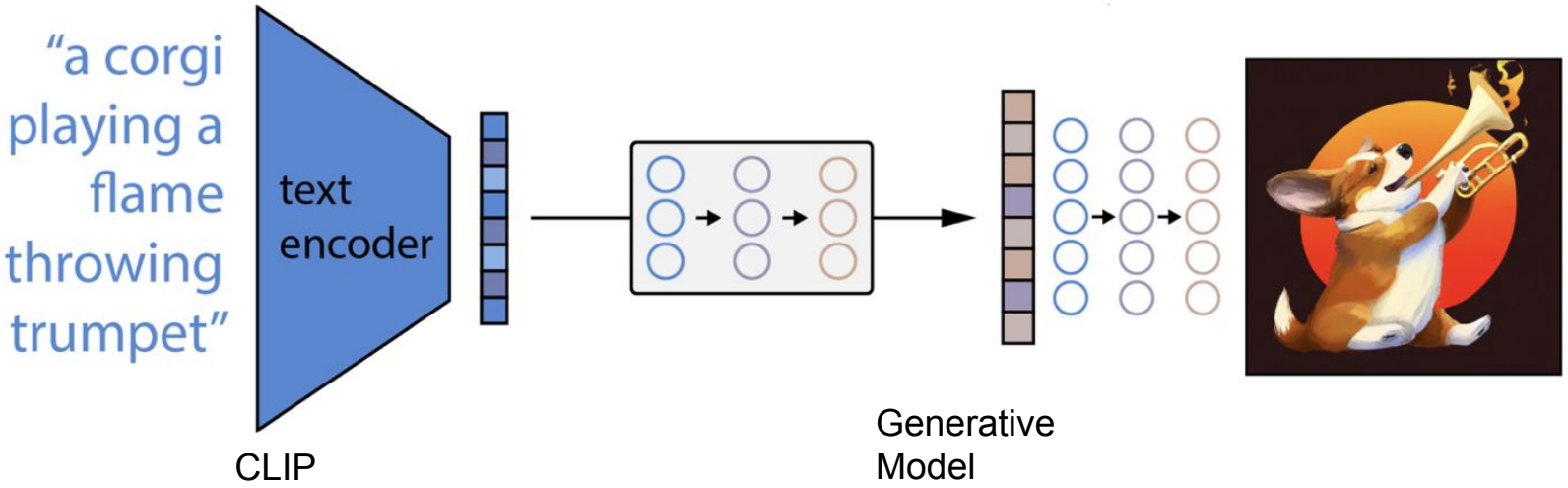
text2img

- finally - хотим по тексту генерить изображения
  - representation + translation + co-learning

- что нам нужно?
  - world knowledge model для текста
    - CLIP
  - датасет пар image + text
    - https://laion.ai/blog/laion-5b/
  - генеративная модель



text2img

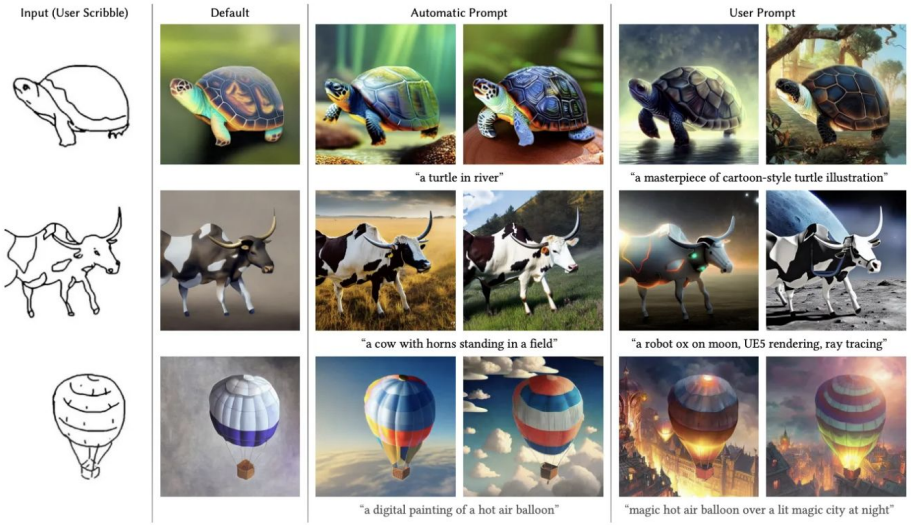


- https://www.assemblyai.com/blog/how-dall-e-2-actually-works/
- https://jalammar.github.io/illustrated-stable-diffusion/
- https://huggingface.co/spaces/stabilityai/stable-diffusion-1 - потыкать можно тут

bonus section - img+text2img

- текста не достаточно, хотим еще и дополнительный картиночный condition засунуть

bonus section - img+text2img



Мы смотрели как делать img2img ранее, тут надо совместить его с готовой генеративной моделью
Для этого надо разобраться как внутри работает conditioning, но это уже другая история :)
В этом примере - модель дообучается с большой text2image модели а не с нуля

https://github.com/lllyasviel/ControlNet-v1-1-nightly - оригинальный репо, https://huggingface.co/blog/controlnet - как проще можно запускать с colab примерами

Саммари

- multimodal - текущий hot topic
  - как склеить 3d, видео, картинки, звук и текстовые модели вместе? как упрощать интерфейсы (например через gpt генерить команды, выполнять их на картиночной модели)
- CLIP - одна из главных моделей последних лет
- Большие world knowledge модели важны
  - все сходится к тому что большинство задач обработки текстов и изображений можно решать через одну общую модель через дообучение или напрямую
- Продуктовые задачи на традиционном cv, nlp и ml (тут традиционным я уже называю не древнее ручное создание фич, а без использования world knowledge моделей) потому что это может быть дешевле, быстрее и качественнее - оверфит на домен полезен