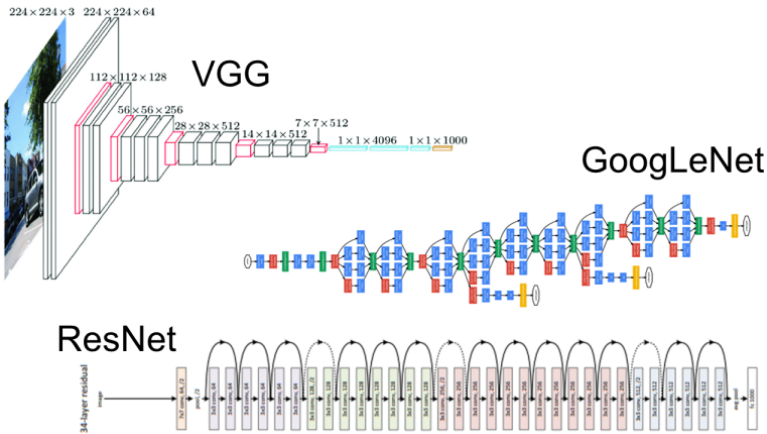
Transformers in CV

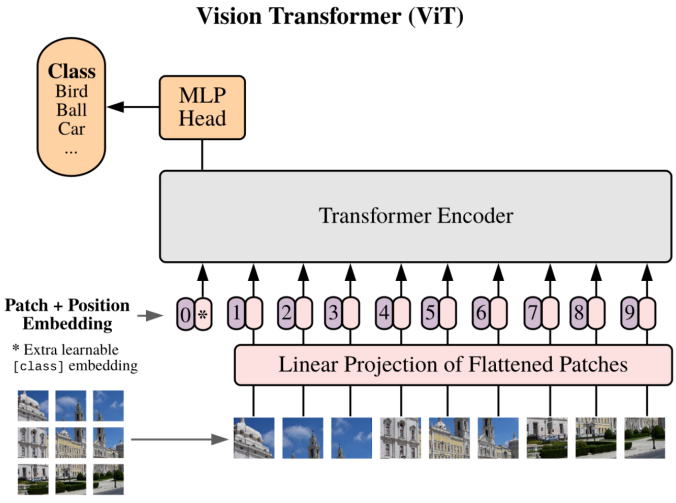
Никита Степанов

April 6, 2021

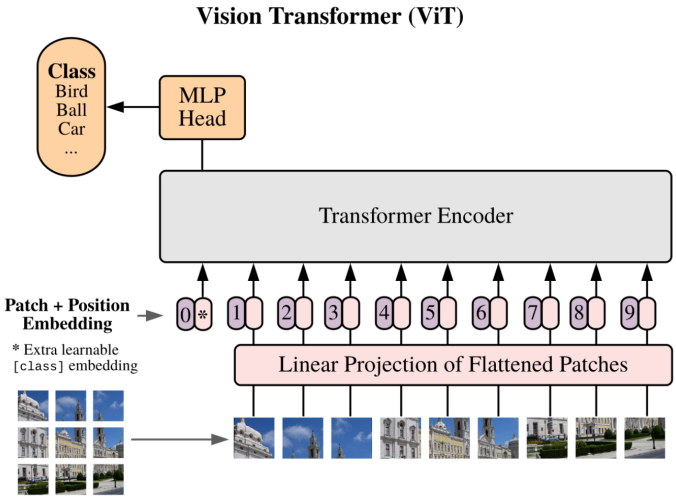
Зачем?


224×224×3
224×224×64
112×112×128
VGG
56×56×256
28×28×512
14×14×512
7×7×512
1×1×4096
1×1×1000
GoogLeNet
ResNet
34-layer residual

Классификация изображений

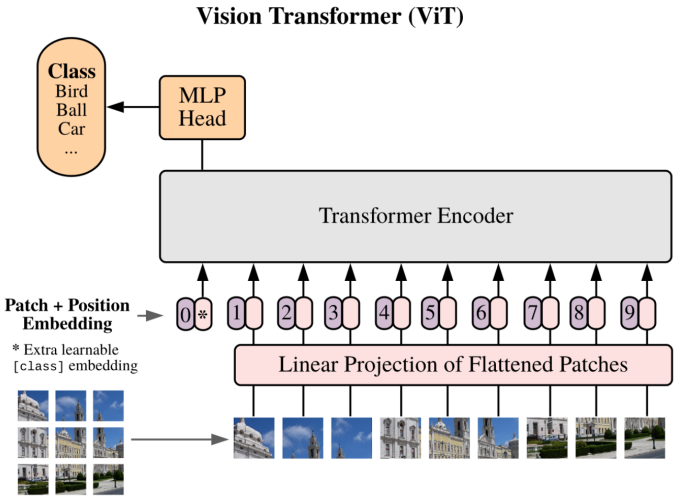


Классификация изображений



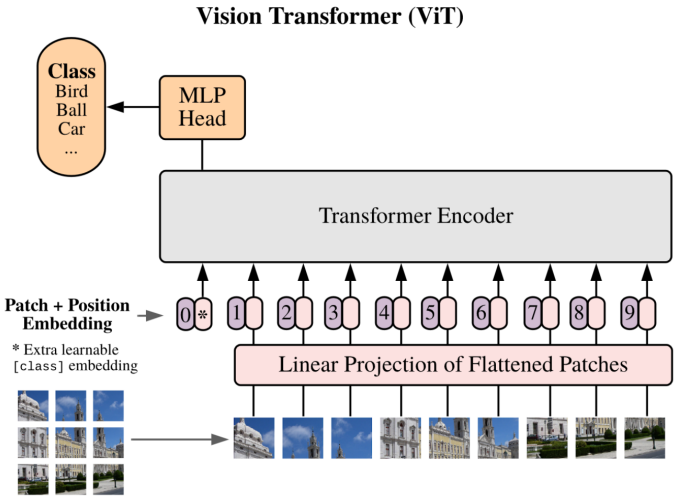
- Изображение “нарезаем“ на кусочки, размер кусочка - гиперпараметр

Классификация изображений



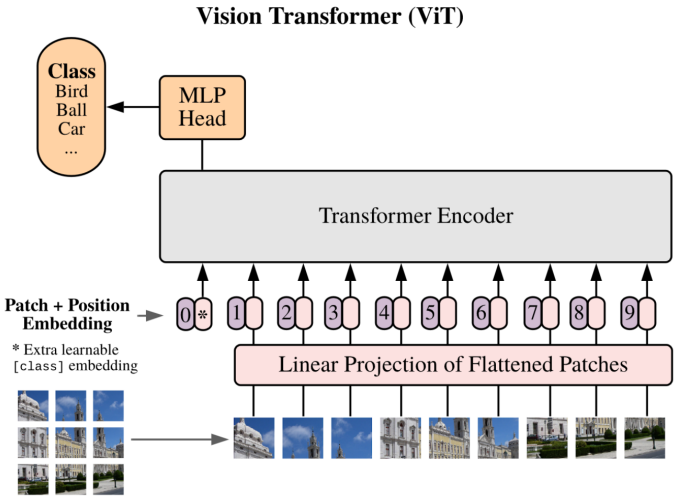
- Изображение “нарезаем“ на кусочки, размер кусочка - гиперпараметр
- Каждый кусочек преобразуем в вектор, затем слева на матрицу

Классификация изображений



- Изображение “нарезаем“ на кусочки, размер кусочка - гиперпараметр
- Каждый кусочек преобразуем в вектор, затем слева на матрицу
- Подаем на вход encoder-у, приписывая в начало фиктивный обучаемый вектор

Классификация изображений



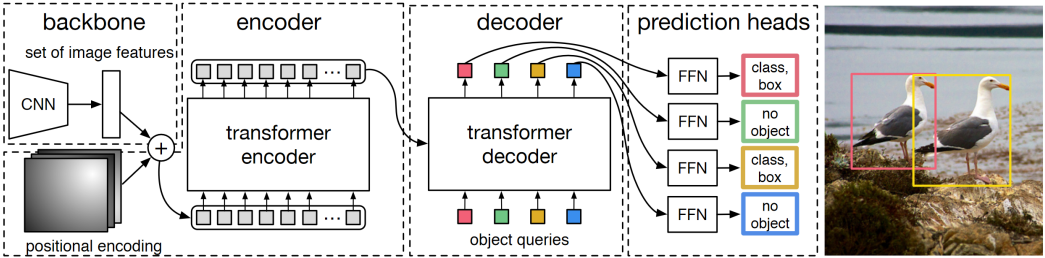
- Изображение “нарезаем“ на кусочки, размер кусочка - гиперпараметр
- Каждый кусочек преобразуем в вектор, затем слева на матрицу
- Подаем на вход encoder-у, приписывая в начало фиктивный обучаемый вектор
- На первом векторе из выхода encoder-а строим классификатор

Результаты

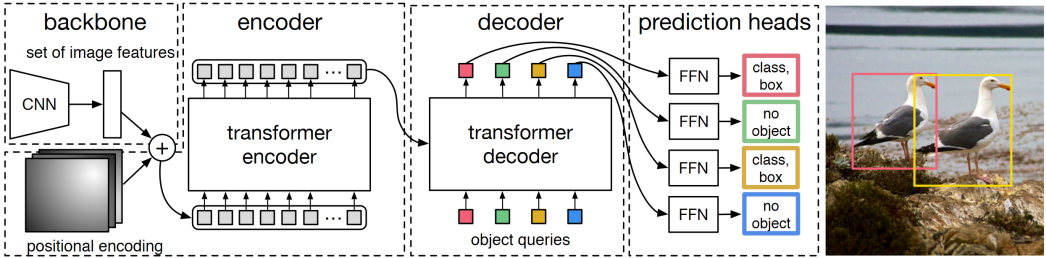
	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-I21K (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	**88.55** ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	**90.72** ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	**99.50** ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	–
CIFAR-100	**94.55** ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	–
Oxford-IIIT Pets	**97.56** ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	–
Oxford Flowers-102	99.68 ± 0.02	**99.74** ± 0.00	99.61 ± 0.02	99.63 ± 0.03	–
VTAB (19 tasks)	**77.63** ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	–
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Object detection

DETR

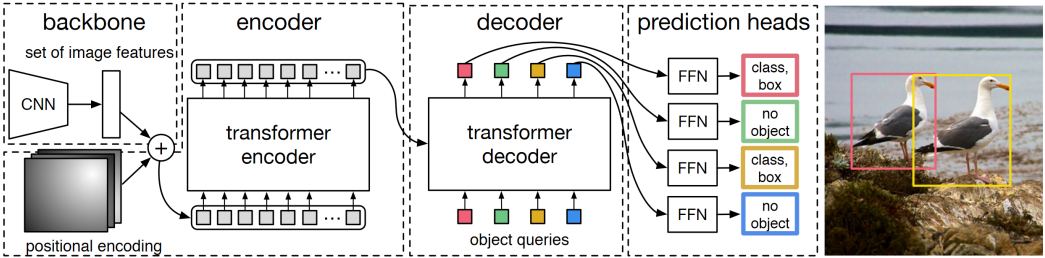

backbone
set of image features
CNN
positional encoding
encoder
transformer encoder
decoder
transformer decoder
object queries
prediction heads
FFN
FFN
FFN
FFN
class, box
no object
class, box
no object

DETR



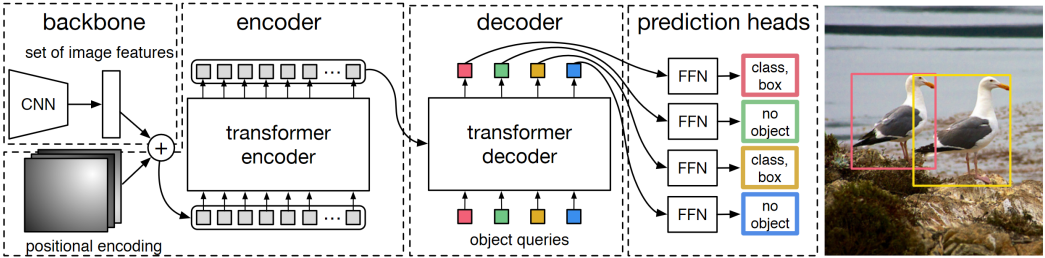
- Получаем карту признаков $\mathbb{R}^{C \times H \times W}$ с помощью предобученной CNN

DETR



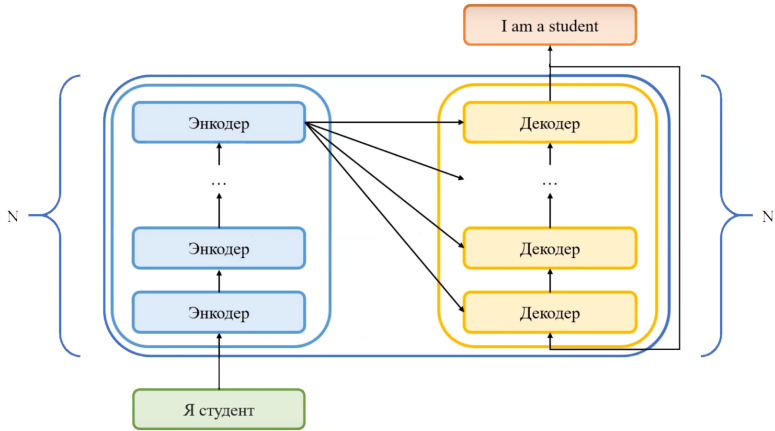
- Получаем карту признаков $\mathbb{R}^{C \times H \times W}$ с помощью предобученной CNN
- Получаем последовательность из HW векторов, подаем на вход encoder-у

DETR

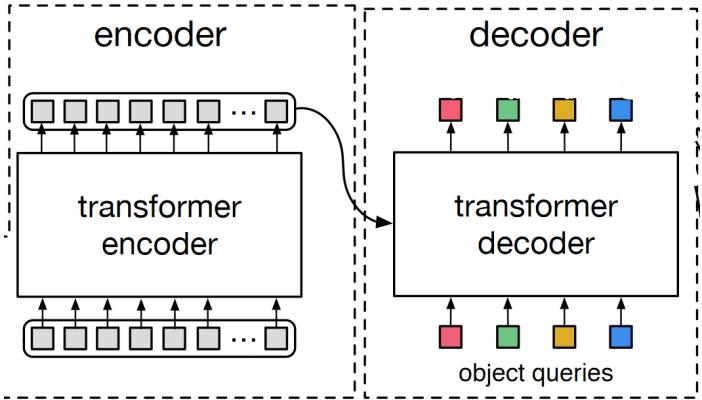


- Получаем карту признаков $\mathbb{R}^{C \times H \times W}$ с помощью предобученной CNN
- Получаем последовательность из HW векторов, подаем на вход encoder-у
- Введем метку [EOS], запускаем decoder, пока классификатор не выдаст [EOS]

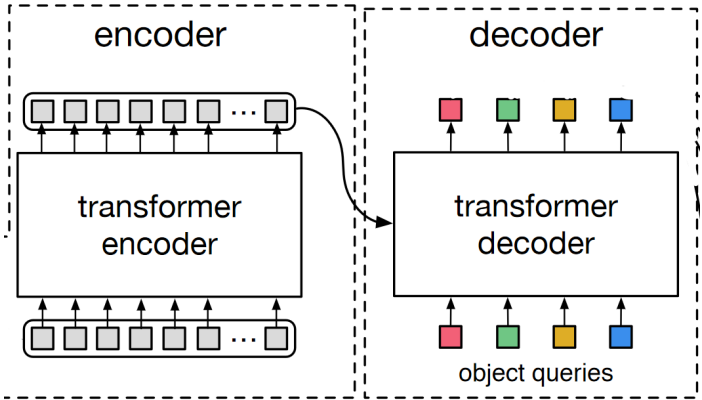
Decoder


I am a student
Энкодер
Энкодер
Энкодер
Декодер
Декодер
Декодер
N
N
Я студент

Decoder

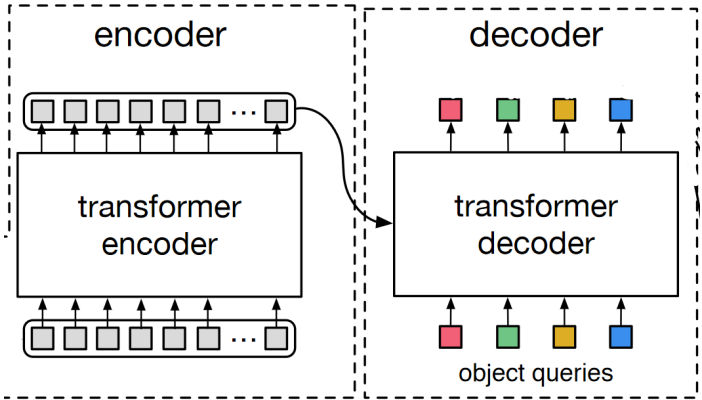

encoder
decoder
transformer
encoder
transformer
decoder
object queries

Decoder



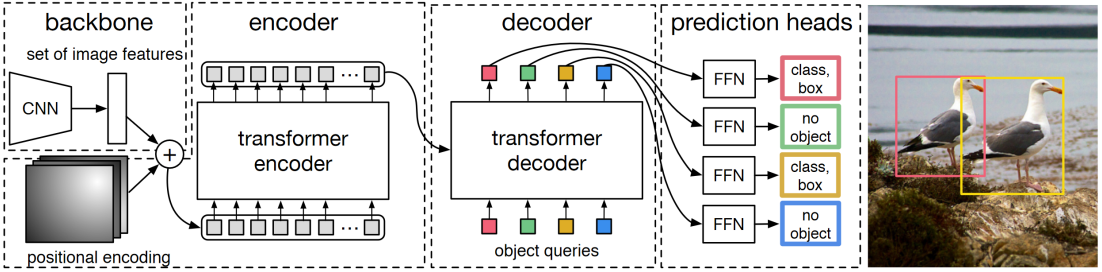
- Выберем N так, чтобы оно было всегда больше, чем число объектов

Decoder

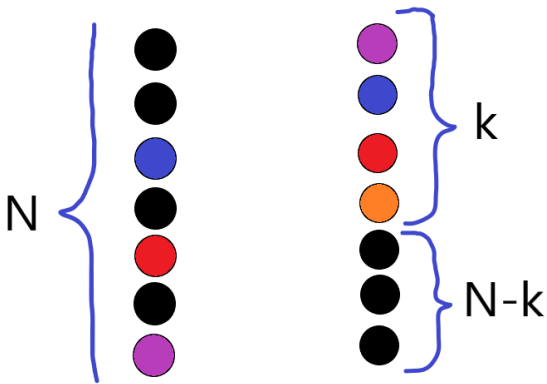


- Выберем N так, чтобы оно было всегда больше, чем число объектов
- Добавим метку [no object]

DETR

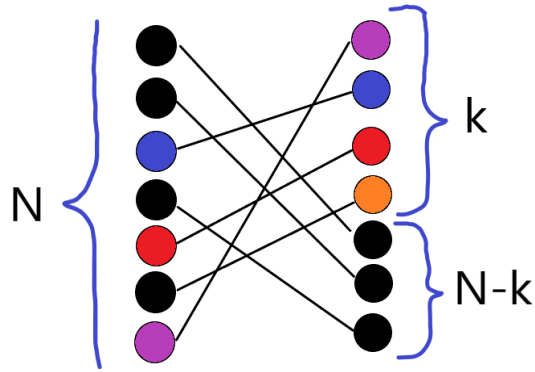

backbone
set of image features
CNN
positional encoding
encoder
transformer encoder
decoder
transformer decoder
object queries
prediction heads
FFN
class, box
FFN
no object
FFN
class, box
FFN
no object

Loss



$$w(u,v) = \begin{cases} 0, \text{ если v — фиктивная} \\ -\mathbb{P}[\text{class}(u) = \text{class}(v)] + \mathcal{L}_{\text{box}}(u,v) \end{cases}$$

Loss



$$\mathcal{L} = \sum_{(u,v)\in M} \left(-\ln \mathbb{P}[\text{class}(u) = \text{class}(v)] + [v \text{ не фиктивная}]\mathcal{L}_{\text{box}}(u, v) \right)$$

Результаты

Model	GFLOPS/FPS	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster RCNN-DC5	320/16	166M	39.0	60.5	42.3	21.4	43.5	52.5
Faster RCNN-FPN	180/26	42M	40.2	61.0	43.8	24.2	43.5	52.0
Faster RCNN-R101-FPN	246/20	60M	42.0	62.5	45.9	25.2	45.6	54.6
Faster RCNN-DC5+	320/16	166M	41.1	61.4	44.3	22.9	45.9	55.0
Faster RCNN-FPN+	180/26	42M	42.0	62.1	45.5	26.6	45.4	53.4
Faster RCNN-R101-FPN+	246/20	60M	44.0	63.9	**47.8**	**27.2**	48.1	56.0
DETR	86/28	41M	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5	187/12	41M	43.3	63.1	45.9	22.5	47.3	61.1
DETR-R101	152/20	60M	43.5	63.8	46.4	21.9	48.0	61.8
DETR-DC5-R101	253/10	60M	**44.9**	**64.7**	47.7	23.7	**49.5**	**62.3**

Список источников

- https://arxiv.org/pdf/2010.11929.pdf
- https://arxiv.org/pdf/2005.12872.pdf
- https://medium.com/visionwizard/detr-b677c7016a47