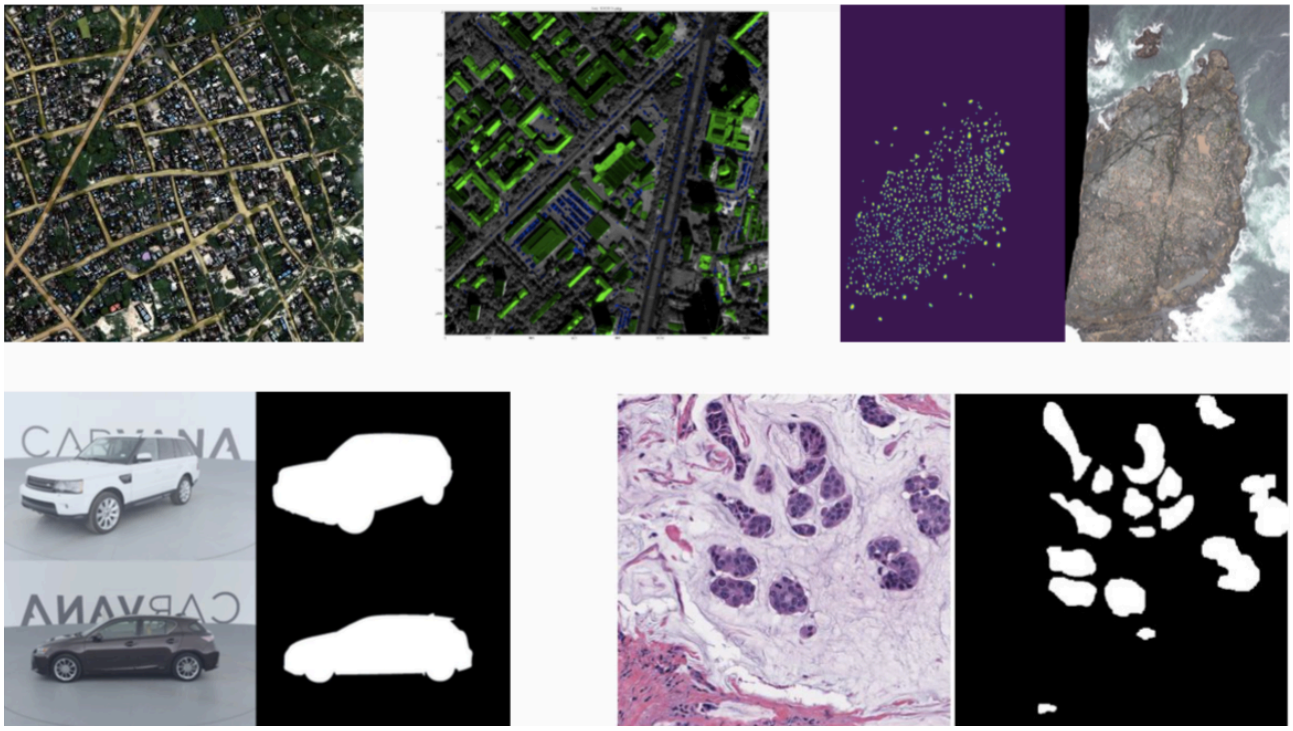
Идеи современных архитектур нейросетей для семантической сегментации изображений

Постановка задачи

Семантическая сегментация изображений - это разделение изображения на отдельные группы пикселей, области, соответствующие одному объекту с одновременным определением типа объекта в каждой области. Задача семантической сегментации является высокоуровневой задачей обработки изображений, относящейся к группе задач т.н. слабого искусственного интеллекта. Она является даже более сложной, чем задача классификации изображений и поиска объектов, что обусловлено не только необходимостью определения классов объектов, но и выявления их структуры, правильного выделения частей объектов на изображении.

Конкретным приложением, для которого важны методы семантической сегментации, является задача анализа аэрофотоснимков высокого разрешения с целью автоматического построения на их основе детальных карт местности или города. При создании карт как раз необходимо точно определить границы объектов на снимке поверхности земли, а так же указать их класс: здание, водоём, дорога, река, автомобиль, растительность. Алгоритмы автоматической семантической сегментации позволят существенно упростить задачу картографов при построении карт и сократить время на обработку данных.



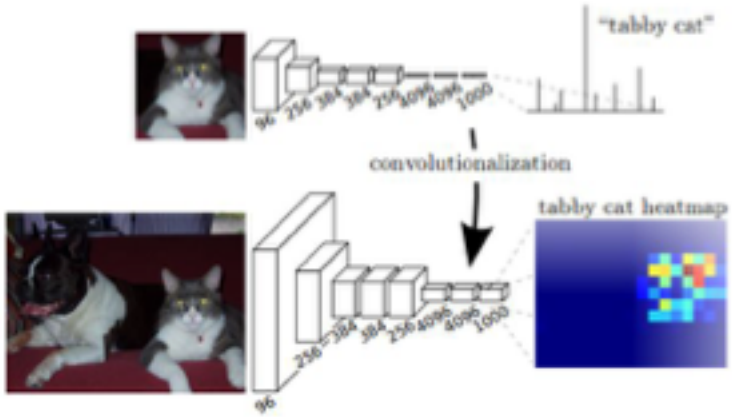
Так же, задача семантической сегментации крайне важна при обработке медицинских изображений для обнаружения опухолей и других патологии.

Основным подходом в настоящее время при решении задач классификации, семантического сегментирования является использование сверточных нейтронных

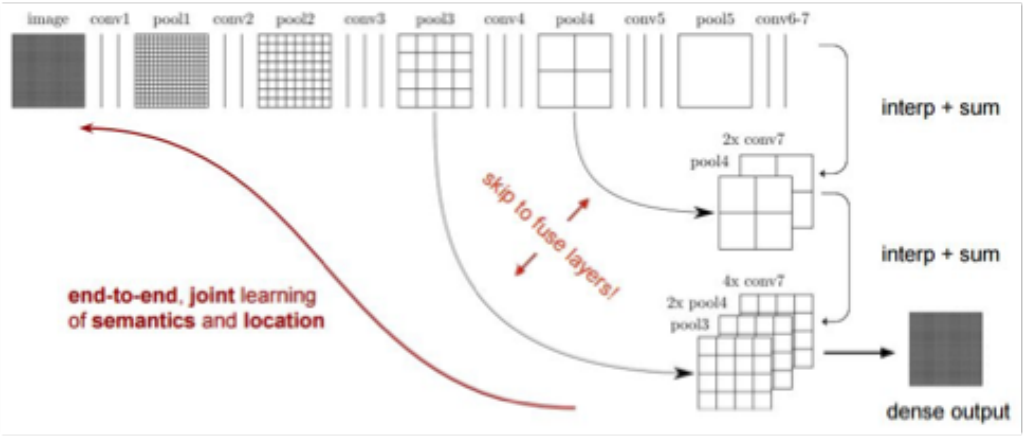
сетей. Данный подход хорошо себя показал в соревновании ImageNet и на данный момент является трендом. Рассмотрим несколько примеров архитектур нейронных сетей для задачи сегментации.

Fully Convolutional Neural Networks

К исходному изображению последовательно применяются операции свертки и unsampling (обычно макспуллиг). Таким образом в изображении находится и локализуется некий объект. После этого, получившееся в результате сверток значения необходимо преобразовать в пиксели исходного изображения (upscaling), построив тем самым карту сегментации.



Понижение размерности за счет операции пуллинга в сверточной сети приводит к увеличению области видимости, но при этом теряется точность позиционирования итогового класса. Поэтому было предложено использовать выходы с ранних слоев для уточнения. Данная идея получила название **skip connections** - ансамбль предсказаний по разным разрешениям.

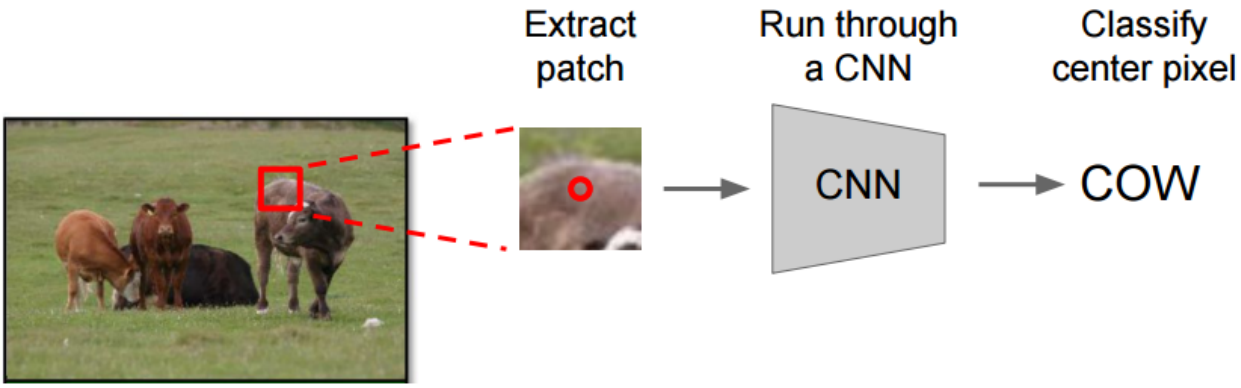


Одним из преимуществ использования полностью сверхточных нейронных сетей для сегментации изображений является то, что можно использовать идею **transfer learning** - взять уже обученную на большой выборке сеть (например, VGG или GoogleNet), обрезать ее по первым сверточным слоям и дообучить для нашей

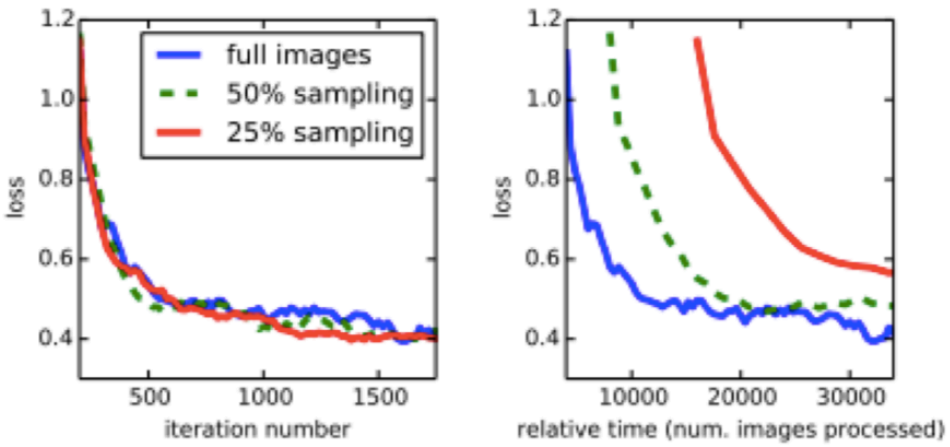
задачи. Таким образом, можно взять сеть, обученную на задаче классификации изображений и адаптировать ее для задачи сегментации. Затем, после встраивания первых слоев можно произвести **fine-tuning** - тонкую настройку параметров последних слоев сети, чтобы получившая карта сегментации была наиболее четкой. При использовании transfer learning можно еще замораживать градиент на первых предобученных слоях, тем самым оптимизируя только последние слои, которые отвечают за задачу сегментации. Обучение происходит быстрее, при этом не теряя в точности.

Поскольку в результате сверток понижается размерность исходного изображения, то получившая в результате карта сегментации может иметь низкое разрешение. Бороться с этим предлагается путем повышения разрешения (**upsampling**). Можно использовать простые методы, такие как unpooling, у которого нет параметров, так и специфичные обучаемые методы, например обратную свертку (deconvolution layer).

Так же, в задачи сегментации можно применять **patch-based** подход. Он заключается в следующем: для каждого пикселя исходного изображения берется кусок с центром в этом пикселе, прогоняется через сверточную сеть для классификации, и этот пиксель красится в цвет соответствующего класса на карте сегментации.



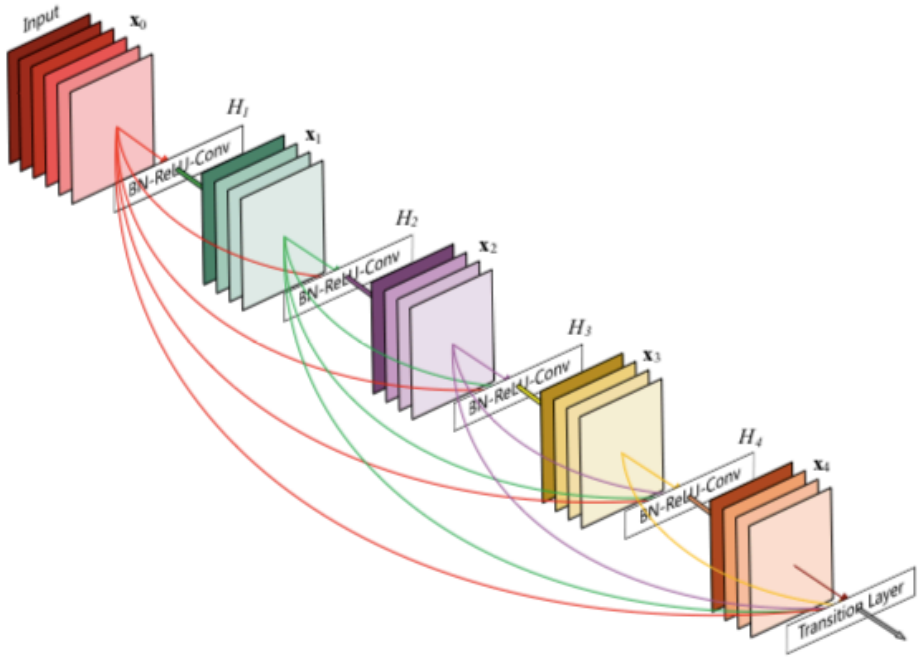
Эту идею можно развить: предлагается считать функцию потерь для задачи сегментации не на всей картинки, а только на некоторых ее частях, например, по случайному набору выходных пикселей. В таком случае, градиент будет пробрасываться только по этим частям, и свертка будет идти тоже только по этим частям. За счет этого обучение происходит быстрее, и, как показывают эксперименты, качество сохраняется.



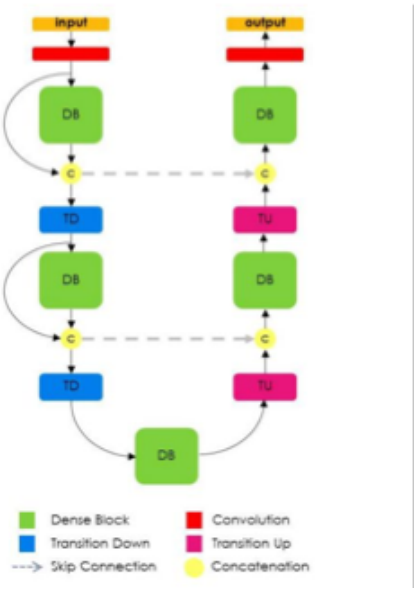
Первоисточник: Fully Convolutional Networks for Semantic Segmentation

Fully Convolutional DenseNet

У сверточных сетей возникает проблема затухания градиента. Было предложено решение, получившее название Dense Convolutional Network (DenseNet): разбить слои сети на блоки и попарно соединить эти блоки друг с другом.



Эту идею можно развить и применить для задачи сегментации изображений, объединив с архитектурой кодировщик-декодировщик. Получается следующая архитектура:



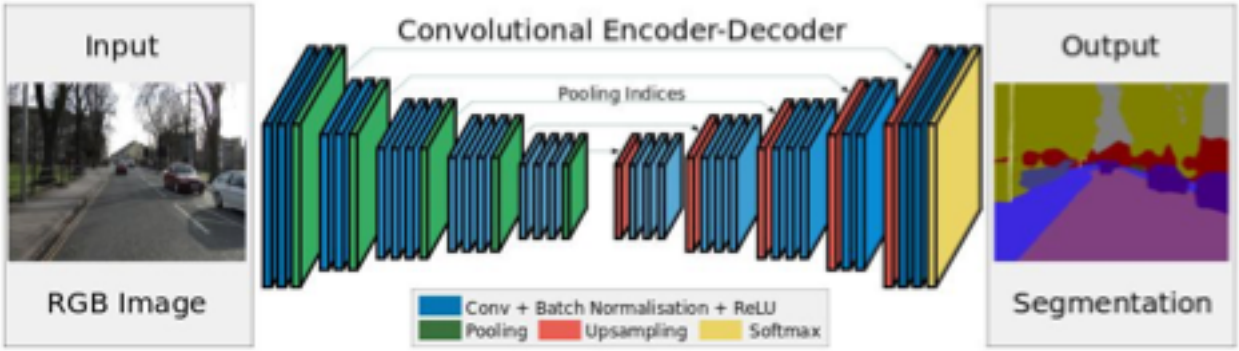
Сравнение качества полученной модели приведено ниже:

Model	Pretrained	# parameters (M)	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Cyclist	Mean IoU	Global accuracy
SegNet [1]	✓	29.5	68.7	52.0	87.0	58.5	13.4	86.2	25.3	17.9	16.0	60.5	24.8	46.4	62.5
Bayesian SegNet [15]	✓	29.5	n/a											63.1	86.9
DeconvNet [21]	✓	252	n/a											48.9	85.9
Visin et al. [36]	✓	32.3	n/a											58.8	88.7
FCN8 [20]	✓	134.5	77.8	71.0	88.7	76.1	32.7	91.2	41.7	24.4	19.9	72.7	31.0	57.0	88.0
DeepLab-LFOV [5]	✓	37.3	81.5	74.6	89.0	82.2	42.3	92.2	48.4	27.2	14.3	75.4	50.1	61.6	–
Dilation8 [37]	✓	140.8	82.6	76.2	89.0	84.0	46.9	92.2	56.3	35.8	23.4	75.3	55.5	65.3	79.0
Dilation8 + FSO [17]	✓	140.8	**84.0**	77.2	91.3	**85.6**	**49.9**	92.5	59.1	**37.6**	16.9	76.0	**57.2**	66.1	88.3
Classic Upsampling	✗	20	73.5	72.2	92.4	66.2	26.9	90.0	37.7	22.7	30.8	69.6	25.1	55.2	86.8
FC-DenseNet56 (k=12)	✗	1.5	77.6	72.0	92.4	73.2	31.8	92.8	37.9	26.2	32.6	79.9	31.1	58.9	88.9
FC-DenseNet67 (k=16)	✗	3.5	80.2	75.4	93.0	78.2	40.9	**94.7**	58.4	30.7	**38.4**	81.9	52.1	65.8	90.8
FC-DenseNet103 (k=16)	✗	9.4	83.0	**77.3**	**93.0**	77.3	43.9	94.5	**59.6**	37.1	37.8	**82.2**	50.5	**66.9**	**91.5**

Table 3. Results on CamVid dataset. Note that we trained our own pretrained FCN8 model

Первоисточник: The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation

SegNet



SegNet иммет архитектуру схожую с автокодировщиков. Сначала идут слои свертки и пуллинга, затем анпуллинга и транспонированной свертки. Поскольку в данной архитектуре не используются полносвязные слои, то она является достаточно легковесной.

Сравнение качества SegNet приведено ниже.

Method	Building	Tree	Sky	Car	Sign-Symbol	Road	Pedestrian	Fence	Column-Pole	Side-walk	Bicyclist	Class avg.	Global avg.	mIoU	BF
SfM+Appearance [28]	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	53.0	69.1	n/a*	
Boosting [29]	61.9	67.3	91.1	71.1	58.5	92.9	49.5	37.6	25.8	77.8	24.7	59.8	76.4	n/a*	
Dense Depth Maps [32]	85.3	57.3	95.4	69.2	46.5	**98.5**	23.8	44.3	22.0	38.1	28.7	55.4	82.1	n/a*	
Structured Random Forests [31]	n/a											51.4	72.5	n/a*	
Neural Decision Forests [64]	n/a											56.1	82.1	n/a*	
Local Label Descriptors [65]	80.7	61.5	88.8	16.4	n/a	98.0	1.09	0.05	4.13	12.4	0.07	36.3	73.6	n/a*	
Super Parsing [33]	87.0	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70.0	19.4	51.2	83.3	n/a*	
SegNet (3.5K dataset training - 140K)	**89.6**	**83.4**	96.1	**87.7**	52.7	96.4	**62.2**	**53.45**	**32.1**	**93.3**	**36.5**	**71.20**	**90.40**	60.10	46.84
CRF based approaches															
Boosting + pairwise CRF [29]	70.7	70.8	94.7	74.4	55.9	94.1	45.7	37.2	13.0	79.3	23.1	59.9	79.8	n/a*	
Boosting+Higher order [29]	84.5	72.6	**97.5**	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	59.2	83.8	n/a*	
Boosting+Detectors+CRF [30]	81.5	76.6	96.2	78.7	40.2	93.9	43.0	47.6	14.3	81.5	33.9	62.5	83.8	n/a*	

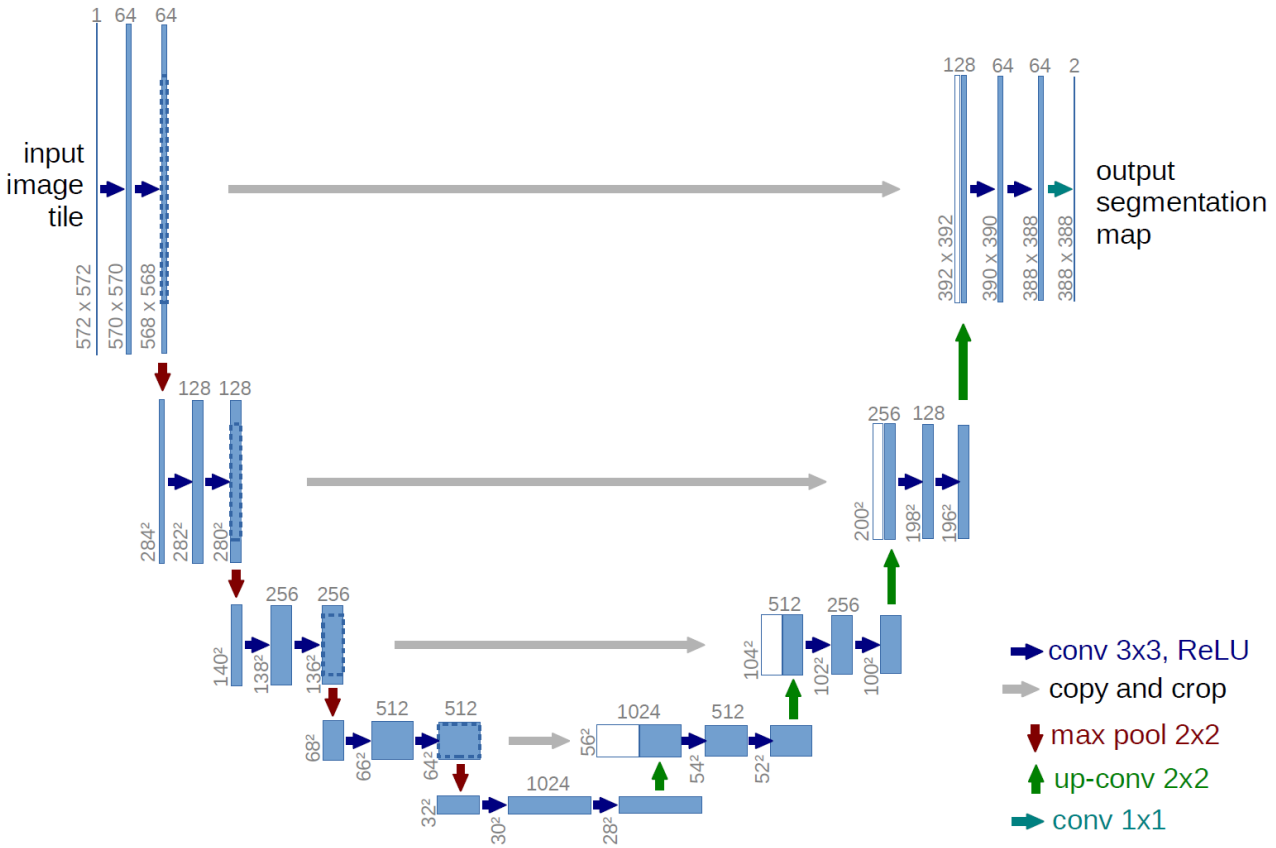
TABLE 2

Quantitative comparisons of SegNet with traditional methods on the CamVid 11 road class segmentation problem [22]. SegNet outperforms all the other methods, including those using depth, video and/or CRF's on the majority of classes. In comparison with the CRF based methods SegNet predictions are more accurate in 8 out of the 11 classes. It also shows a good $\approx 10\%$ improvement in class average accuracy when trained on a large dataset of 3.5K images. Particularly noteworthy are the significant improvements in accuracy for the smaller/thinner classes. * Note that we could not access predictions for older methods for computing the mIoU, BF metrics.

Первоисточник: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation

U-net

Данная архитектура была придумана для сегментации медицинских изображений. Особенностью медицинских изображений является то, что они более простые, чем визуальные сцены, однако выборка из таких изображений очень маленькая. Поэтому U-net существенно использует идею skip connections - промежуточные выходы кодировщика конкатенируются с промежуточными выходами декодировщика с применением upsampling. Это позволяет не терять информацию на этапах свертки, а использовать ее по максимуму, поскольку выборка маленькая.



Результаты U-net на выборках PhC-U373 (35 изображений для обучения) и DIC-Hela (20 изображений):

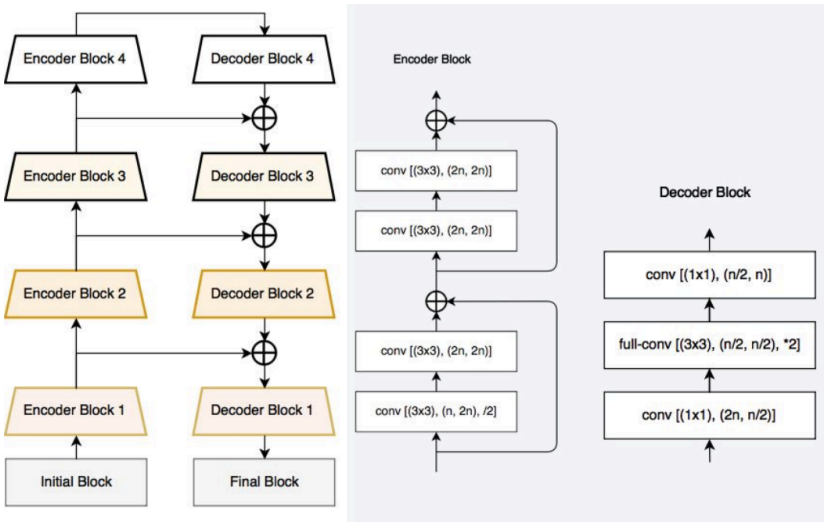
Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	**0.9203**	**0.7756**

Первоисточник: U-Net: Convolutional Networks for Biomedical Image Segmentation

Link-net

Еще одна архитектура типа кодировщик-декодировщик. Выходы кодировщика и декодировщика суммируются, а не конкатенируются как в U-net. Декодировщик действует следующим образом:

1. Свертками 1x1 понижается размерность в 4 раза
2. Transposed convolution вместо простого upsampling увеличивает разрешение в 2 раза
3. Несколько сверточных слоев возвращает признаковую размерность.



Данная архитектура хорошо себя зарекомендовала в различных конкурсах на сегментацию изображений.

Сравние моделей:

TABLE IV: Cityscapes val set results (* on test set)

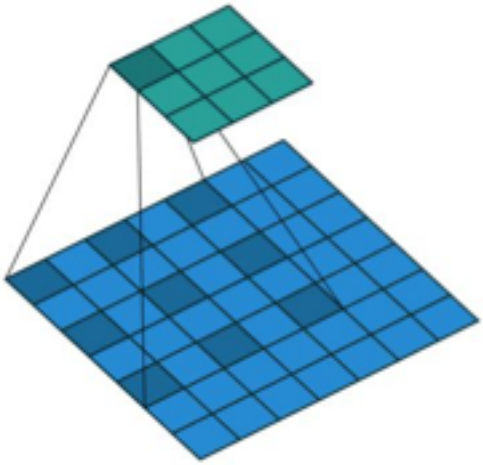
Model	Class IoU	Class iIoU
SegNet*	56.1	34.2
ENet*	58.3	34.4
Dilation10	68.7	-
Deep-Lab CRF (VGG16)	65.9	-
Deep-Lab CRF (ResNet101)	71.4	42.6
LinkNet without bypass	72.6	51.4
LinkNet	**76.4**	**58.6**

Первоисточник: LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation

Dilated convolutions

Помимо использования различных техник для восстановления размерности развивались также подходы, которые ставили целью уменьшить ее снижение. Были придуманы dilated и atrous свертки, которые имеют большую разрешающую

способность без пулинга, однако увеличивают теоретическую и практическую сложность сети.



Сравние моделей:

	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Bicyclist	mean IoU
ALE	73.4	70.2	**91.1**	64.2	24.4	91.1	29.1	31.0	13.6	72.4	28.6	53.6
SuperParsing	70.4	54.8	83.5	43.3	25.4	83.4	11.6	18.3	5.2	57.4	8.9	42.0
Liu and He	66.8	66.6	90.1	62.9	21.4	85.8	28.0	17.8	8.3	63.5	8.5	47.2
SegNet	68.7	52.0	87.0	58.5	13.4	86.2	25.3	17.9	16.0	60.5	24.8	46.4
DeepLab-LFOV	81.5	74.6	89.0	82.2	42.3	**92.2**	48.4	27.2	14.3	**75.4**	50.1	61.6
Dilation8	**82.6**	**76.2**	89.9	**84.0**	**46.9**	**92.2**	**56.3**	**35.8**	**23.4**	75.3	**55.5**	**65.3**

Table 5: Semantic segmentation results on the CamVid dataset. Our model (Dilation8) is compared to ALE (Ladicky et al., 2009), SuperParsing (Tighe & Lazebnik, 2013), Liu and He (Liu & He, 2015), SegNet (Badrinarayanan et al., 2015), and the DeepLab-LargeFOV model (Chen et al., 2015a). Our model outperforms the prior work.

Первоисточник: Multi-Scale Context Aggregation by Dilated Convolutions