Paying attention to Attention

Piotr Mazurek

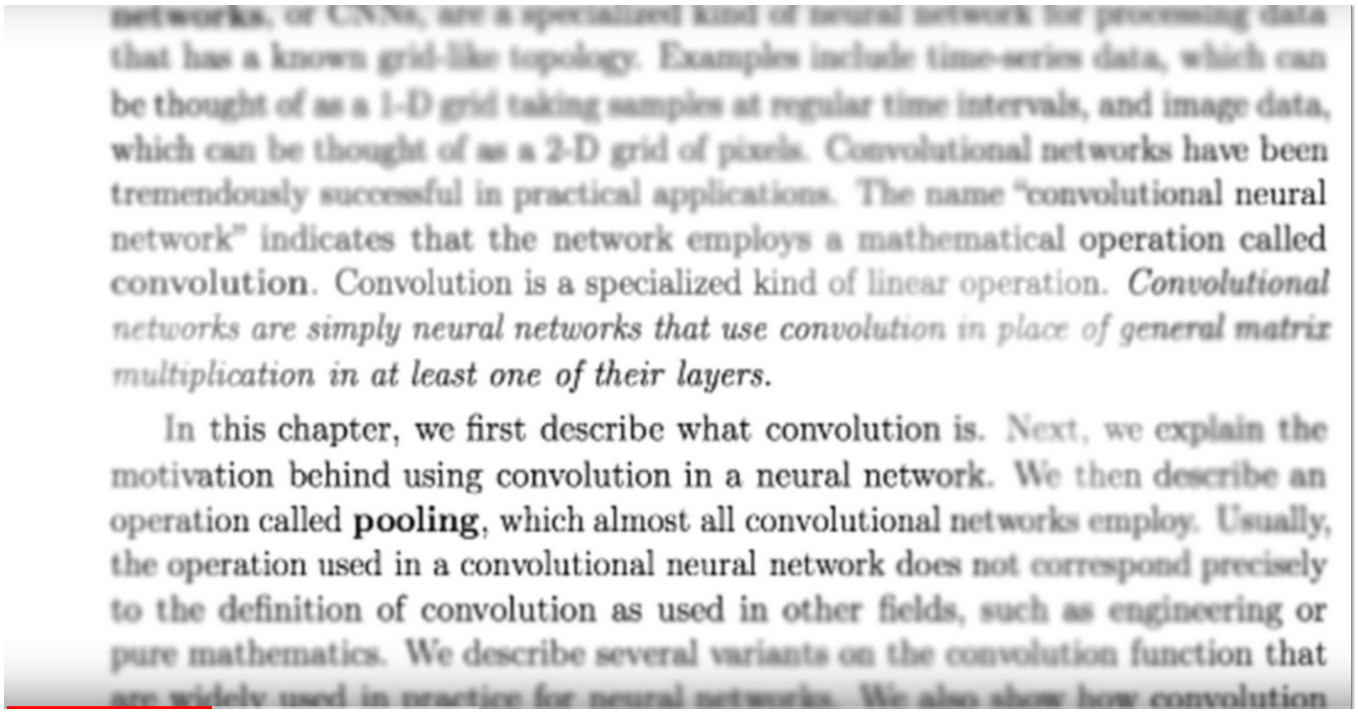
April 15, 2020

Agenda

1. Introduction to Attention
2. Introduction to Neural Networks
3. Basics of the Attention mechanism
4. Deep dive into Attention
5. Attention applications

Introduction to Attention

Basic intuition



networks, or CNNs, are a specialized kind of neural network for processing data that has a known grid-like topology. Examples include time-series data, which can be thought of as a 1-D grid taking samples at regular time intervals, and image data, which can be thought of as a 2-D grid of pixels. Convolutional networks have been tremendously successful in practical applications. The name "convolutional neural network" indicates that the network employs a mathematical operation called **convolution**. Convolution is a specialized kind of linear operation. ***Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.***

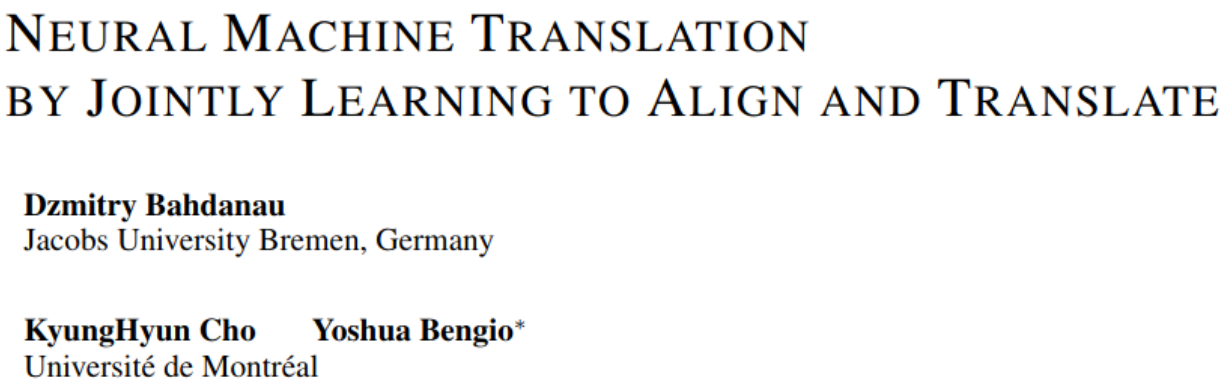
In this chapter, we first describe what convolution is. Next, we explain the motivation behind using convolution in a neural network. We then describe an operation called **pooling**, which almost all convolutional networks employ. Usually, the operation used in a convolutional neural network does not correspond precisely to the definition of convolution as used in other fields, such as engineering or pure mathematics. We describe several variants on the convolution function that are widely used in practice for neural networks. We also show how convolution

Figure: We teach a neural network to focus on some part of the data[1]

[1] https://www.youtube.com/watch?v=W2rWgXJBZhU&t=180s

Introduction to Attention

Who, where and when?



NEURAL MACHINE TRANSLATION
BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

Figure: Idea of the Attention first time mentioned (three times in the two consecutive lines, but still counts), Sept. 2014[1]

Introduction to Attention

Who, where and when?

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Figure: Attention is all You need paper, Dec. 2017[3]

Introduction to Attention

Boom in Attention papers

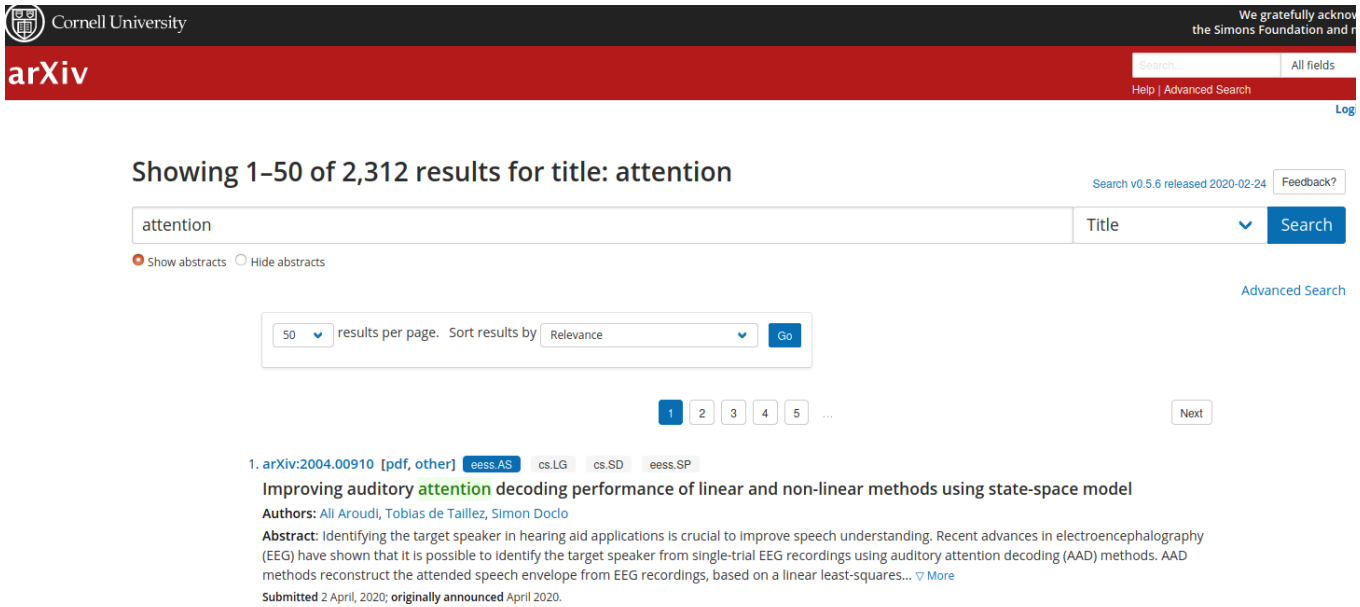


Figure: Attention - a hot research topic

Introduction to Attention

Better start paying attention now

Not All Attention Is Needed: Gated Attention Network for Sequence Data

Lanqing Xue,[1] **Xiaopeng Li,**[2*] **Nevin L. Zhang**[1,3]
[1]The Hong Kong University of Science and Technology, Hong Kong
[2]Amazon Web Services, WA, USA
[3]HKUST-Xiaoi Joint Lab, Hong Kong
lxueaa@cse.ust.hk,xiaopel@amazon.com,lzhang@cse.ust.hk

Figure: Not All Attention Is Needed, Dec. 2019

[12]

Neural Networks - basic introduction

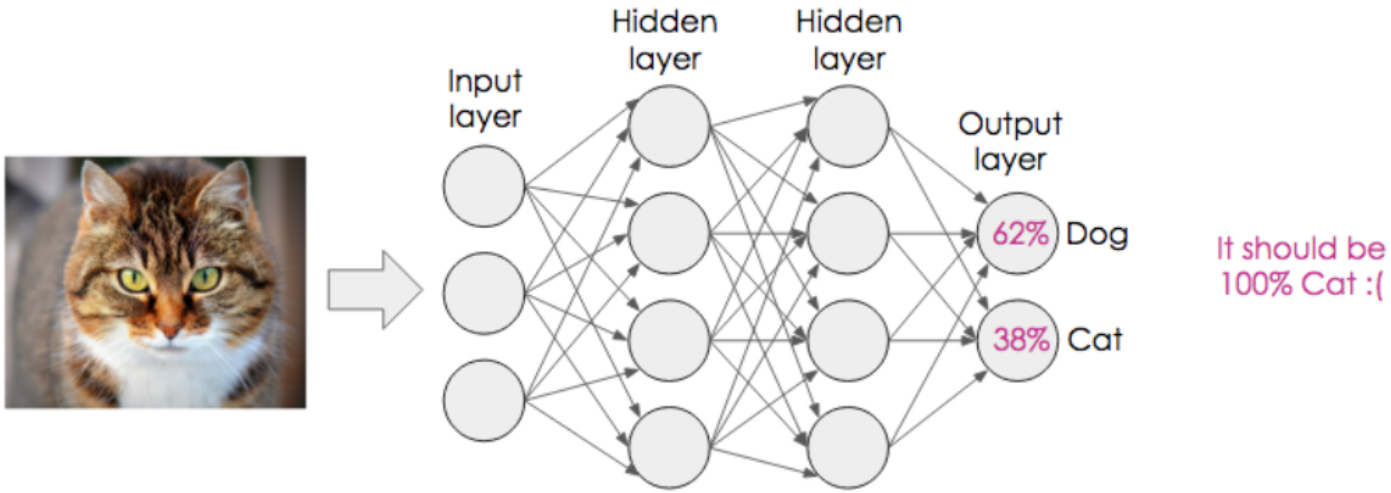


Figure: Not trained neural network[2]

[2] https://www.fromthegenesis.com/artificial-neural-network-part-5/

Neural Networks

Linear Algebra

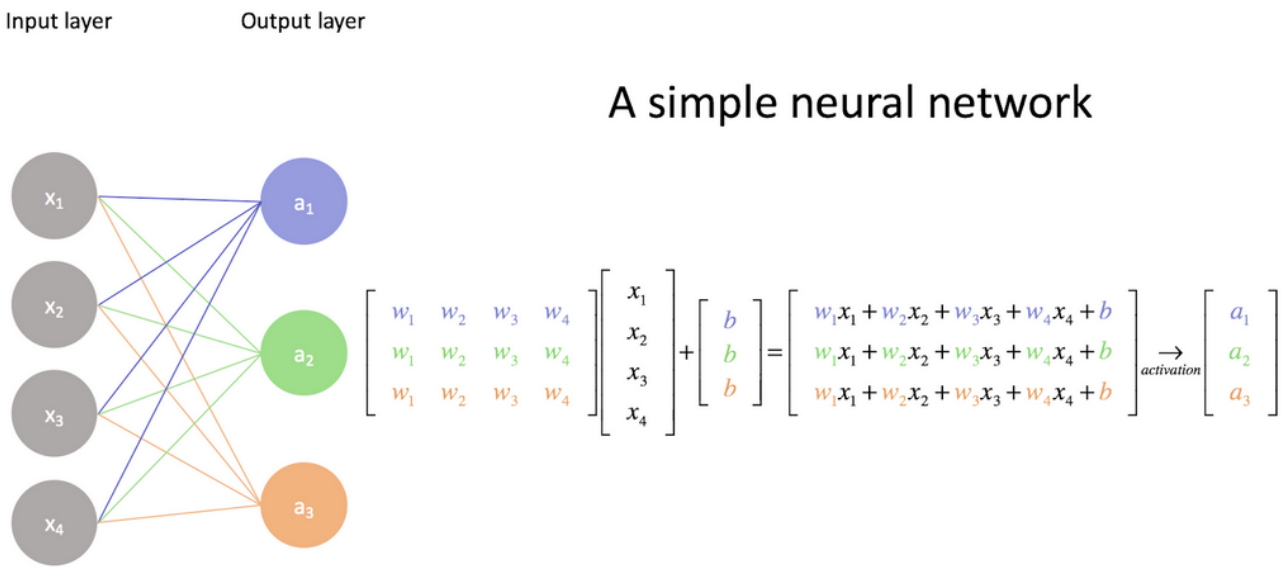
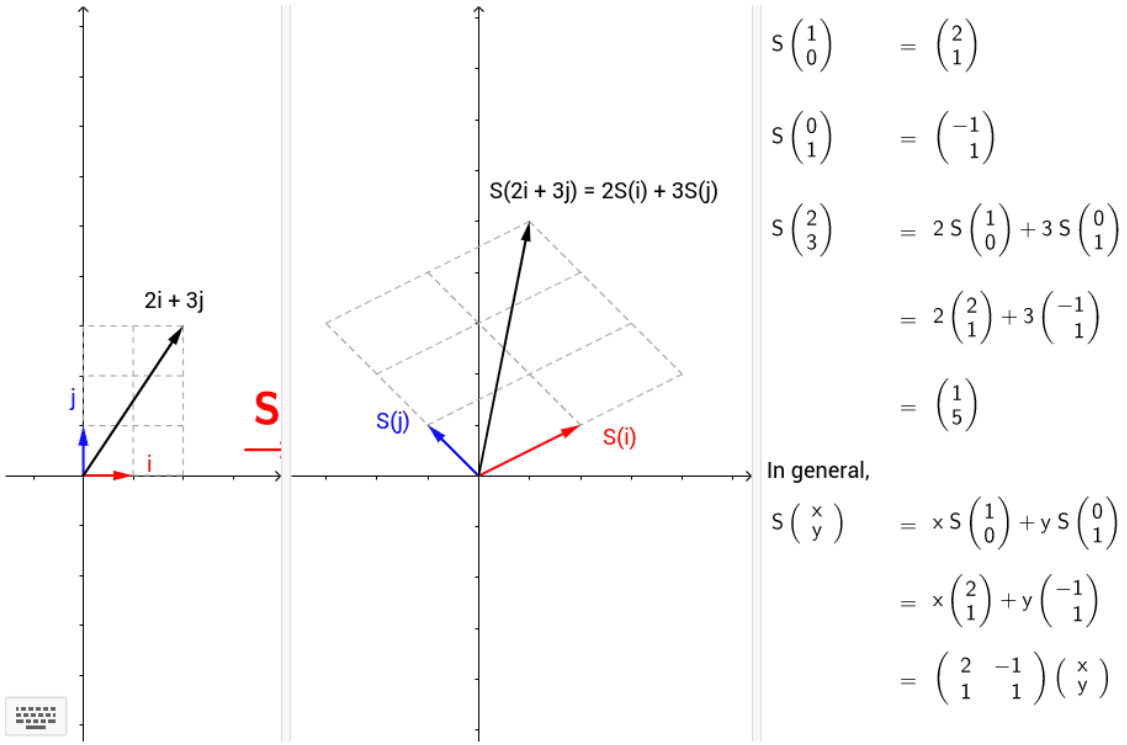


Figure: How matrices are used to compute values in neurons [3]

[3] https://www.jeremyjordan.me/intro-to-neural-networks/

Neural Networks

Algebraic intuition



$$S\begin{pmatrix}1\\0\end{pmatrix} = \begin{pmatrix}2\\1\end{pmatrix}$$

$$S\begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}-1\\1\end{pmatrix}$$

$$S\begin{pmatrix}2\\3\end{pmatrix} = 2S\begin{pmatrix}1\\0\end{pmatrix} + 3S\begin{pmatrix}0\\1\end{pmatrix} = 2\begin{pmatrix}2\\1\end{pmatrix} + 3\begin{pmatrix}-1\\1\end{pmatrix} = \begin{pmatrix}1\\5\end{pmatrix}$$

$$S\begin{pmatrix}x\\y\end{pmatrix} = xS\begin{pmatrix}1\\0\end{pmatrix} + yS\begin{pmatrix}0\\1\end{pmatrix} = x\begin{pmatrix}2\\1\end{pmatrix} + y\begin{pmatrix}-1\\1\end{pmatrix} = \begin{pmatrix}2 & -1\\1 & 1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

Figure: Linear Transformation [4]

[4] https://www.geogebra.org/m/B9Nbwh7w

Mathematical Formalization

Oversimplified

Neural network in general

$$f(x, \theta) = \hat{y} \tag{1}$$

where:

x: input data

Θ: model parameters

$\hat{y}$: distribution of probability over set of classes

Mathematical Formalization

Oversimplified

Neural network in general

$$f(x, \theta) = \hat{y} \tag{1}$$

where:
x: input data
Θ: model parameters
$\hat{y}$: distribution of probability over set of classes

Example

$$f(x, w, b) = \sigma(x^T w + b) \tag{2}$$

Loss Function

How to find θ

How to find θ

$$\underset{\theta}{\text{minimize}}\ J(\theta) \tag{3}$$

For example we can use SGD optimizer

Loss Function

How to find θ

How to find θ

$$\underset{\theta}{\text{minimize}}\ J(\theta) \tag{3}$$

For example we can use SGD optimizer

Example loss function

$$J(\theta) = \frac{1}{2}\mathbb{E}_{\mathbf{x},\mathbf{y}\sim\hat{p}_{\text{data}}}\|\boldsymbol{y} - f(\boldsymbol{x};\boldsymbol{\theta})\|^2 \tag{4}$$

Latent space

Auto-encoder

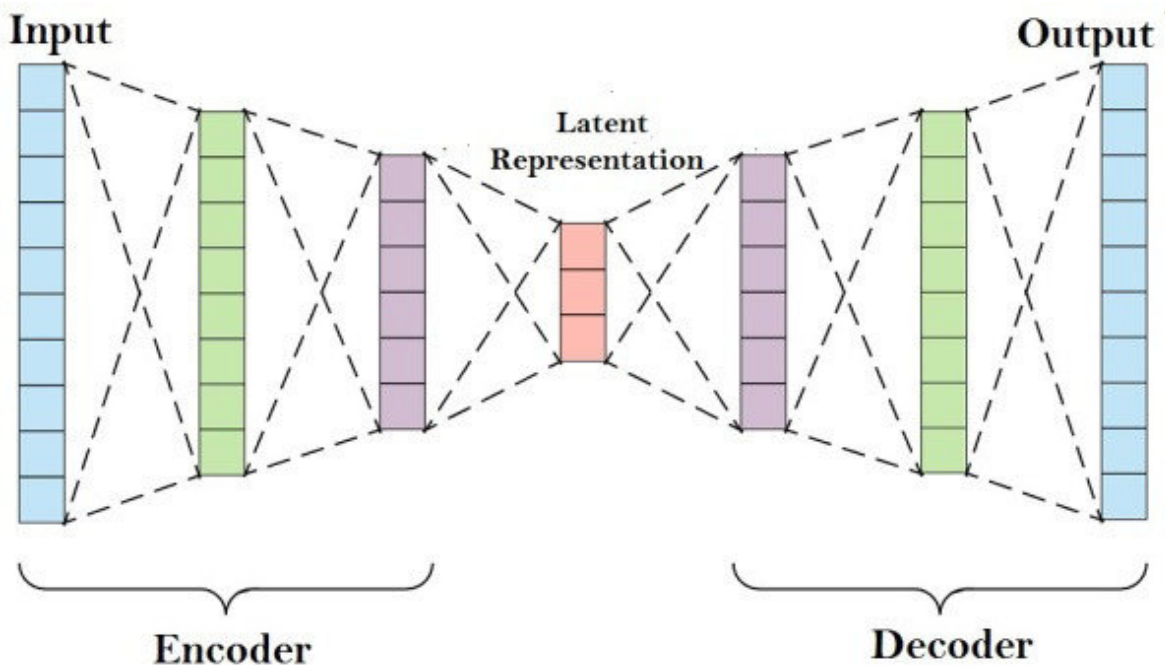


Figure: Example auto-encoder[5]

[5] https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

Latent space

The big picture

- Used to find the hidden representation of data
- By training model on a particular set of data we create some kind of general knowledge

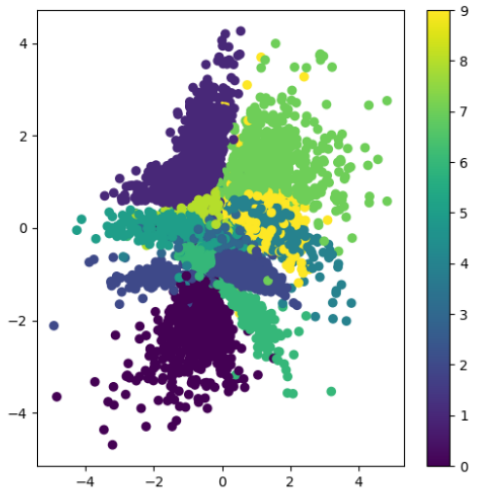


Figure: Finding information in chaos

Recurrent Neural Networks (RNNs)

Oversimplified

- Used for sequential data.
  E.g for text analysis, video recognition etc.

Recurrent Neural Networks (RNNs)

Oversimplified

- Used for sequential data. E.g for text analysis, video recognition etc.
- Have so-called "local memory" or "state memory"

Recurrent Neural Networks (RNNs)

Oversimplified

- Used for sequential data. E.g for text analysis, video recognition etc.
- Have so-called "local memory" or "state memory"
- Computationally expensive to train

Recurrent Neural Networks (RNNs)

Oversimplified

- Used for sequential data. E.g for text analysis, video recognition etc.
- Have so-called "local memory" or "state memory"
- Computationally expensive to train
- Problems with vanishing gradient

Recurrent Neural Networks (RNNs)

Oversimplified

- Used for sequential data. E.g for text analysis, video recognition etc.
- Have so-called "local memory" or "state memory"
- Computationally expensive to train
- Problems with vanishing gradient

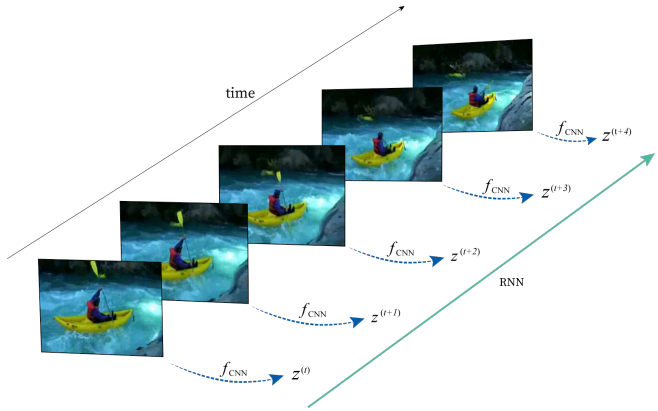
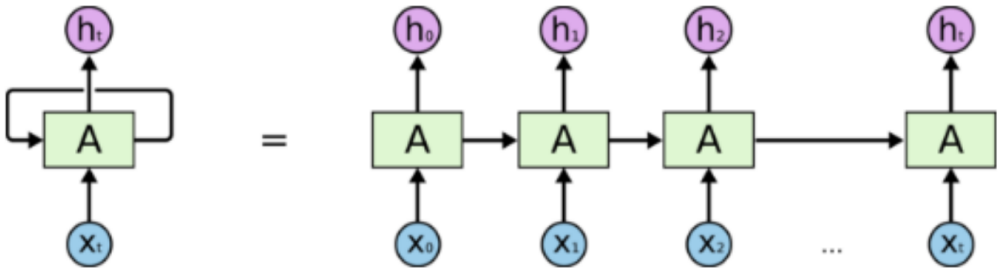


Figure: Sequence of frames[a]

[a] https://awesomeopensource.com/project/HHTseng/video-classificat

Recurrent Neural Networks (RNNs)

Oversimplified



An unrolled recurrent neural network.

Figure: Folded and unfolder RNN [6]

[6] https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e

Types of RNN

Vector to sequence model

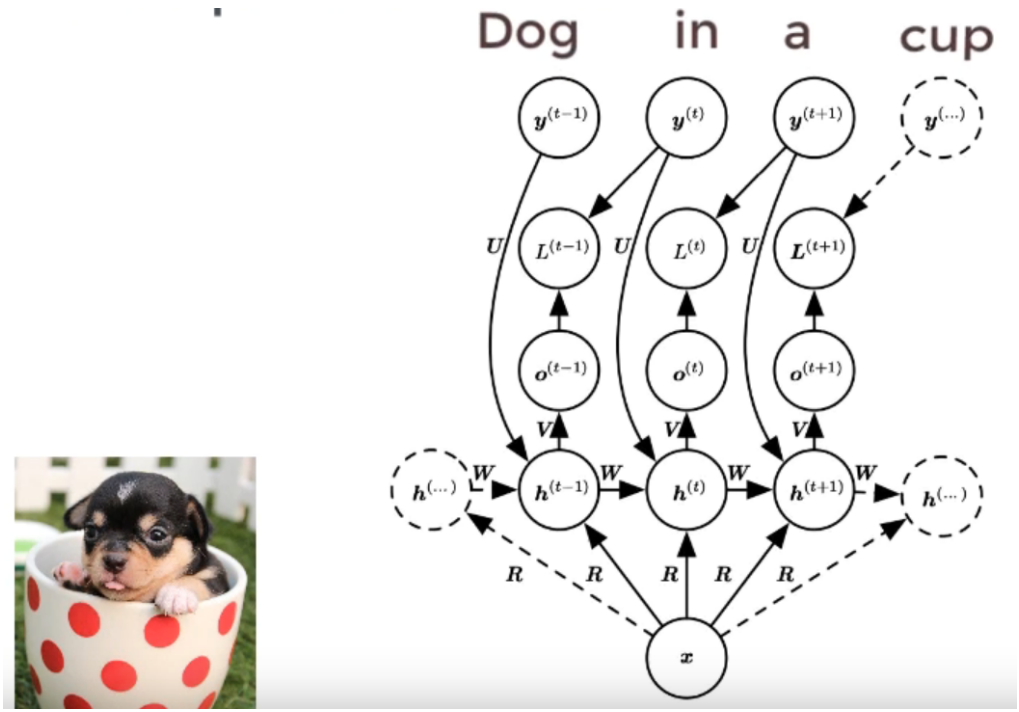


Figure: From an image (vector) generate a caption(sequence) [7]

[7] https://www.youtube.com/watch?v=TQQlZhbC5ps&t=610s

Types of RNN

Sequence to vector model

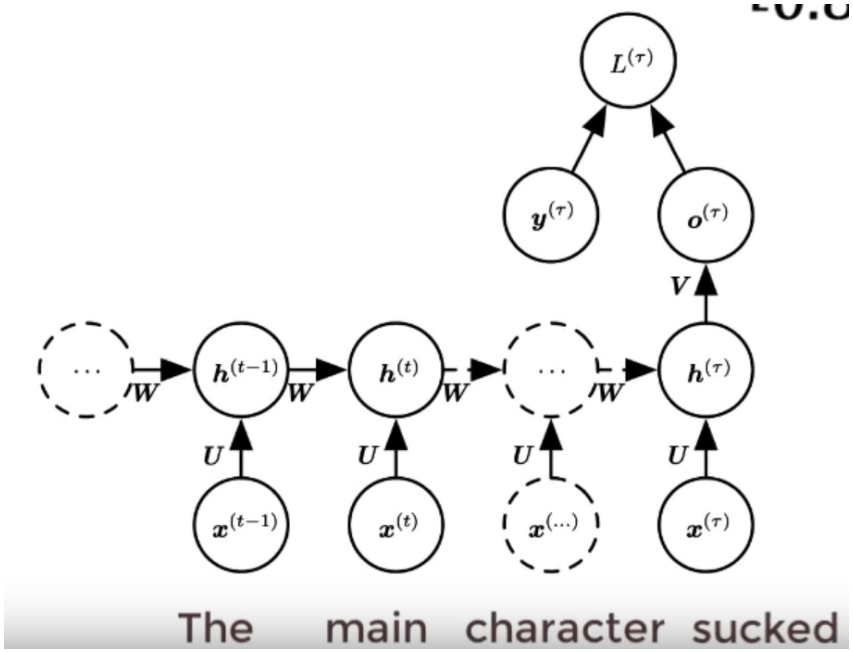


Figure: Take a review (sequence) and predict whether it is positive or negative (vector)[7]

Types of RNN

Sequence to sequence (seq2seq) model

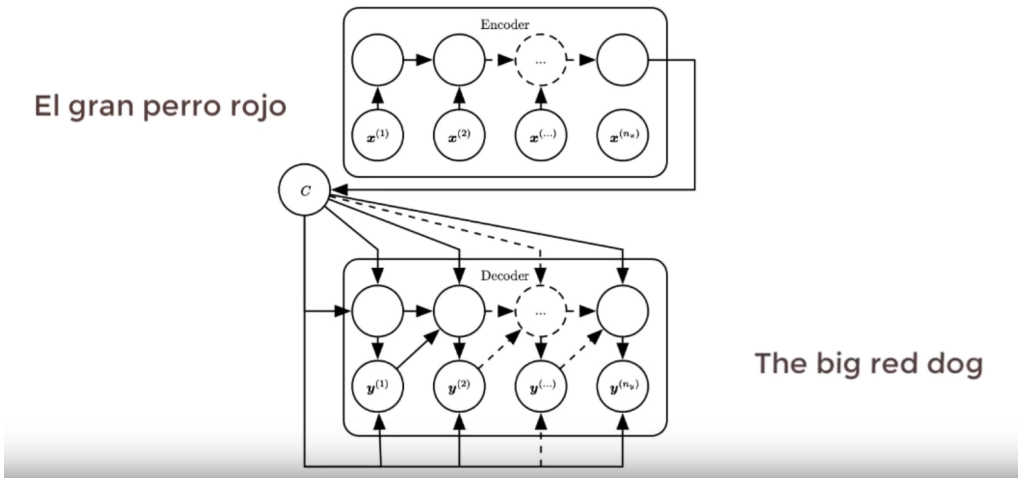
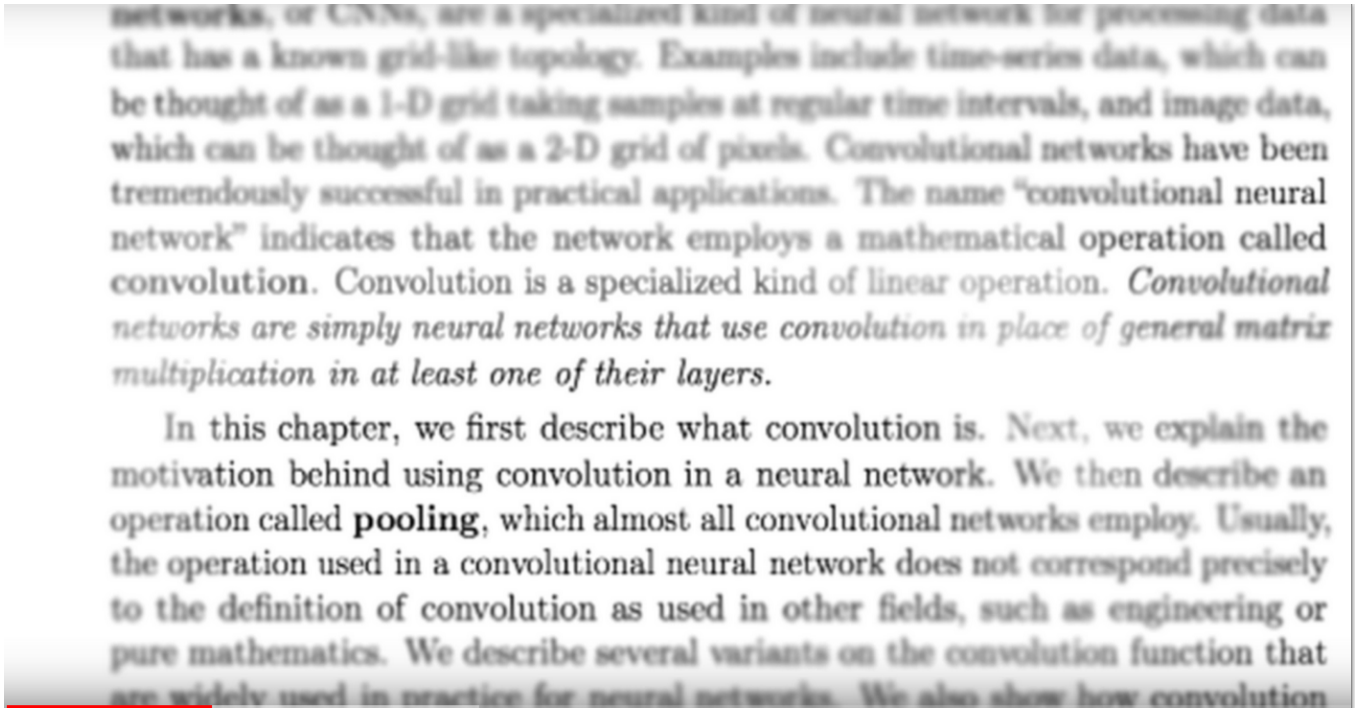


Figure: Translate a sentence in Spanish (sequence) into a sentence in English (sequence)[7]

Attention basic concept - recap



networks, or CNNs, are a specialized kind of neural network for processing data that has a known grid-like topology. Examples include time-series data, which can be thought of as a 1-D grid taking samples at regular time intervals, and image data, which can be thought of as a 2-D grid of pixels. Convolutional networks have been tremendously successful in practical applications. The name "convolutional neural network" indicates that the network employs a mathematical operation called convolution. Convolution is a specialized kind of linear operation. *Convolutional networks are simply neural networks that use convolution in place of general matrix multiplication in at least one of their layers.*

In this chapter, we first describe what convolution is. Next, we explain the motivation behind using convolution in a neural network. We then describe an operation called **pooling**, which almost all convolutional networks employ. Usually, the operation used in a convolutional neural network does not correspond precisely to the definition of convolution as used in other fields, such as engineering or pure mathematics. We describe several variants on the convolution function that are widely used in practice for neural networks. We also show how convolution

Figure: We teach a neural network to focus on the part of the data[1]

Attention basic framework

EEAP

1. Embed

Attention basic framework

EEAP

1. Embed
2. Encode

Attention basic framework

EEAP

1. Embed
2. Encode
3. Attend

Attention basic framework

EEAP

1. Embed
2. Encode
3. Attend
4. Predict

Pet problem to work with

Encoder

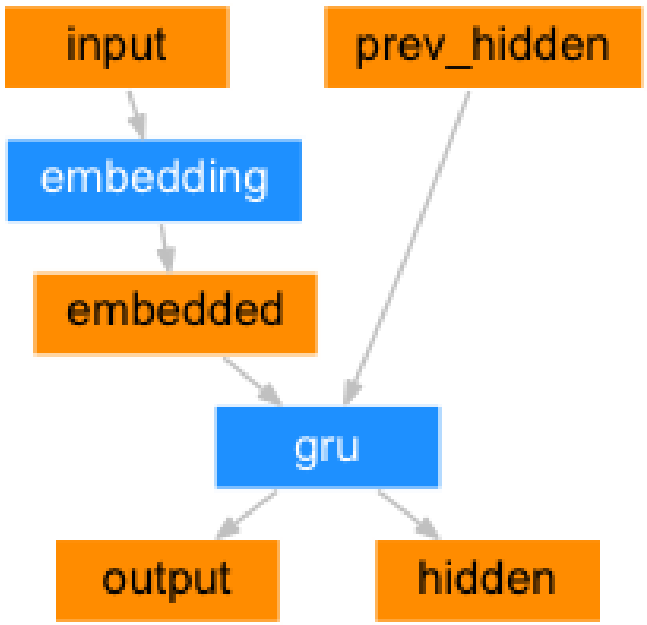


Figure: Our encoder model[8]

[8] https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html

Pet problem to work with

Attention

- We calculate attention weights based on **decoder input** and a hidden state
- We multiply it with **encoder outputs** to create a weighted combination
- It will help the decoder to "find out" which part of encoder output is "responsible" for which part of decoder output

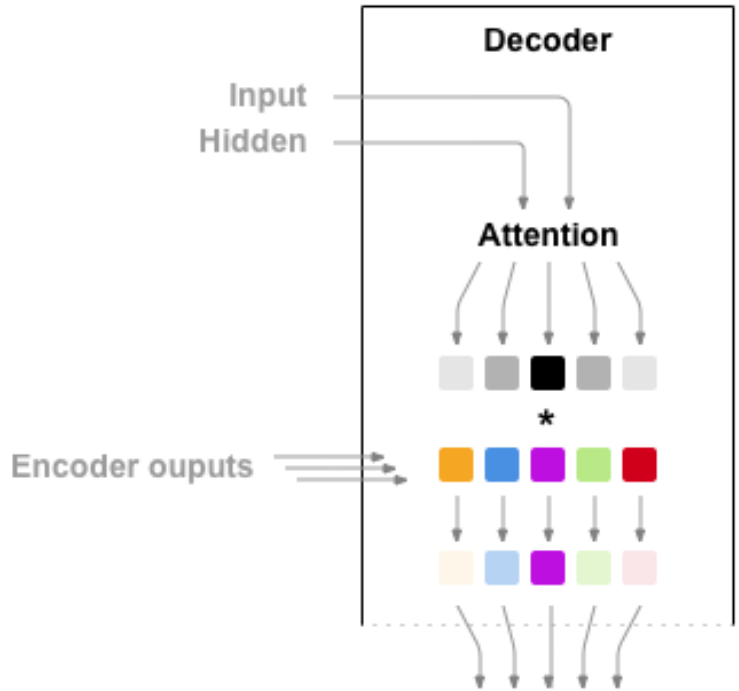


Figure: Way of computing attention[8]

Pet problem to work with

Decoder

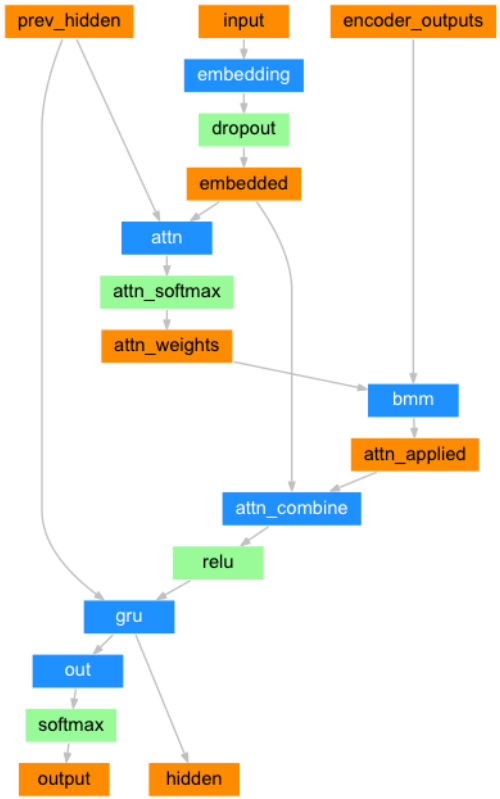


Figure: Our attention-decoder model[8]

Time to code

Now we will do a little bit of live coding

Where can You find the code?

`https://github.com/tugot17/paying-attention-to-attention`

Attention applications

Automatic image captioning

- Nearly the same solution as previously
- Instead of encoder output, latent space from a pre-trained state-of-the-art network (e.g Inception V3)
- Attention method similar as before (e.g Bahdanau Attention)

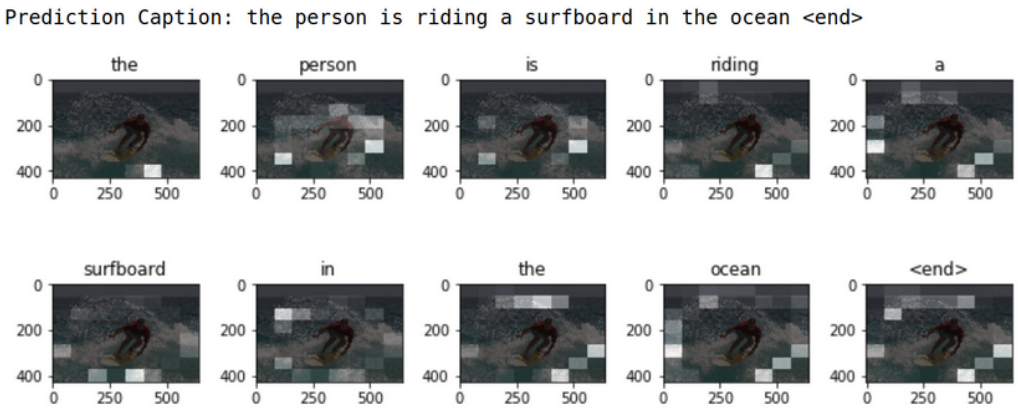


Figure: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[6]

Attention applications

Attention UNet

- State-of-the-art image segmentation solution
- Improving model sensitivity and accuracy by attaching attention gates on top of the standard U-Net

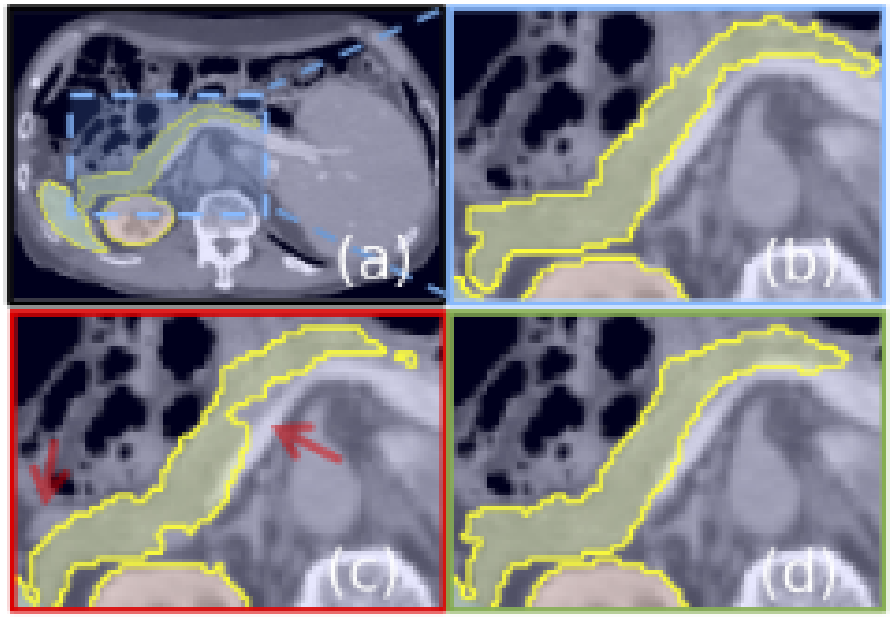


Figure: Segmentation for medical images with U-Net[7]

Attention applications

Latex code generation

- Seq2seq very similar to the Image Captioning
- Uses attention based RNN to generate Latex Code
- Real world application: `https://mathpix.com/`

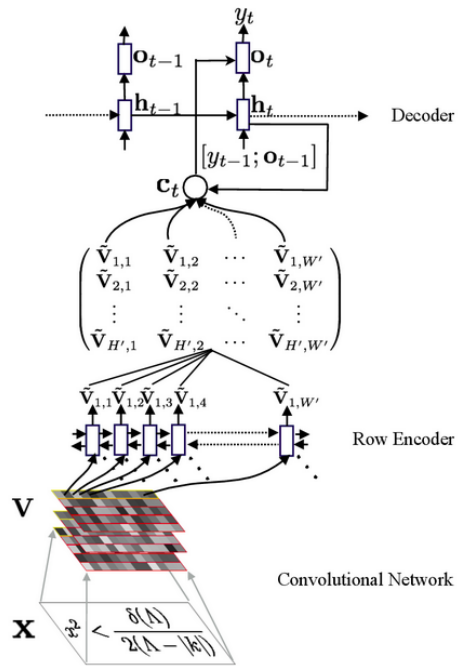


Figure: What You Get Is What You See:A Visual Markup Decompiler[8]

Attention applications

Many others, mainly some kind of seq2seq models

- Bert[5], RoBerta[9], AlBert[10], *Bert*
- GPT 2[2]
- Transformer - Attention is all You need[3]
- Residual Attention Network for Image Classification[4]
- TreeGen: A Tree-Based Transformer Architecture for Code Generation[11]
- and many others ...

Summary

Many others, mainly some kind of seq2seq models

- Attention = teach Your NN to on what part of data should it be focused on

Summary

Many others, mainly some kind of seq2seq models

- Attention = teach Your NN to on what part of data should it be focused on
- Attention is all You need - we can solve a wide variety of problems using attention

Summary

Many others, mainly some kind of seq2seq models

- Attention = teach Your NN to on what part of data should it be focused on
- Attention is all You need - we can solve a wide variety of problems using attention
- We can use nearly same solution for many problems

Summary

Many others, mainly some kind of seq2seq models

- Attention = teach Your NN to on what part of data should it be focused on
- Attention is all You need - we can solve a wide variety of problems using attention
- We can use nearly same solution for many problems
- Embed, Encode, Attend, Predict Framework

Summary

Many others, mainly some kind of seq2seq models

- Attention = teach Your NN to on what part of data should it be focused on
- Attention is all You need - we can solve a wide variety of problems using attention
- We can use nearly same solution for many problems
- Embed, Encode, Attend, Predict Framework
- Hot research topic - It's hard to stay up to date but it is relatively easy to publish

Summary

Many others, mainly some kind of seq2seq models

- Attention = teach Your NN to on what part of data should it be focused on
- Attention is all You need - we can solve a wide variety of problems using attention
- We can use nearly same solution for many problems
- Embed, Encode, Attend, Predict Framework
- Hot research topic - It's hard to stay up to date but it is relatively easy to publish
- There are several methods for calculating attention (e.g Bahdanau Attention, Luong attention, etc.)

References I

[1] Y. B. D. Bahdanau, K. Cho.
Neural machine translation by jointly learning to align and translate.
2014.

[2] A. R. et al.
Language models are unsupervised multitask learners.
2019.

[3] A. V. et al.
Attention is all you need.
2017.

References II

[4] F. W. et al.
Residual attention network for image classification.
2019.

[5] J. D. et al.
Bert: Pre-training of deep bidirectional transformers for language understanding.
2018.

[6] K. X. et al.
Show, attend and tell: Neural image captiongeneration with visual attention.
2016.

References III

[7] O. O. et al.
Attention u-net: Learning where to look for the pancreas.
2018.

[8] Y. D. et al.
What you get is what you see:a visual markup decompiler.
2016.

[9] Y. L. et al.
Roberta: A robustly optimized bert pretraining approach.
2019.

References IV

[10] Z. L. et al.
Albert: A lite bert for self-supervised learning of language representations.
2019.

[11] Z. S. et al.
Treegen: A tree-based transformer architecture for code generation.
2019.

[12] N. L. Z. L. Xue, X. Li.
Not all attention is needed: Gated attention network.
2019.

Thank You for Your attention