



Analysis and Design of Deep Neural Networks

Ahmad Kalhor
Associate Professor
School of Electrical and Computer Engineering
University of Tehran
Spring 2022

Part1

Design of DNNs, by analyzing the well known layers, blocks, modules, and architectures

1. Structure analysis in DNNs

1.1 Layers, Blocks and Modules

- Fully Connected layers and blocks
- Convolution Layers-Blocks-Modules
- Recurrent Layers-Modules
- Attention Layers-Modules
- Pooling Layers
- Normalization Layers

1.2 Architectures

- CNNs
- R-CNNs
- RNNs and Transformers
- AEs and VAEs
- GANs

1.1 Layers, Blocks and Modules

1.1.1. Fully Connected layers and blocks

1.1.2 Convolution Layers-Blocks-Modules

1.1.3 Recurrent and Attention Layers –Modules

1.1.4 Attention Layers –Modules

1.1.5 Pooling and Down-sampling/Up-sampling Layers

1.1.6 Normalizing Layers

1.1.1 Fully Connected layers and their blocks

Ideal to make partitions, maps and encoded/decoded data from a set of distinct inputs.

- One FC layer
- A block of two FC layers
- A block of three FC layers
- A block of more than three FC layers

One FC layer

Definition

- $y = f(Wx + b)$, $W = [w_{ji}]$ $b = [b_j]$
- $x \in \mathbb{R}^n$ $y \in \mathbb{R}^m$ $i \in \{1, \dots, n\}$ $j \in \{1, \dots, m\}$
- Activation function:
- $f \in \{\text{sign}, \text{step}, \text{tanh}, \text{sigmoid}, \text{Relu}, \text{identity} \dots\}$

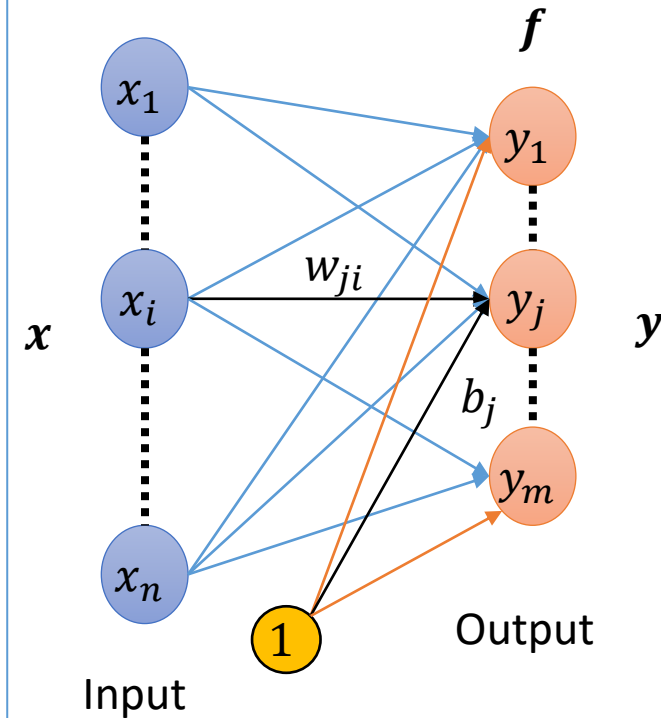
Functionality

(1) Forming an " n "dimensional hyper plane:

$$y_{inj} = w_{j1}x_1 + \dots + w_{jn}x_n + b_j$$

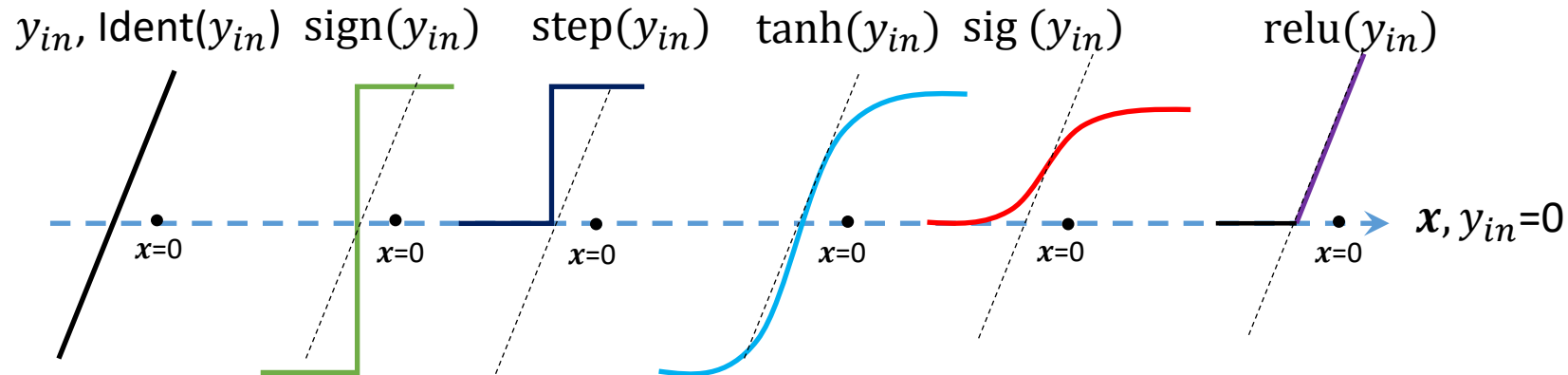
(2) Folding (hard/soft), Rectifying... on the hyper plane:

$$y_j = f(y_{inj})$$

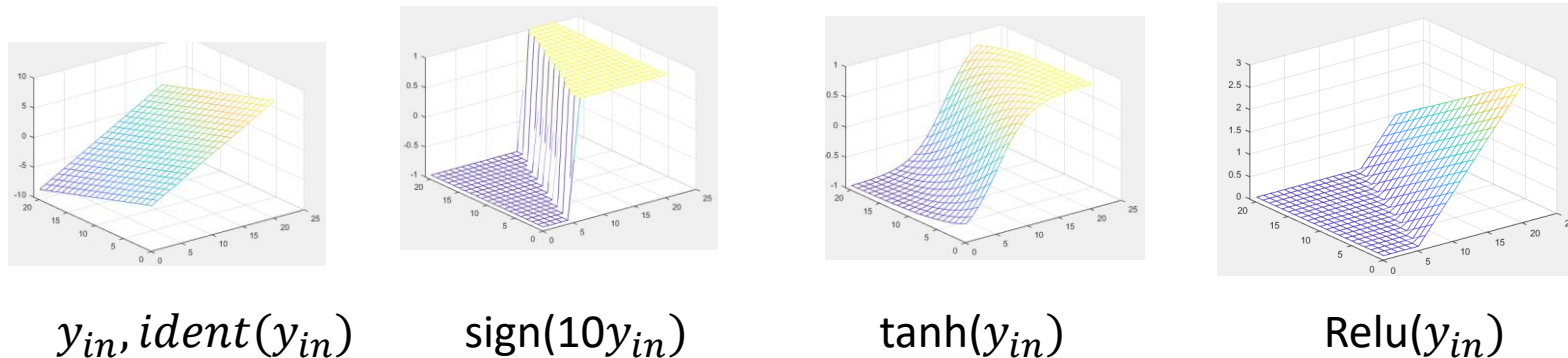


z

Example(1) for $n = 1$, $f(y_{in})$, $y_{in} = 2x + 1$ w :slope parameter , b : ($up \uparrow$ $down \downarrow$) shift parameter



Example(2) for $n = 2$, $f(y_{in})$, $y_{in} = 2x_2 - x_1 + 0$



❖ A “one FC layer” can transform the input space to a (hard or soft) folded or rectified hyper plane

Some Notes about one FC layer

1. One FC layer (with n inputs and m outputs) is formed from m neurons at output, where each neuron is connected to all n input units (fully connected).
2. Each input send a **weighted** signal to each neuron.
3. **For each neuron, the summation of weighted inputs makes an independent hyper-plane just before activated by it.**
4. **The parameters of all hyper-planes are opted through a learning process.**
5. **For each neuron, the corresponding hyper-plane is hardly or softly folded or rectified just after activating by it.**
6. **Indeed, taking in inputs as a " n " dimensional vector, the FC layer provides " m " folded, or rectified hyper-planes at the output.**
7. **Assuming the inputs are binary or bipolar, and the activation functions make binary or bipolar values, a neuron in one FC layer can operate as a simple logic gate like "and", "or, etc. (M&P neuron)**
8. **Assuming the activation functions of the neurons are identity, a FC layer operates as a linear transformer, which can approximate a linear regression model in a regression problem.**
9. **Assuming " r " independent linear or nonlinear correlations among inputs, the order of the formed hyper plan is " $n-r$ ".**

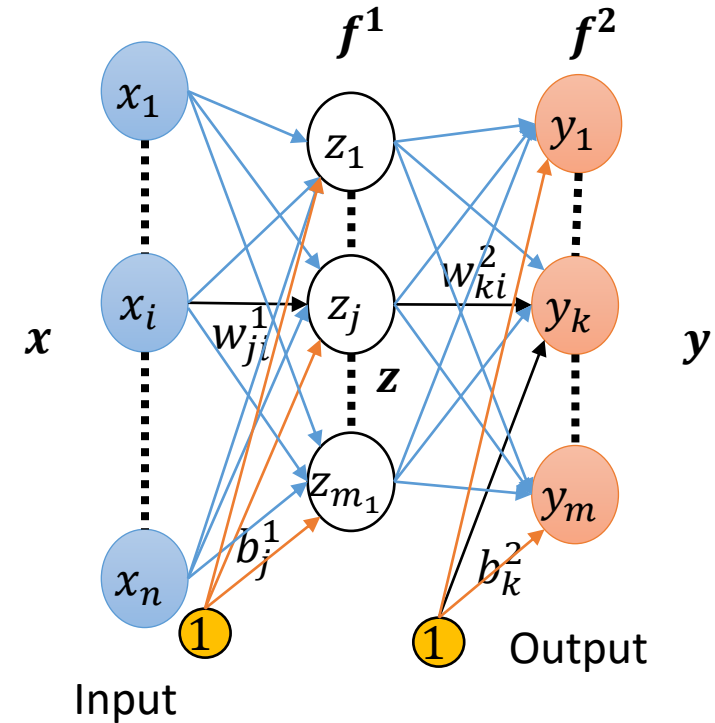
A block of two FC layers

Definition

- $y = f^2 \left(W^2 \underbrace{f^1(W^1 x + b^1)}_z + b^2 \right)$
- $x \in \mathbb{R}^n, z \in \mathbb{R}^{m_1}, y \in \mathbb{R}^m \quad i \in \{1, \dots, n\} \quad j \in \{1, \dots, m_1\} \quad k \in \{1, \dots, m\}$
- Activation functions:
- $f^{1,2} \in \{\text{sign}, \text{step}, \text{tanh}, \text{sigmoid}, \text{Relu}, \text{identity} \dots\}$

Overall Functionality

- (1) A hyper plane: $z_{in_j} = w_{j1}^1 x_1 + \dots + w_{jn}^1 x_n + b_j^1$
- (2) Folding, Rectifying,.. on the hyper plane: $z_j = f^1(z_{in_j})$
- (3) A hyper plane: $y_{in_k} = w_{k1}^2 z_1 + \dots + w_{km_1}^2 z_{m_1} + b_k^2$
- (4) Folding, Rectifying,.. on the hyper plane: $y_k = f^2(y_{in_k})$



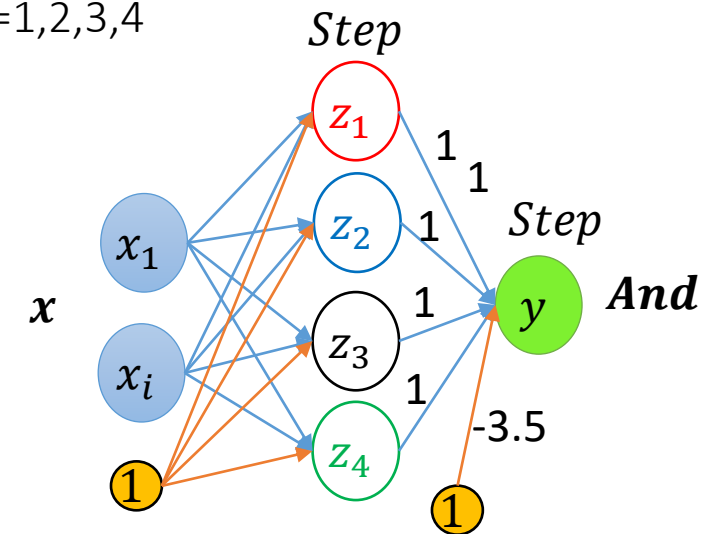
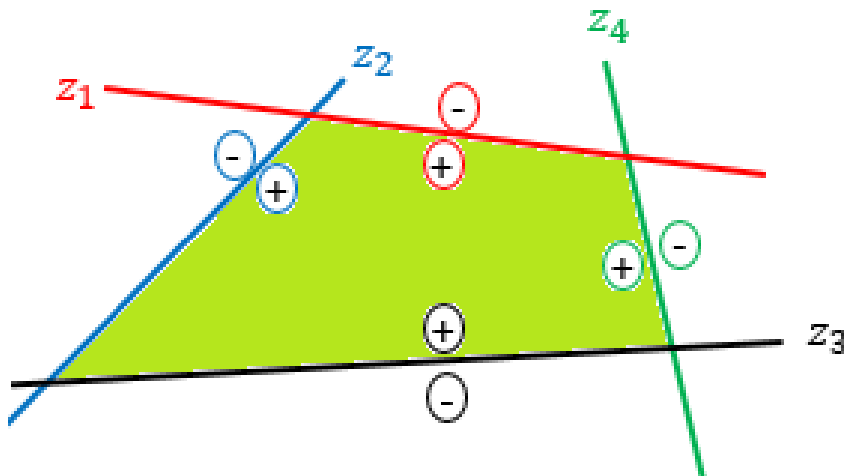
A Desired Functionality in classification /Regression problems

(1) A hyper plane: $z_{in_j} = w_{j1}^1 x_1 + \dots + w_{jn}^2 x_n + b_j^1$

(2) Folding the hyper plane: $z_j = f^1(z_{in_j})$

(3) An “and” or “or” logic gate for former folded hyper-planes is defined by “*kth*” neuron of last layer, by which “*kth*” convex hyper-polygon will be resulted.

Example for $n = 2, m_1 = 4, m=1$ $z_j = w_{j1}^1 x_1 + w_{j2}^1 x_2 + b_j^1, \quad j=1,2,3,4$



- ❖ A block of two FC layers can transform the input space to multi convex hyper- polygon partitions
- ❖ Using soft activation functions, a block of two FC layers can transform the input space to multi soft convex hyper- polygon maps

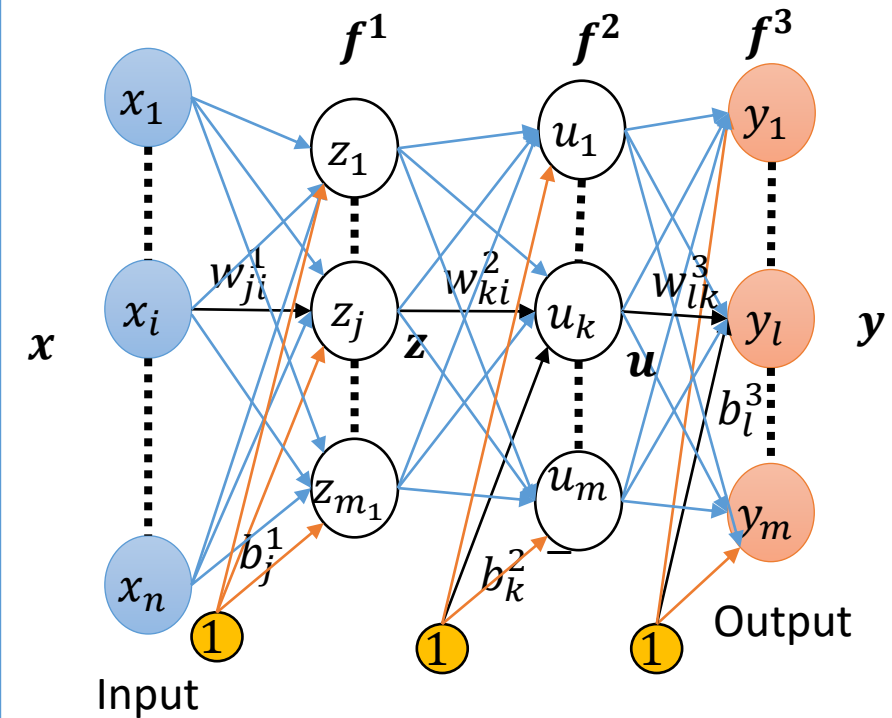
A block of three FC layers

Definition

- $y = f^3 \left(\underbrace{W^3 f^2 \left(\underbrace{W^2 \underbrace{f^1(W^1 x + b^1)}_z + b^2}_u \right) + b^3} \right)$
- $x \in \mathbb{R}^n \quad z \in \mathbb{R}^{m_1} \quad u \in \mathbb{R}^{m_2} \quad y \in \mathbb{R}^m$
- $i \in \{1, \dots, n\} \quad j \in \{1, \dots, m_1\} \quad k \in \{1, \dots, m_2\} \quad l \in \{1, \dots, m\}$
- Activation functions:
- $f^{1,2,3} \in \{\text{sign}, \text{step}, \text{tanh}, \text{sigmoid}, \text{Relu}, \text{identity} \dots\}$

Overall Functionality

- (1) A hyper plane: $z_{in_j} = w_{j1}^1 x_1 + \dots + w_{jn}^1 x_n + b_j^1$
- (2) Folding, Rectifying... on the hyper plane: $z_j = f^1(z_{in_j})$
- (3) A hyper plane: $u_{in_k} = w_{k1}^2 z_1 + \dots + w_{km_1}^2 z_{m_1} + b_k^2$
- (4) Folding, Rectifying... on the hyper plane: $u_k = f^2(u_{in_k})$
- (5) A hyper plane: $y_{in_l} = w_{l1}^3 u_1 + \dots + w_{lm_2}^3 u_{m_2} + b_l^3$
- (6) Folding, Rectifying... on the hyper plane: $y_l = f^3(y_{in_l})$



A Desired Functionality in classification /Regression problems

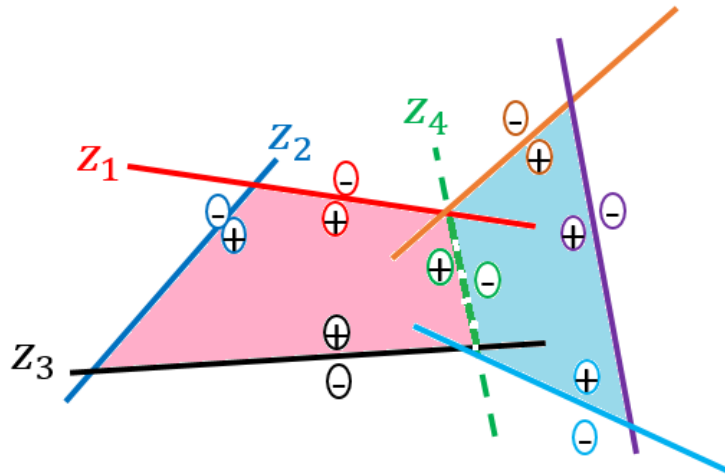
(1) A hyper plane: $z_{in_j} = w_{j1}^1 x_1 + \dots + w_{jn}^1 x_n + b_j^1$

(2) Folding the hyper plane: $z_j = f^1(z_{in_j})$

(3) An “and” or “or logic gate for former folded hyper-planes is defined by “ k th” neuron of second hidden layer 3, by which “ k th” convex hyper-polygon will be resulted.

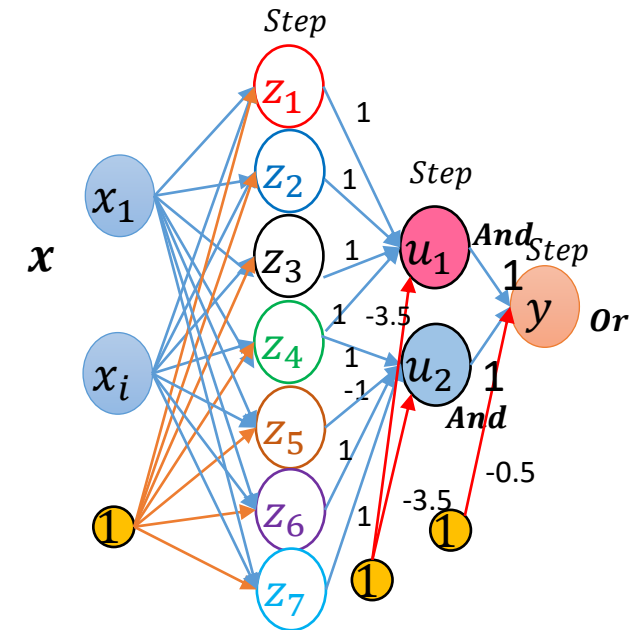
(4) An “and” or “or logic gate for former convex hyper-planes is defined by “ l th” neuron of last layer, by which “ l th” non-convex hyper-polygon will be resulted.

Example for $n = 2, m_1 = 7, m_2 = 2, m = 1$



$$z_j = w_{j1}^1 x_1 + w_{j2}^1 x_2 + b_j^1, \quad j=1,2,3,4,\dots,7$$

$$u_k = w_{k1}^2 z_1 + w_{k2}^2 z_2 + b_k^2, \quad k=1,2$$

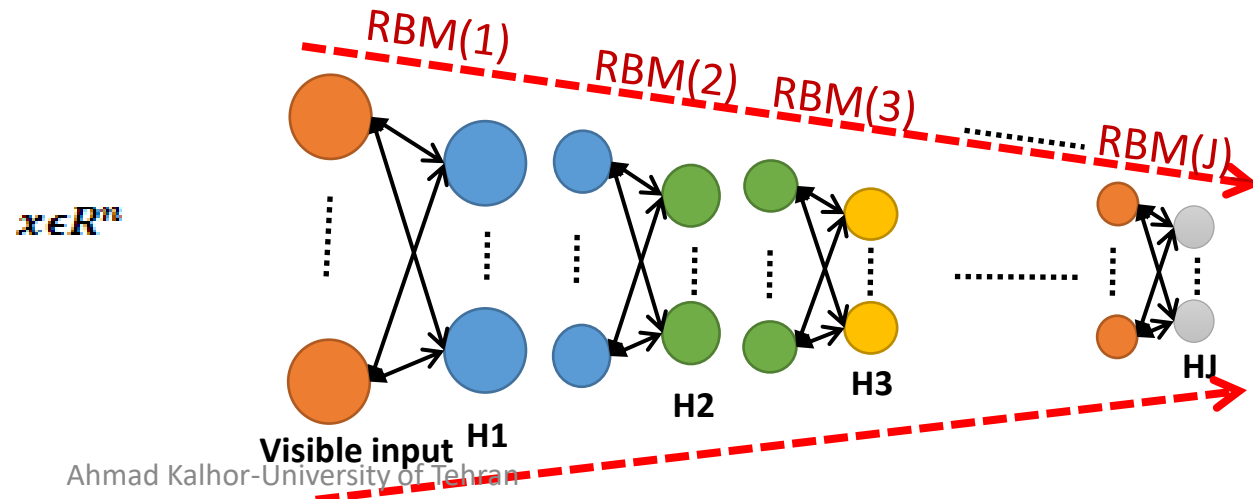
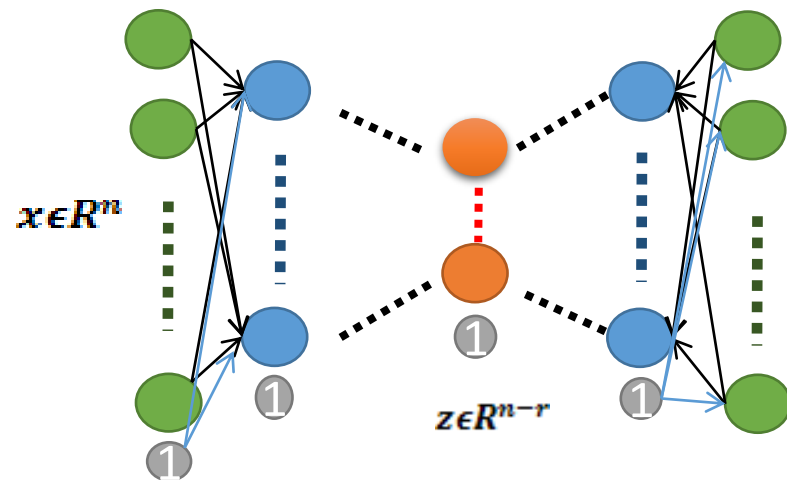


- ❖ A block of two FC layers can transform the input space to multi non-convex hyper- polygon partitions
- ❖ Using soft activation functions, a block of two FC layers can transform the input space to multi non-convex hyper- polygon volume maps

Some Notes about block of FC layers

1. A block of two or three FC layers are known as universal function approximator, through which any function by any desired accuracy can be approximated.
2. In comparison to a block of two FC layers, a block of three FC layers has better extrapolation (generalization) for unbounded regions and non-convex regions.

A block of FC layers with more than three layers are conventionally used in deep auto-encoders and cascaded restricted Boltzmann machines. In such blocks, each layer learns to encode or decode the taken data with a low change.



Some notes about FC layers

1. Fully connected layers are applied to a set of inputs reshaped as a vector; the spatial or temporal correlations among the inputs are not considered in its operation.
2. For high dimensional data, due to their massive wirings, they suffer from large memories, high computation load, and overfitting.
3. Although, they are appropriate for partitioning and mapping purposes, they are not ideal to remove disturbances, and filter and extract features from temporal, spatial, and multi modal signals.

1.1.2 Convolution Layers-Blocks-Modules*

Ideal to filter and encode/decode various, spatial, temporal and multi modal signals

- Simple Convolution
- Tiled Convolution
- Dilated Convolution
- Deconvolution (Transposed convolution)
- 1x1 Convolutions
- Flattened Convolutions
- Spatial and Cross-Channel convolutions
- Depth-wise Separable Convolutions
- Residual Blocks and types
- Grouped Convolutions
- Shuffled Grouped Convolutions

* Most of examples are from <https://ikhlestov.github.io/pages/machine-learning/convolutions-types/>, Illarion Khlestov.

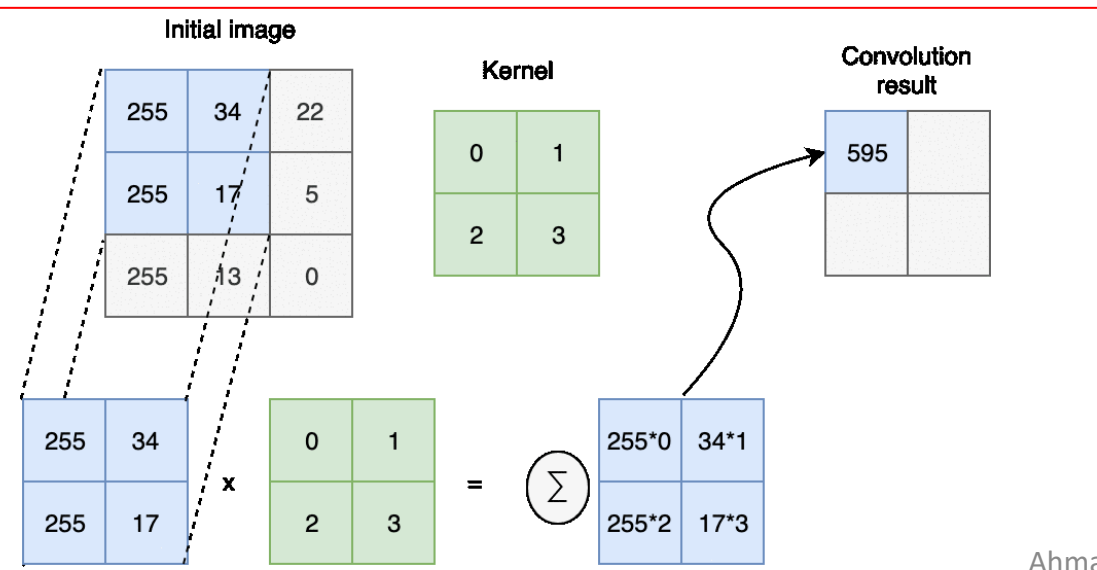
Ahmad Kalhor-University of Tehran

Simple Convolution

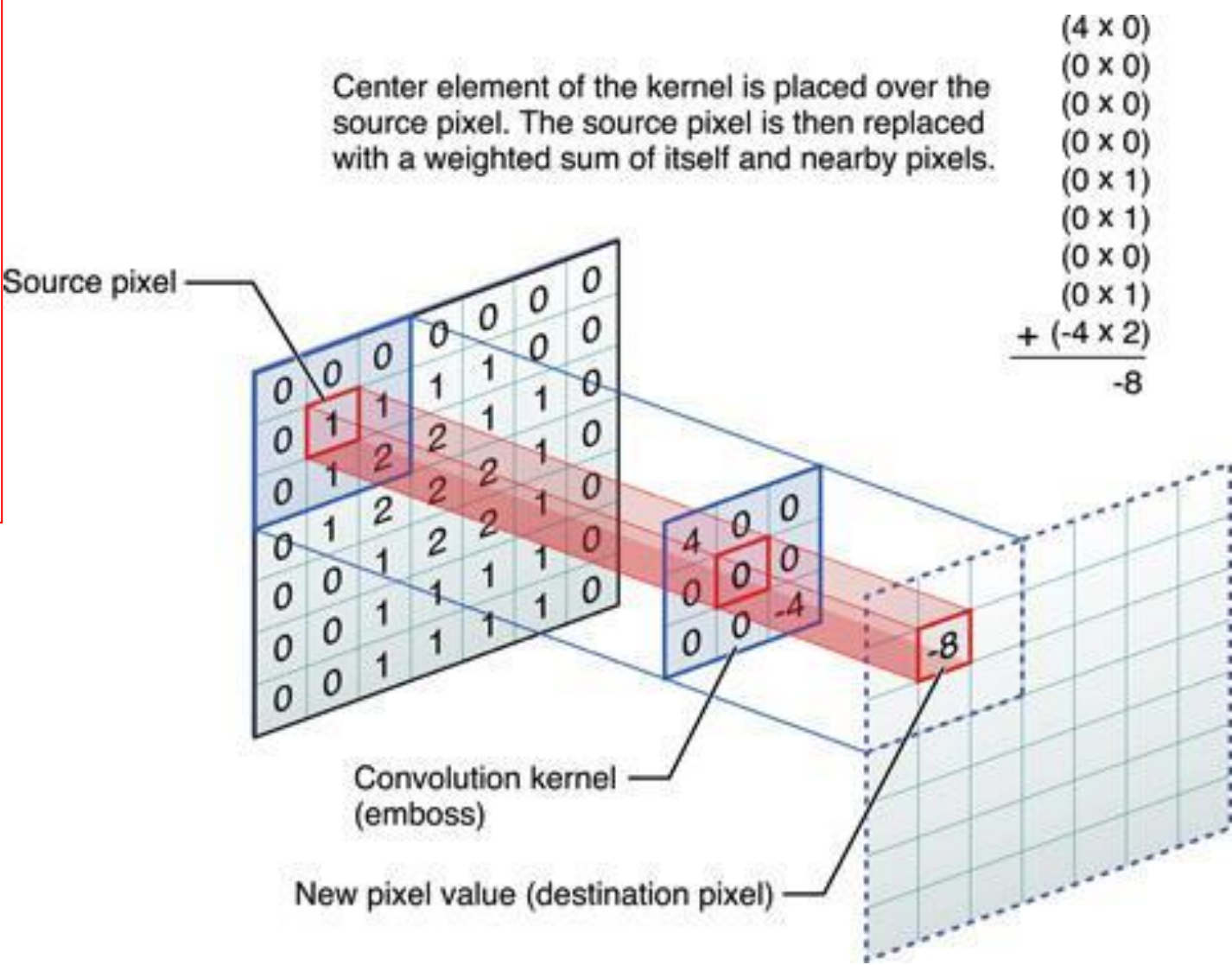
Take multiply dot products with same filter with some width/height shift.

Interesting because:

- Weights sharing and local connection
- it can capture and intensify all those patches which are adequately similar to the kernel and remove other ones by “relu”
- It can be interpreted as a filter that remove all patches which have weak linear correlation with the kernel



Example of convolution on two dimensional signal (image)

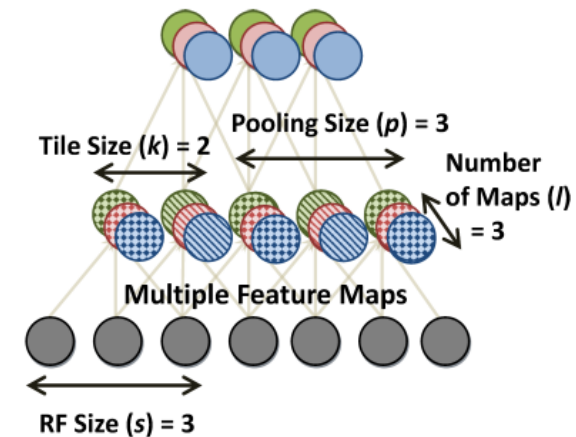
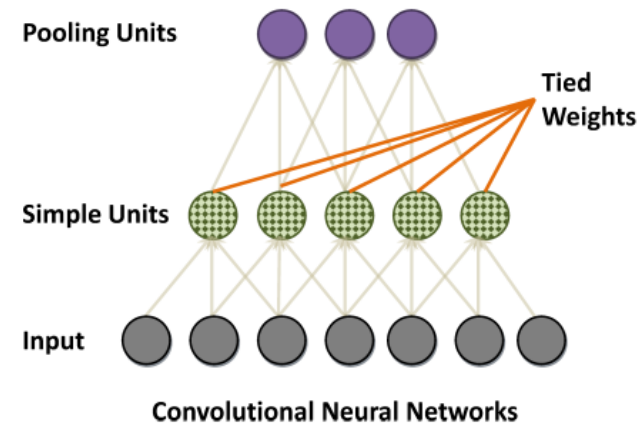
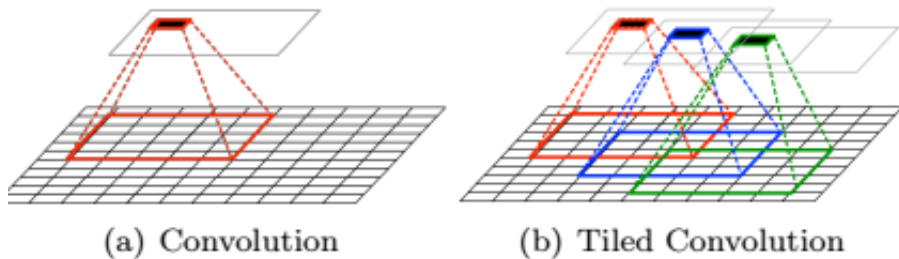


Tiled Convolution

- It means that we can learn multiples feature maps rather than one feature map by considering several different filters
- In tiled convolution we can provide different kinds of invariance.
- Separate kernels are learned within the same layer

Example of Tiled convolution on one dimensional signal (time series)

Example of Tiled convolution on two dimensional signal



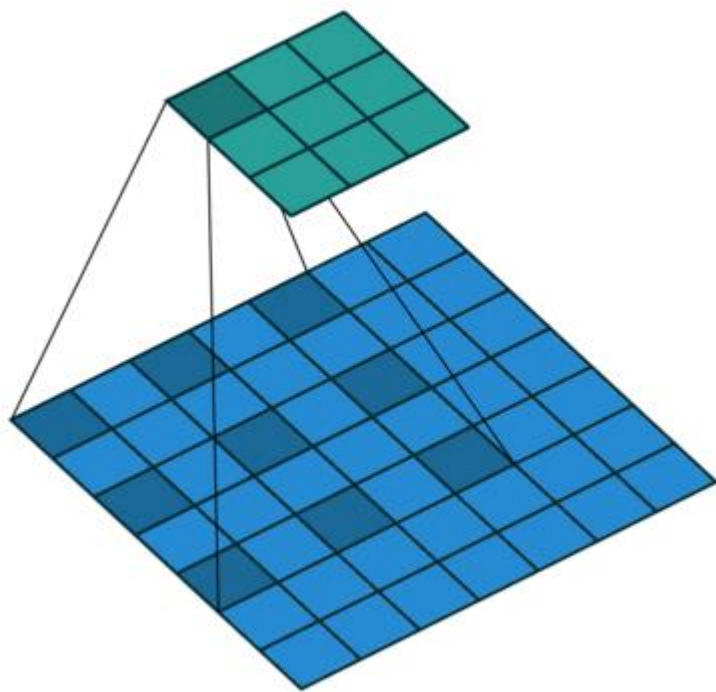
Some notes about the convolution layer

1. The convolution layer is a variant of the convolution operation used in LTI systems
2. A convolution layer actually transform an input (spatial, temporal, or multi modal) signal to several feature maps.
3. All units of feature maps just after convolution may be activated by applying an activation function like "relu".
4. Using a filter, each feature map captures a certain feature from different local regions of the former signal.
5. Features, which are captured by filters, are actually frequently repeating sub-patterns within a signal.
6. The kernel size and the number of filters depend to the size and the number of the existing independent features, respectively.
7. Using stride=1 and padding, the size of a feature map is equal to the spatial size of the signal but using stride>1 (stride<1), the signal becomes down-sampled (up-sampled) by the convolution layer.
8. The resulting feature maps from a convolution layer, can be convolved again through a new convolution layer.
9. Through using a block of sequenced convolution layers (may come with activation functions, pooling and normalizing layers), actually redundancies and disturbances are removed and exclusive features for appropriate classification or regression problem will be appeared.

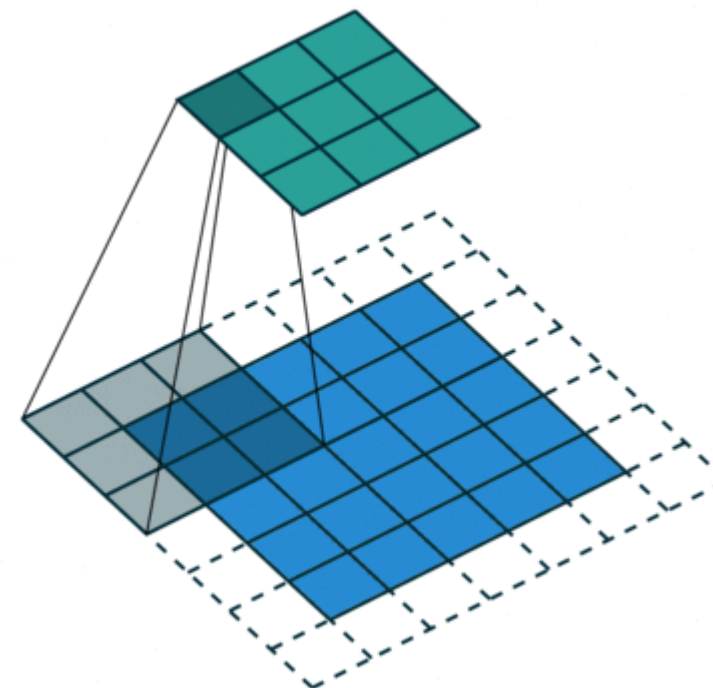
Dilated Convolution

To convolve and down-sample signals with very large spatial/temporal size

It is a **technique that expands the kernel (input) by inserting holes between its consecutive elements**. In simpler terms, it is same as convolution but it involves pixel skipping, so as to cover a larger area of the input. ... In essence, normal convolution is just 1-dilated convolution



Dilated Convolution ($l=2$)



Standard Convolution ($l=1$)

Deconvolution (Transposed Convolution)

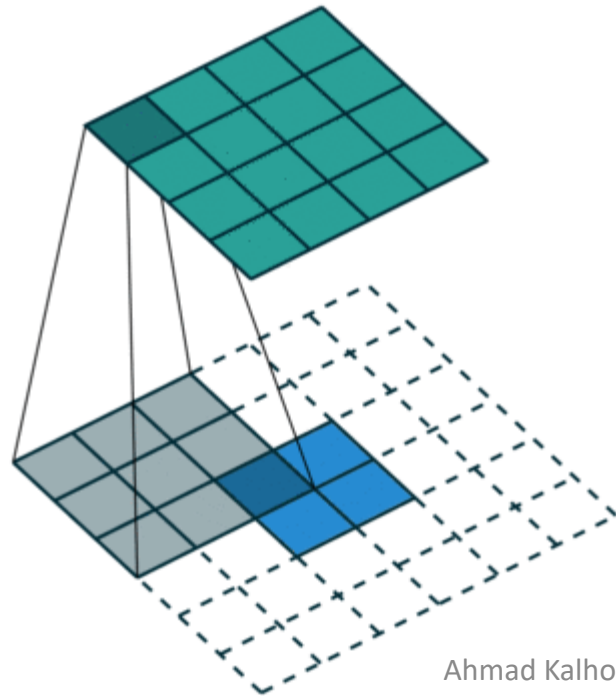
Transposed convolution can be seen as the backward pass of a corresponding traditional convolution.

Transpose Convolution is a convolution layer which reverses the operation done by the corresponding convolution.

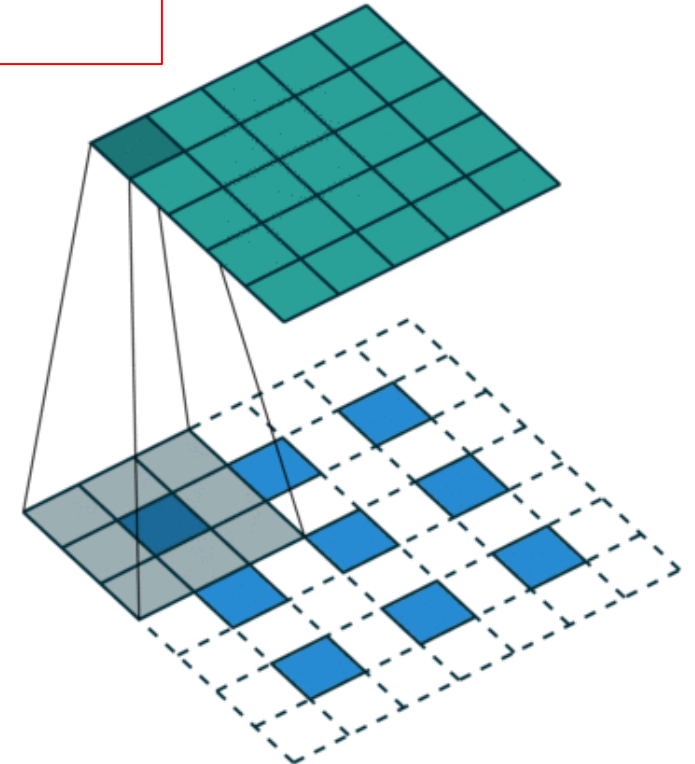
High pass Filter (corresponding Conv.) \rightarrow Low pass Filter (Tran. Conv.)

Smooth Filter (corresponding Conv.) \rightarrow Difference Filter (Tran. Conv.)

Visually, for a transposed convolution with stride one and no padding, we just pad the original input (blue entries) with zeroes (white entries).



In case of stride two and padding, the transposed convolution would look like this

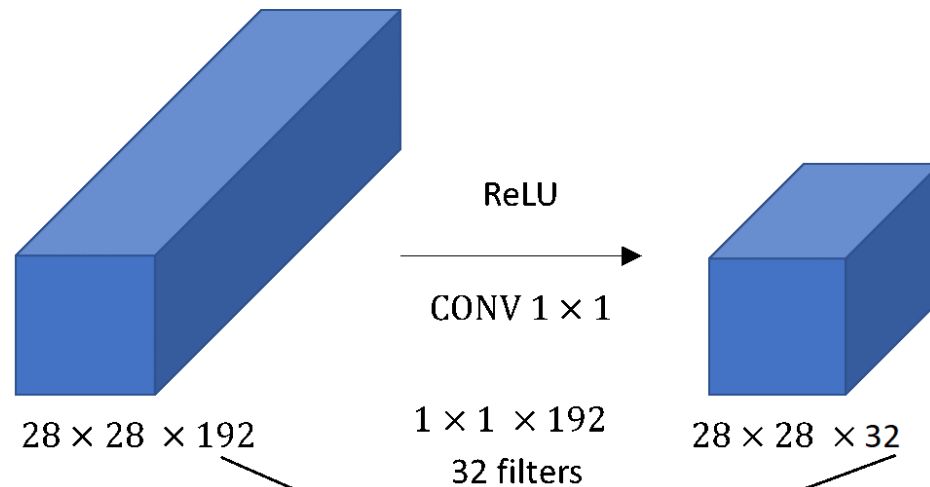


1x1 Convolutions

Initially 1x1 convolutions were proposed at [Network-in-network\(NiN\)](#). After they were highly used in [GoogleNet architecture](#). Main features of such layers:

- Reduce or increase dimensionality
- Apply nonlinearity again after convolution
- Can be considered as “feature pooling”

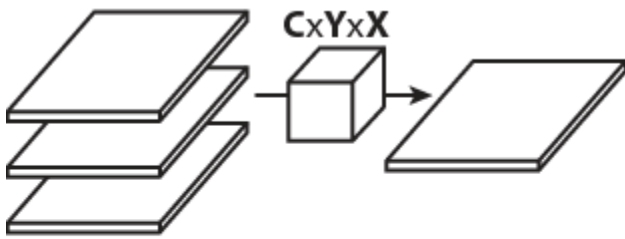
They are used in such way: we have image with size $28 \times 28 \times 192$, where 192 means features, and after applying 32 1x1 convolutions filters we will get images with $28 \times 28 \times 32$ dimensions.



Flattened Convolutions

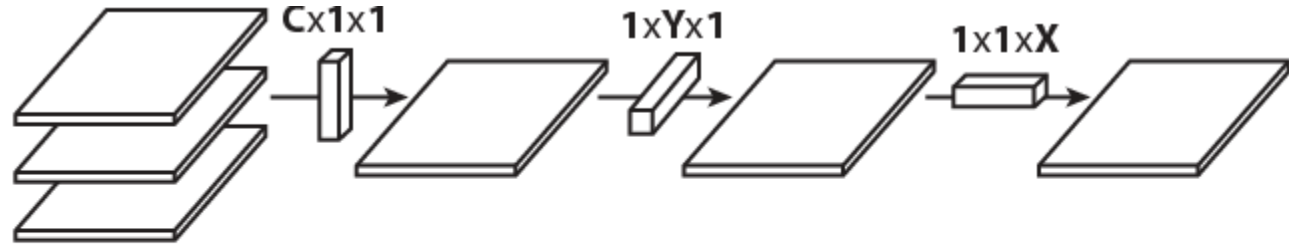
Were published in [Flattened Convolutional Neural Networks for Feedforward Acceleration](#).

Reason of usage same as 1×1 convs from NiN networks, but now not only features dimension set to 1, but also one of another dimensions: width or height.



(a) 3D convolution

The number of operations at each feature map is $= a * a * (C * Y * X)$

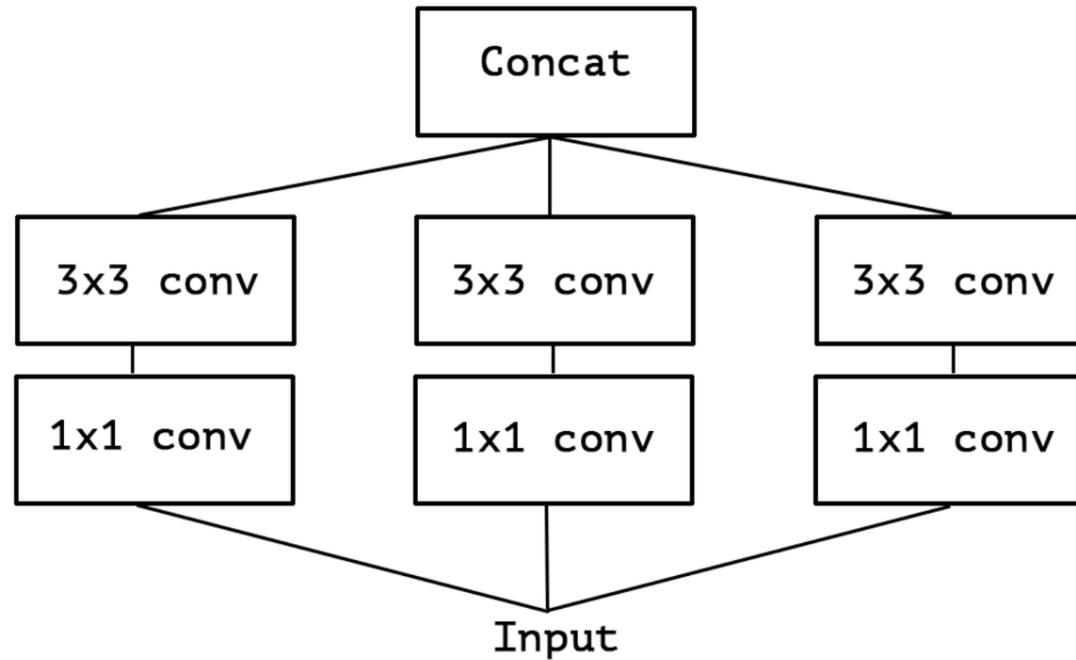


(b) 1D convolutions over different directions

The number of operations at each feature map is $= a * a * (C + Y + X)$

Spatial and Cross-Channel convolutions

First this approach was widely used in **Inception network**. Main reason is to split operations for cross-channel correlations and at spatial correlations into a series of independently operations. Spatial convolutions means convolutions performed in **spatial dimensions** - **width** and **height**



- ❖ Grouping convolutions in independent parallel paths cause acceleration in computations and reduction in the required memory

Inception Module

Inception module is introduced by Szegedy *et al.* * which can be seen as a logical culmination of NIN. They use variable filter sizes to capture different visual patterns of different sizes, and approximate the optimal sparse structure by the inception modul

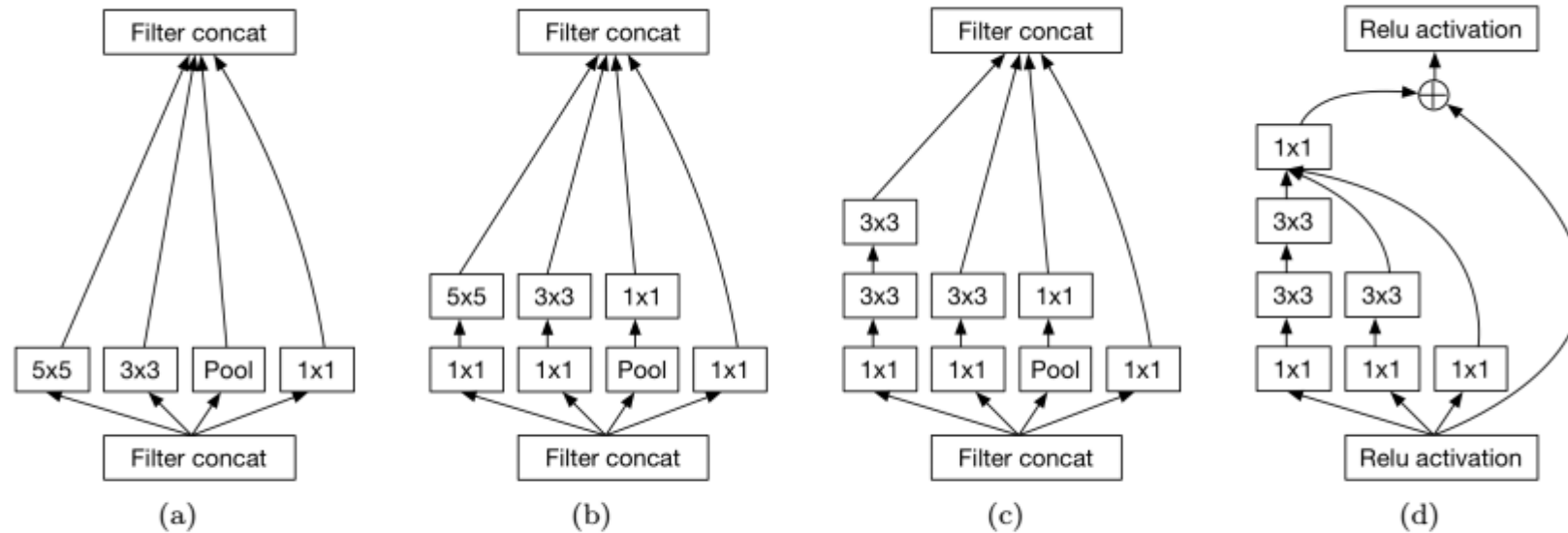


Figure 5: (a) Inception module, naive version. (b) The inception module used in [10]. (c) The improved inception module used in [41] where each 5×5 convolution is replaced by two 3×3 convolutions. (d) The Inception-ResNet-A module used in [42].

❖ Using parallel convolution paths with different kernel sizes increases the chance of better feature extraction

* C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

Depthwise Separable Convolutions

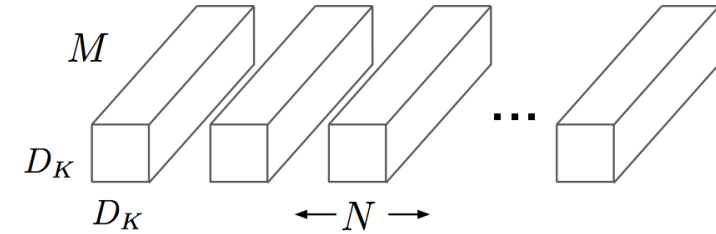
A lot about such convolutions published in the ([Xception paper](#)) or ([MobileNet paper](#)).

Consist of:

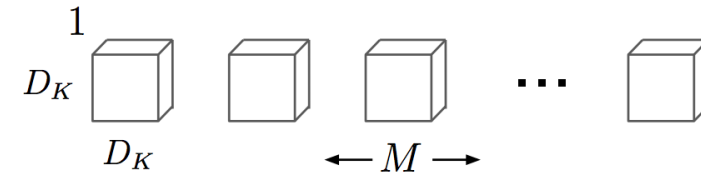
- **Depthwise convolution**, i.e. a spatial convolution performed independently over each channel of an input.
- **Pointwise convolution**, i.e. a 1×1 convolution, projecting the channels output by the depthwise convolution onto a new channel space.

Difference between Inception module and separable convolutions:

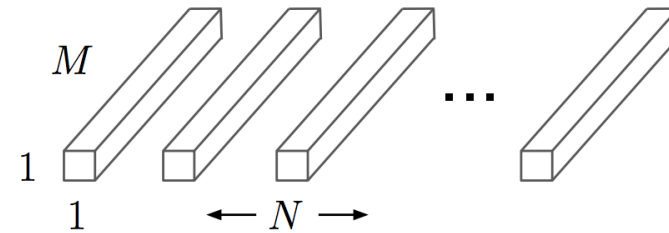
- Separable convolutions perform first channel-wise spatial convolution and then perform 1×1 convolution, whereas Inception performs the 1×1 convolution first.
- depthwise separable convolutions are usually implemented without non-linearities



(a) Standard Convolution Filters



(b) Depthwise Convolutional Filters



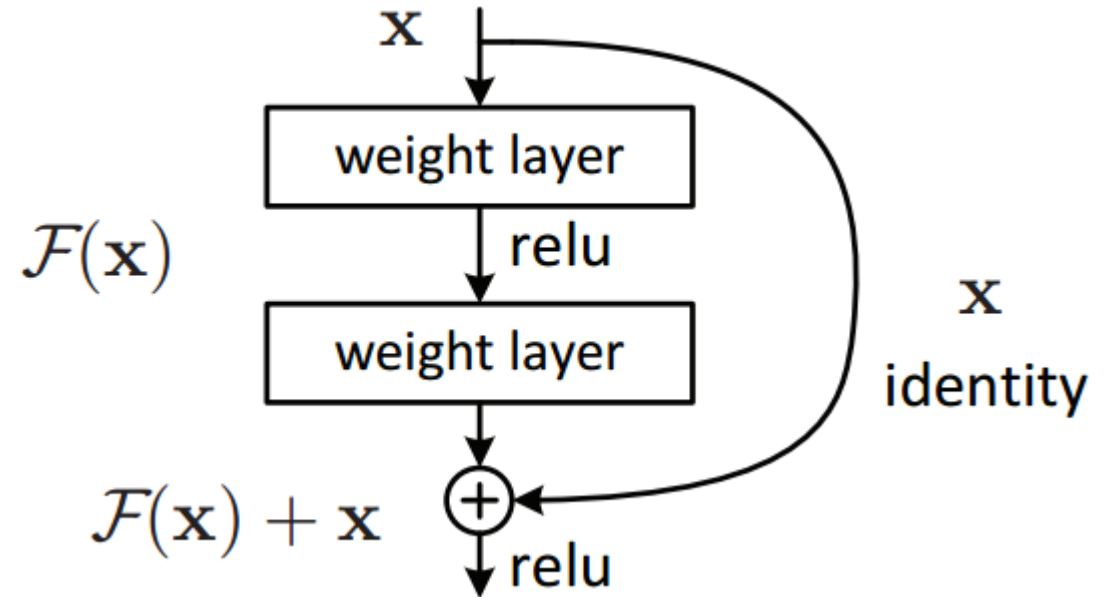
(c) 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

❖ Using **Depthwise** convolution decreases the computation load of 3D-convolution

Residual Blocks

To avoid missing information and provide a solution for vanishing gradient

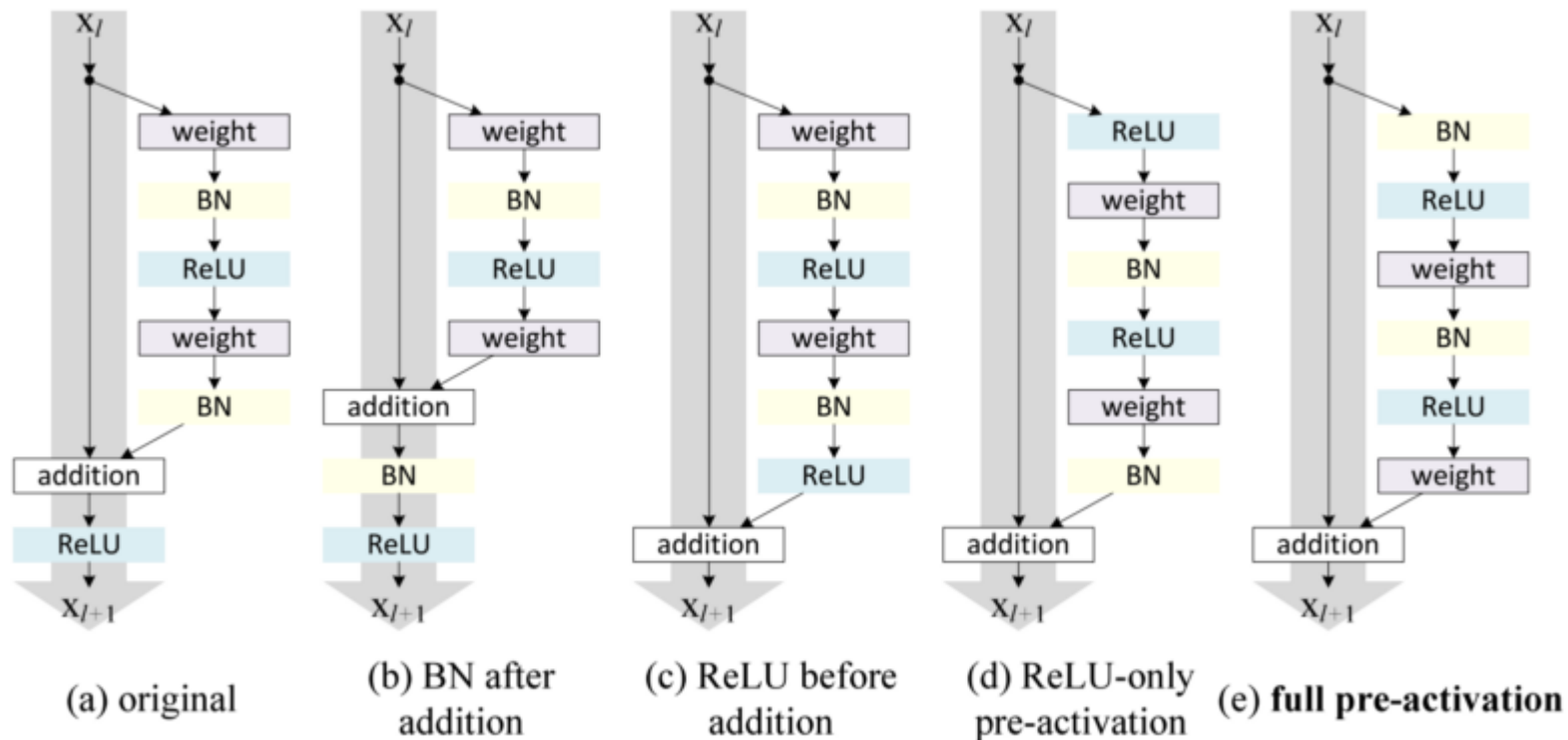
In traditional neural networks, each layer feeds into the next layer. In a network with residual blocks, each layer feeds into the next layer and directly into the layers about 2–3 hops away. That's it. But understanding the intuition behind why it was required in the first place, why it is so important, and how similar it looks to some other state-of-the-art architectures is where we are going to focus on. There is more than one interpretation of why residual blocks are awesome and how & why they are one of the key ideas that can make a neural network show state-of-the-art performances on a wide range of tasks.



❖ Applying an identity path to convolution, residual block causes that features to be captured in a difference evolving learning manner without missing information.

Different forms of residual blocks

The image below shows how to arrange the residual block and identity connections for the optimal gradient flow. It has been observed that pre-activations with batch normalizations generally give the best results (i.e., the right-most residual block in the image below gives the most promising results).



Grouped Convolutions

Grouped convolutions were initially mentioned in AlexNet, and later reused in [ResNeXt](#). Main motivation of such convolutions is to reduce computational complexity while dividing features on groups.

The image below shows multiple interpretations of a residual block.

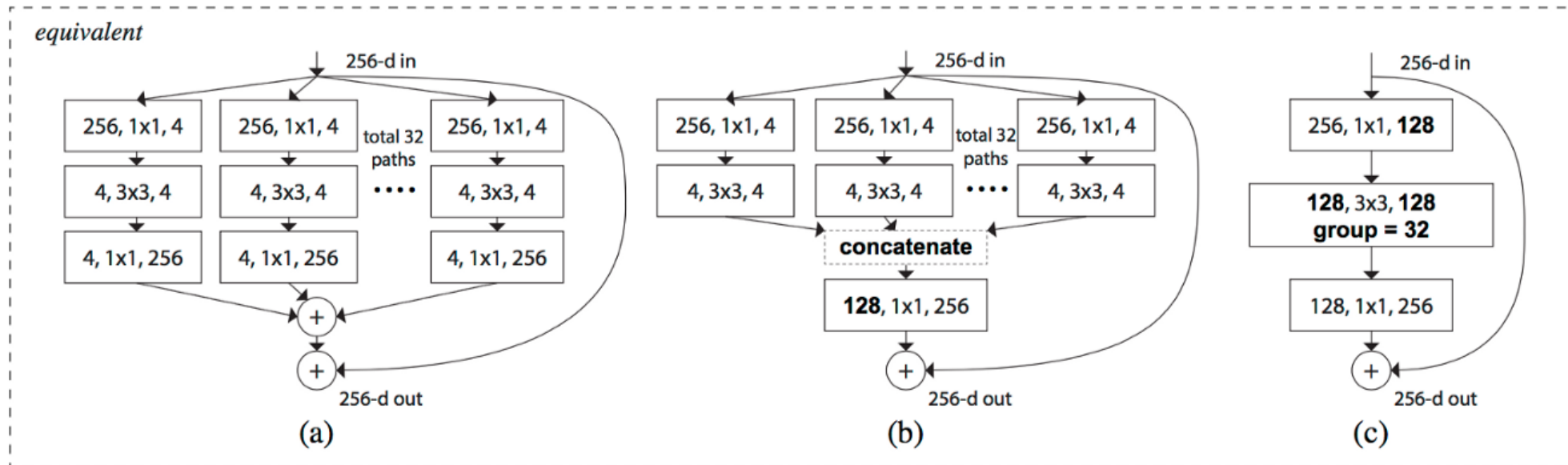


Figure 3. Equivalent building blocks of ResNeXt. **(a)**: Aggregated residual transformations, the same as Fig. 1 right. **(b)**: A block equivalent to (a), implemented as early concatenation. **(c)**: A block equivalent to (a,b), implemented as grouped convolutions [24]. Notations in **bold** text highlight the reformulation changes. A layer is denoted as (# input channels, filter size, # output channels).

Shuffled Grouped Convolutions

[Shuffle Net](#) proposed how to eliminate main side effect of the grouped convolutions that “outputs from a certain channel are only derived from a small fraction of input channels”.

They proposed shuffle channels in such way(layer with $g \times n \times n$ channels whose output has $g \times n \times n$ channels):

- reshape the output channel dimension into $(g, n)(g, n)$
- transpose output
- flatten output back

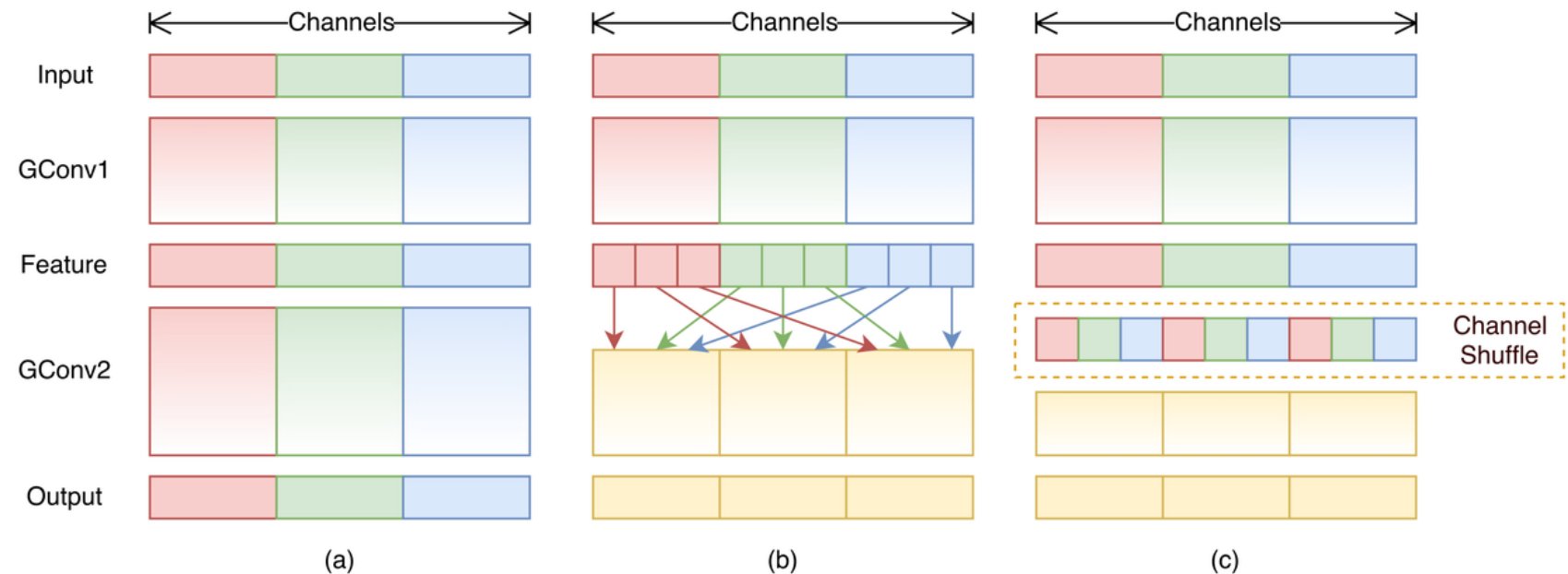


Figure 1: Channel shuffle with two stacked group convolutions. GConv stands for group convolution. a) two stacked convolution layers with the same number of groups. Each output channel only relates to the input channels within the group. No cross talk; b) input and output channels are fully related when GConv2 takes data from different groups after GConv1; c) an equivalent implementation to b) using channel shuffle.

Some notes about convolution layers

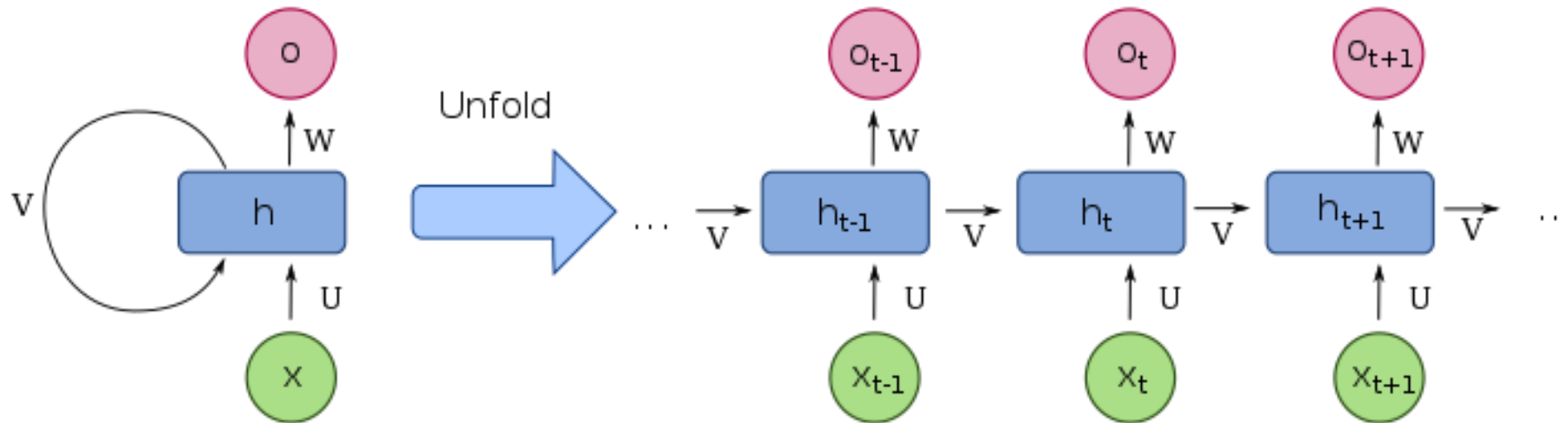
1. There are many extensions of simple convolutional layers.
2. Deconvolution layers are introduced as the backward pass of their corresponding convolution.
3. Residual blocks use a direct skip connection from inputs to outputs(as an identity function) to avoid missing information and to solve the problem of vanishing gradient. In addition, many other extended convolution layers use such skip connection in their structures.
4. Some convolutional layers such as "NiN" and flattened convolutions provide a considerable drop in computations by using point-wise and depth-wise convolution instead of using simple 3D convolution operations.
5. Some convolutional layers such as inception module by using some parallel convolutions with different resolutions have tried to provide better accuracy.
6. Grouped convolutions as well as shuffled grouped convolutions use several similar form and parallel convolution blocks (with an identity skip connection) in their structures. Each convolution block includes point-wise and depth-wise convolutions.

1.1.3 Recurrent Neural Networks

Ideal to retrieve short/long dependencies among the sequenced samples of a signals

- Simple RNN Module
- LSTM Module
- GRU Module
- Bidirectional RNN layer
- Bidirectional Deep RNNs
- TimeDistributed layer
- ConvLSTM layer

Simple RNN Module



$$\mathbf{h}_t = \mathbf{f}(\mathbf{V}\mathbf{h}_{t-1} + \mathbf{U}\mathbf{x}_t)$$

$$\mathbf{o}_t = \mathbf{g}(\mathbf{W}\mathbf{h}_t)$$

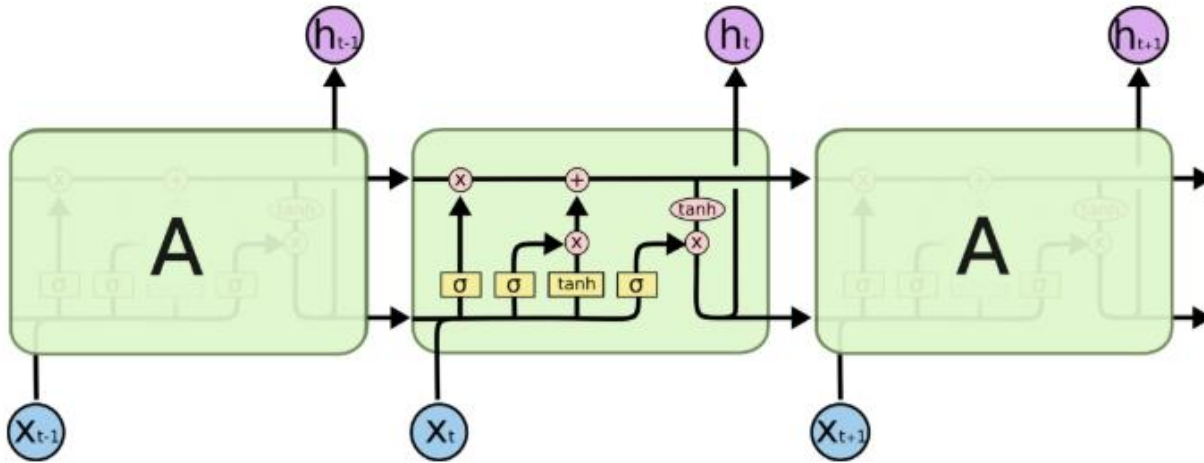
$$\mathbf{o}_t = \mathbf{g}\left(\mathbf{W}\mathbf{f}\left(\mathbf{V}\mathbf{f}\left(\mathbf{V}\dots\left(\mathbf{V}\mathbf{f}\left(\mathbf{V}\mathbf{h}_0 + \mathbf{U}\mathbf{x}_1\right)\dots + \mathbf{U}\mathbf{x}_{t-1}\right)\right) + \mathbf{U}\mathbf{x}_t\right)\right), t = 1, 2, \dots$$

Some notes

1. RNN actually make a set of nonlinear first order difference equations.
2. The current state \mathbf{h}_t is affected by both exogenous input \mathbf{x}_t and the former state \mathbf{h}_{t-1} .
3. The output \mathbf{o}_t is a function of hidden state \mathbf{h}_t at each time t .
4. Such equations can provide a memory from the short dependencies in a sequenced patterns
5. *Activation functions:* $f, g \in \{Relu, tanh, sigm, sign, identity \dots\}$

LSTM (long short term memory)

Hochreitor & Shmidhuber 1997



$$f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma (W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

$$o_t = \sigma (W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh (C_t)$$

$$\hat{y}_t = f(V \cdot h_t + b_v)$$

Some notes:

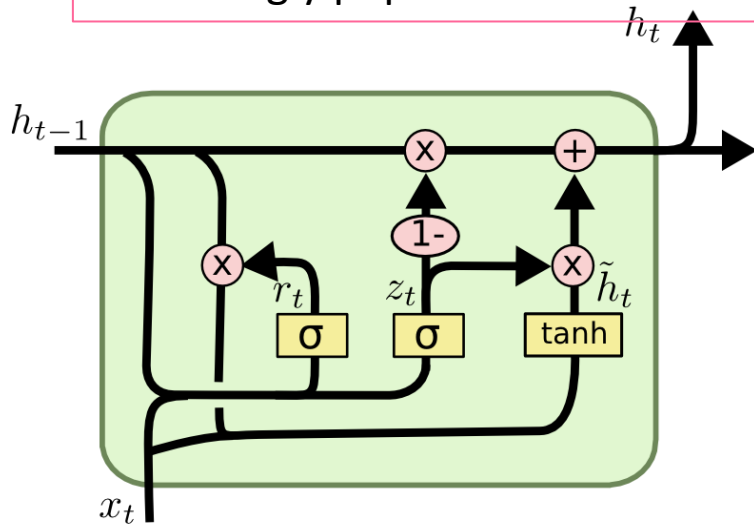
1. LSTM is a module of *four interacted layers*.
2. A cell state is a key concept in a LSTM proving (**packaging and carrying**) all informative data required to save long/short dependencies among the sequenced patterns.
3. Using exogenous input and the hidden state, cell state interacts by three independent gates: **forget**, **write**, and **read** gates.
4. Some information on the cell state is forgotten by the **forget gate**.
5. Using exogenous input and the hidden state, a package of informative data is added to the cell state by **write gate**.
6. The hidden state is provided from the cell state by **read gate**.
7. The output is a function of the hidden state.

GRU (Gated Recurrent Unit)

Cho, et al. (2014)

Many extensions of the LSTM have been introduced. A slightly more dramatic variation on the LSTM is the Gated Recurrent Unit, or GRU.

GRU combines the forget and input gates into a single “update gate.” It also merges the cell state and hidden state, and makes some other changes. The resulting model is simpler than standard LSTM models, and has been growing increasingly popular.

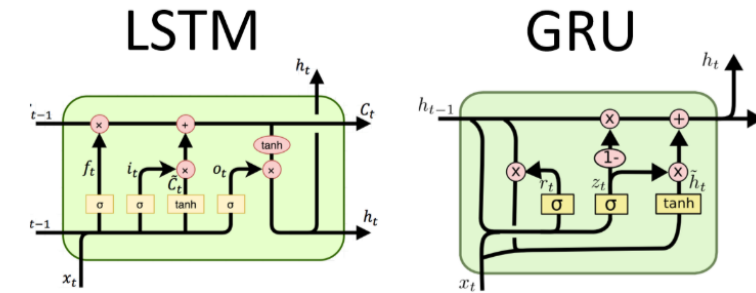


$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t])$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t])$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$



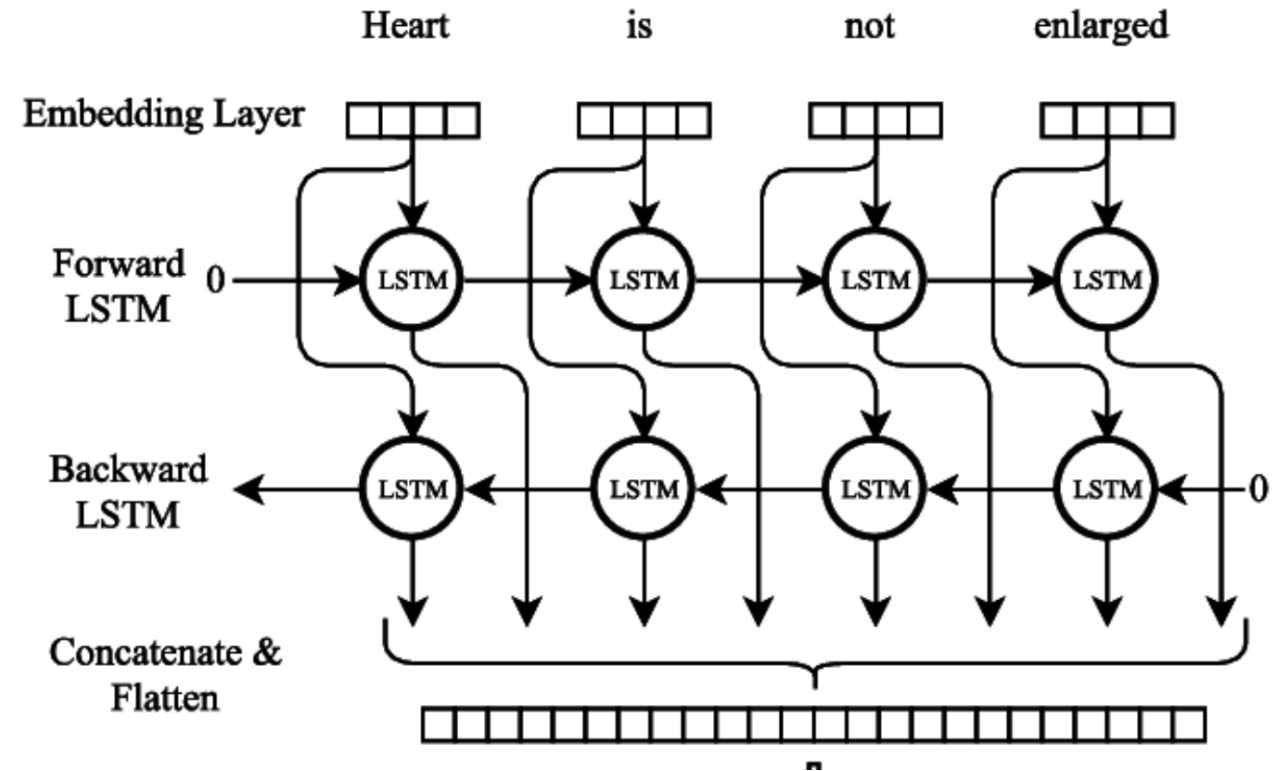
Bidirectional RNNs

A **Bidirectional LSTM**, or **biLSTM**, is a sequence processing model that consists of two LSTMs:

One taking the input in a forward direction, and the other in a backwards direction.

BiLSTMs effectively increase the amount of information available to the network, improving the context available to the algorithm (e.g. knowing what words immediately follow *and* precede a word in a sentence).

Image Source: Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks, Cornegruta et al.

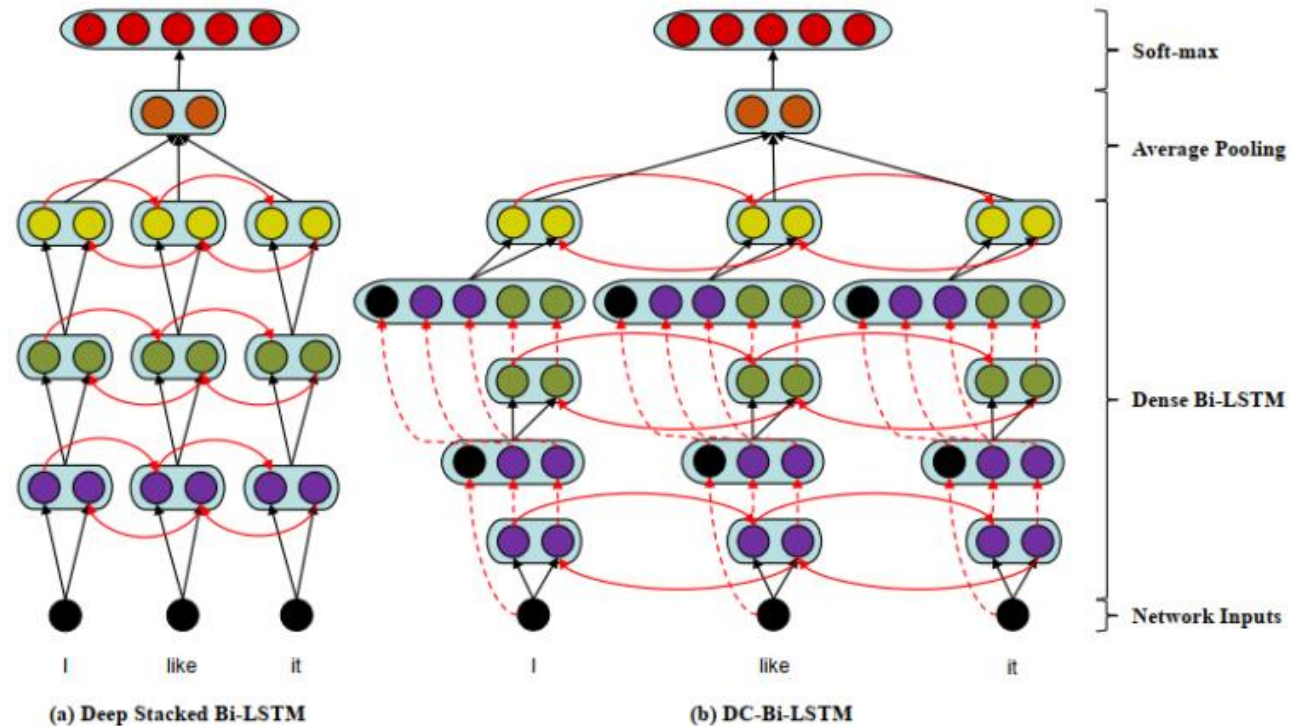


It can be used in categorization
Missed words prediction and so on.

Deep (Bidirectional) RNNs

A block of one or more than one LSTM or Bi-LSTM layers

- **Deep (Bidirectional) RNNs** are similar to Bidirectional RNNs, only that we now have multiple layers per time step. In practice this gives us a higher learning capacity (but we also need a lot of training data)



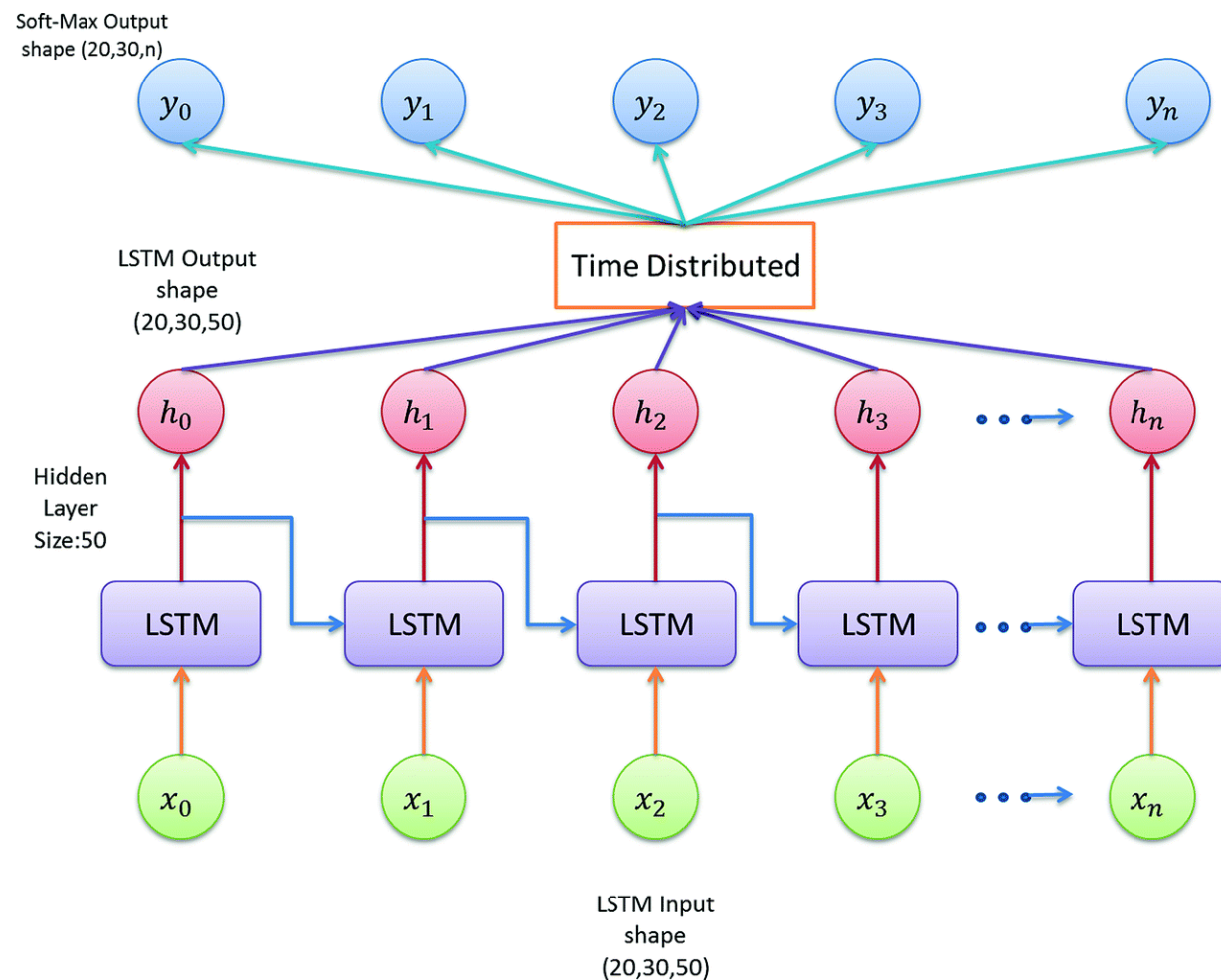
Deep Stacked Bi-LSTM

Densely Connected Bidirectional

Time distributed Layer

to reduce the required memory in processing multi sequenced samples(images) to predict the target

Time Distributed layer is very useful to work with time series data or video frames. It allows **to use a layer for each input**. That means that instead of having several input “models”, we can use “one model” applied to each input. Then GRU or LSTM can help to manage the data in “time”.



Time step: the number of past samples of a sequence which are used in a LSTM to predict the targets.

ConvLSTM

Xingjian Shi Zhourong, et al (2015)

- To provide memory with sequential images, one approach is using **ConvLSTM** layers.
- **ConvLSTM** is a Recurrent layer, just like the LSTM, but internal matrix multiplications are exchanged with convolution operations. As a result, the data that flows through the ConvLSTM cells keeps the input dimension (3D in our case) instead of being just a 1D vector with features.
- Using shared weights, convolution provide more appropriate processing and less computations in comparison to matrix product.

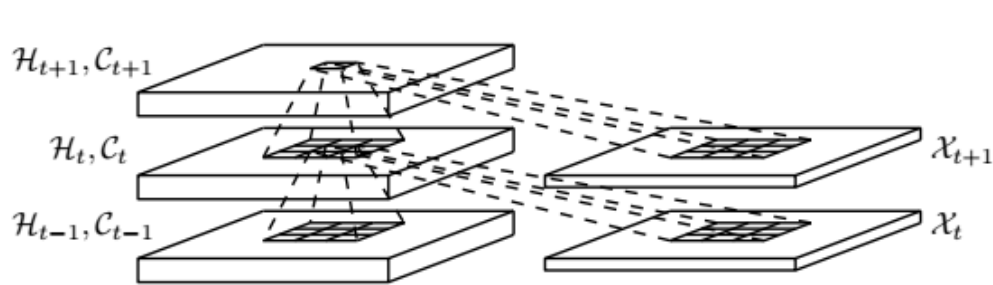
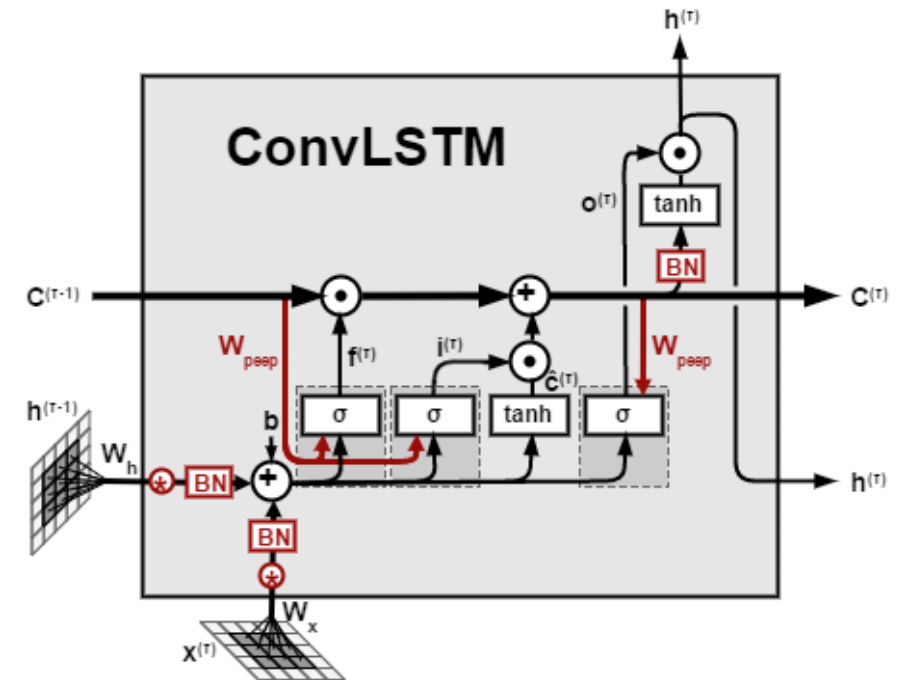


Figure 2: Inner structure of ConvLSTM

$$\begin{aligned} i_t &= \sigma(W_{xi} * \mathcal{X}_t + W_{hi} * \mathcal{H}_{t-1} + W_{ci} \circ C_{t-1} + b_i) \\ f_t &= \sigma(W_{xf} * \mathcal{X}_t + W_{hf} * \mathcal{H}_{t-1} + W_{cf} \circ C_{t-1} + b_f) \\ C_t &= f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * \mathcal{X}_t + W_{hc} * \mathcal{H}_{t-1} + b_c) \\ o_t &= \sigma(W_{xo} * \mathcal{X}_t + W_{ho} * \mathcal{H}_{t-1} + W_{co} \circ C_t + b_o) \\ \mathcal{H}_t &= o_t \circ \tanh(C_t) \end{aligned}$$

where $*$ denotes the convolution operator and \circ , as before, denotes the Hadamard product (element-wise product)



1.1.4 Attention Layers-Modules

An improving mechanism for RNNs by applying an interface connecting the encoder and decoder

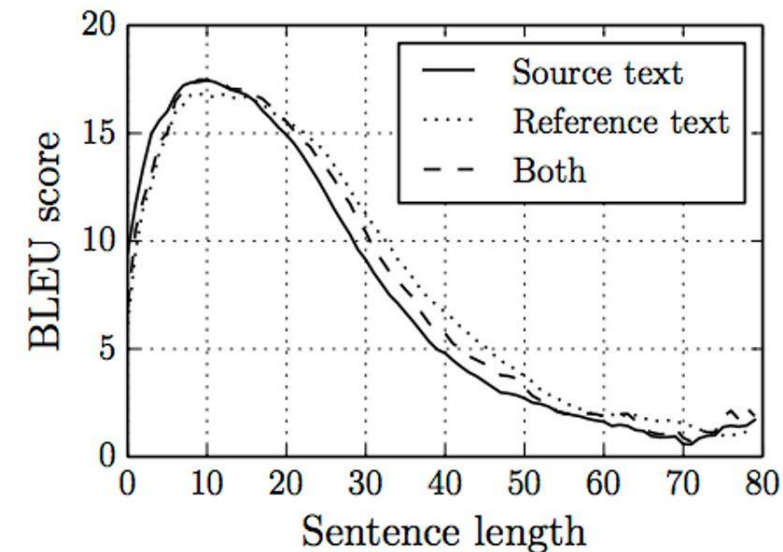
- In translation machines, LSTM/GRU as a seq2seq model can not predict successful translation for a sentence, when its length increases.

What is attention layer in Deep Learning?

- Every RNN/LSTM can be interpreted as an Encoder and Decoder.
 - Encoder encodes the input samples of a chronological signal to hidden state.
 - Decoder decodes the hidden state to the output.

Attention is an interface connecting the encoder and decoder that provides the decoder with information from every encoder hidden state.

With this framework, the model is able to selectively focus on valuable parts of the input sequence and hence, learn the association between them.



BLEU:
Bilingual Evaluation Understudy

*Neural Machine Translation by Jointly Learning to Align and Translate

[Dzmitry Bahdanau](#), [Kyunghyun Cho](#), [Yoshua Bengio](#)

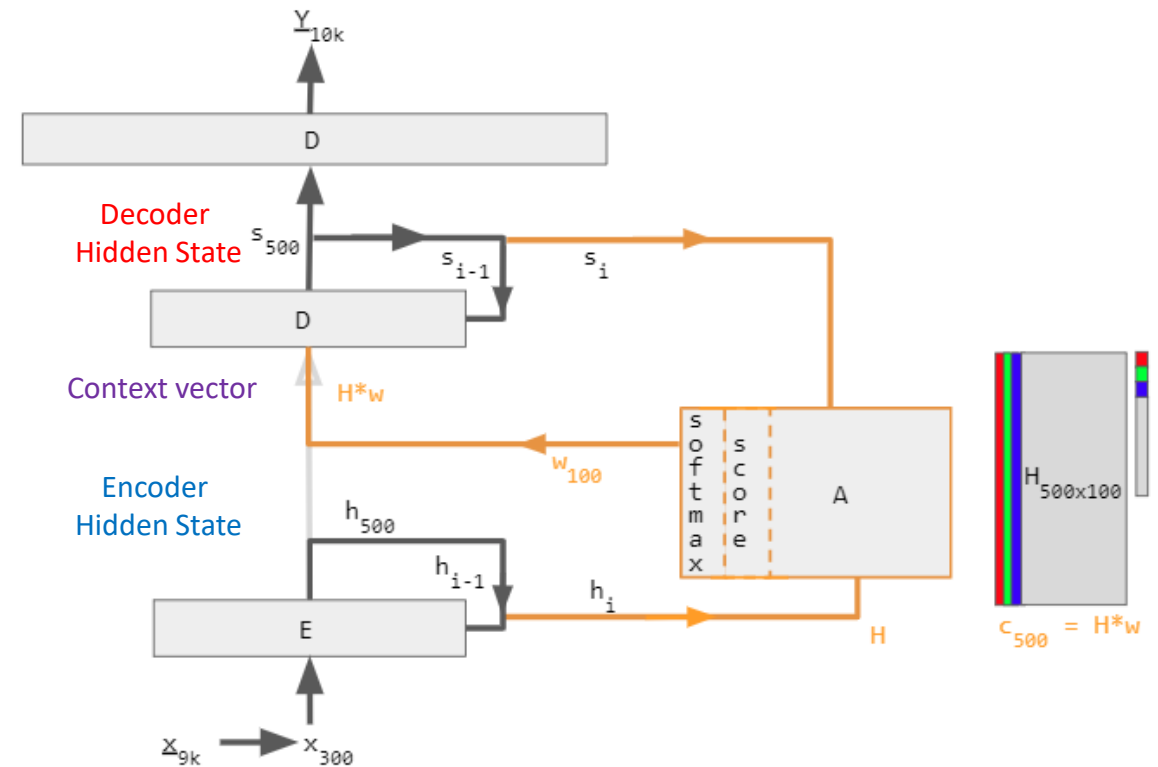
Ahmad Kalhor-University of Tehran

A language translation example

To build a machine that translates **English-to-French** (see the shown diagram), one starts with **an Encoder-Decoder** and **grafts an attention unit to it**. In the simplest case such as the shown example, **the attention unit is just lots of dot products of recurrent layer states and does not need training**. In practice, the attention unit consists of 3 fully connected neural network layers that needs to be trained. The 3 layers are called Query, Key, and Value.

Encoder-Decoder with attention. This diagram uses specific values to relieve an already cluttered notation alphabet soup. **The left part (in black) is the Encoder-Decoder, the middle part (in orange) is the attention unit, and the right part (in grey & colors) is the computed data.** Grey regions in H matrix and w vector are zero values.

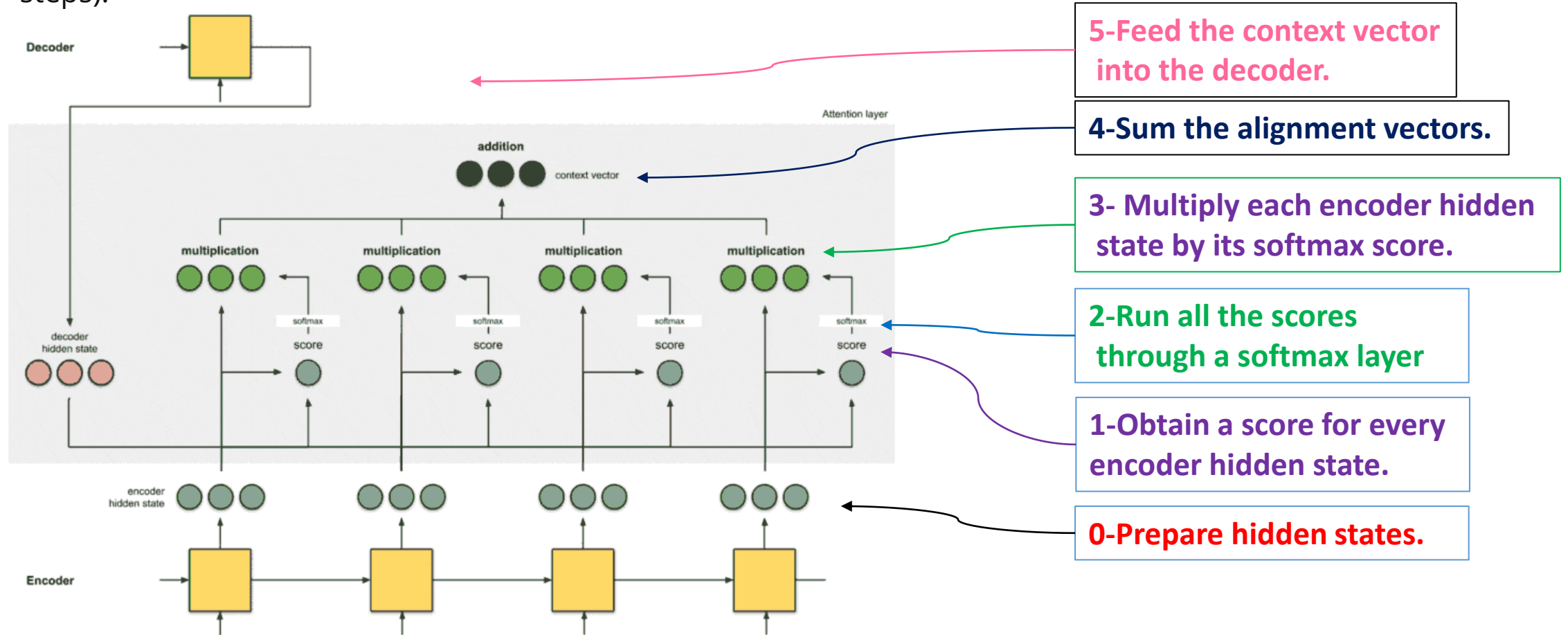
Numerical subscripts are examples of vector sizes. Lettered subscripts i and $i-1$ indicate time step.



label	description
100	max sentence length
300	embedding size (word dimension)
500	length of hidden vector
9k, 10k	dictionary size of input & output languages respectively.
\underline{x} , \underline{y}	9k and 10k 1-hot dictionary vectors. $\underline{x} \rightarrow x$ implemented as a lookup table rather than vector multiplication. \underline{y} is the 1-hot maximizer of the linear Decoder layer D; that is, it takes the argmax of D's linear layer output.
x	300-long word embedding vector. The vectors are usually pre-calculated from other projects such as GloVe or Word2Vec .
h	500-long encoder hidden vector. At each point in time, this vector summarizes all the preceding words before it. The final h can be viewed as a "sentence" vector, or a thought vector as Hinton calls it.
s	500-long decoder hidden state vector.
E	500 neuron RNN encoder. 500 outputs. Input count is 800–300 from source embedding + 500 from recurrent connections. The encoder feeds directly into the decoder only to initialize it, but not thereafter; hence, that direct connection is shown very faintly.
D	2-layer decoder. The recurrent layer has 500 neurons and the fully connected linear layer has 10k neurons (the size of the target vocabulary). ^[7] The linear layer alone has 5 million (500 * 10k) weights -- ~10 times more weights than the recurrent layer.
score	100-long alignment score
w	100-long vector attention weight. These are "soft" weights which changes during the forward pass, in contrast to "hard" neuronal weights that change during the learning phase.
A	Attention module — this can be a dot product of recurrent states, or the Query-Key-Value fully connected layers. The output is a 100-long vector w.
H	500x100. 100 hidden vectors h concatenated into a matrix
c	500-long context vector = $H * w$. c is a linear combination of h vectors weighted by w.

Unfolded attention layer

The below figure demonstrates an Encoder-Decoder architecture with an attention layer (All time steps).



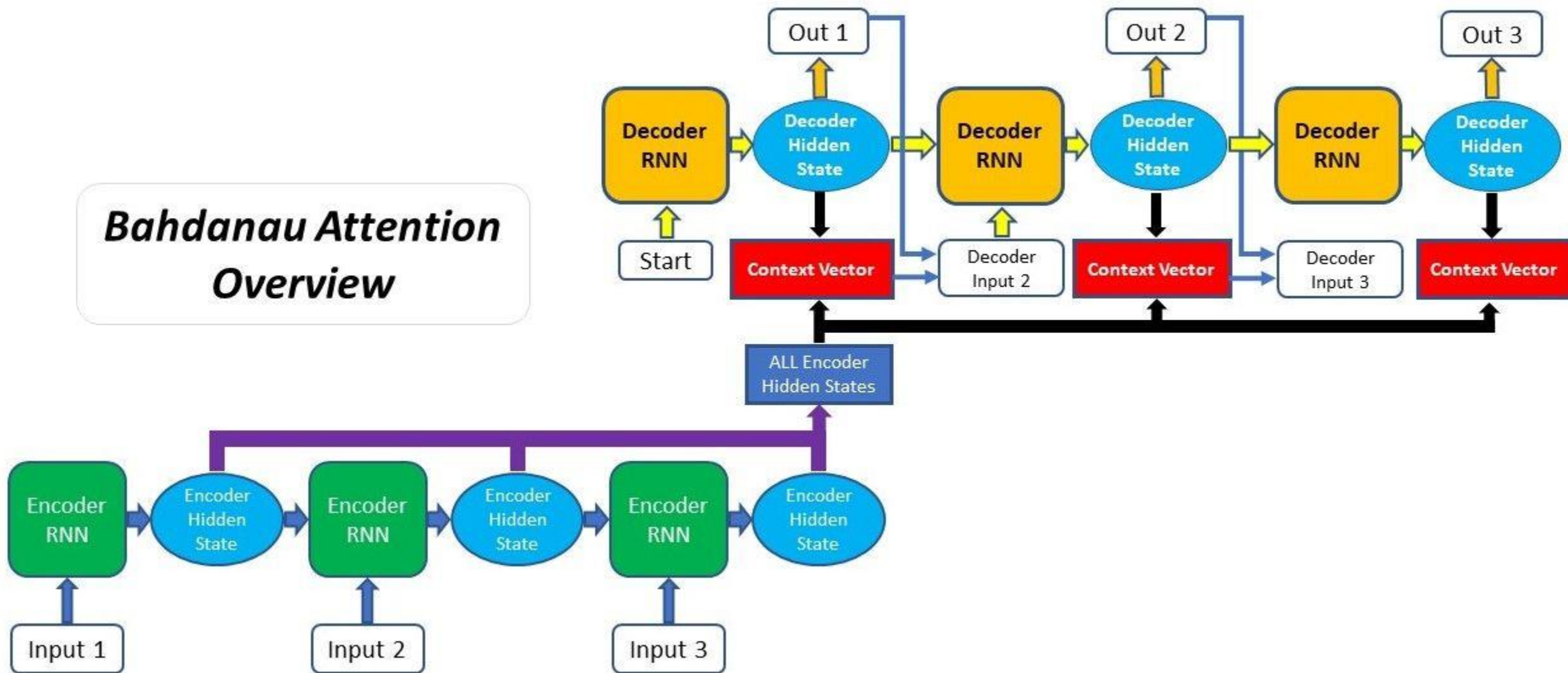
*Encoder-Decoder Recurrent Neural Network Models for Neural Machine Translation
by [Jason Brownlee](#) on January 1, 2018 in [Deep Learning for Natural Language Processing](#)

The implementations of an attention layer can be broken down into 4 steps.

For each time step:

- **Step 0: Prepare hidden states.**
 - First, prepare all the available encoder hidden states (green) and the first decoder hidden state (red). In our example, we have 4 encoder hidden states and the current decoder hidden state. (Note: the last consolidated encoder hidden state is fed as input to the first time step of the decoder. The output of this first time step of the decoder is called the first decoder hidden state.)
- **Step 1: Obtain a score for every encoder hidden state.**
 - A score (scalar) is obtained by a score function (also known as alignment score function or alignment model). In this example, the score function is a dot product between the decoder and encoder hidden states.
- **Step 2: Run all the scores through a softmax layer.**
 - We put the scores to a softmax layer so that the softmax scores (scalar) add up to 1. These softmax scores represent the attention distribution.
- **Step 3: Multiply each encoder hidden state by its softmax score.**
 - By multiplying each encoder hidden state with its softmax score (scalar), we obtain the alignment vector or the annotation vector. This is exactly the mechanism where alignment takes place.
- **Step 4: Sum the alignment vectors.**
 - The alignment vectors are summed up to produce the context vector. A context vector is an aggregated information of the alignment vectors from the previous step.
- **Step 5: Feed the context vector into the decoder.**

Bahdanau Attention Overview



*Neural Machine Translation by Jointly Learning to Align and Translate

[Dzmitry Bahdanau](#), [Kyunghyun Cho](#), [Yoshua Bengio](#)

Attention variants

1. encoder-decoder dot product
2. encoder-decoder QKV (Query-Key-Value)
3. encoder-only dot product
4. encoder-only QKV
5. Pytorch tutorial

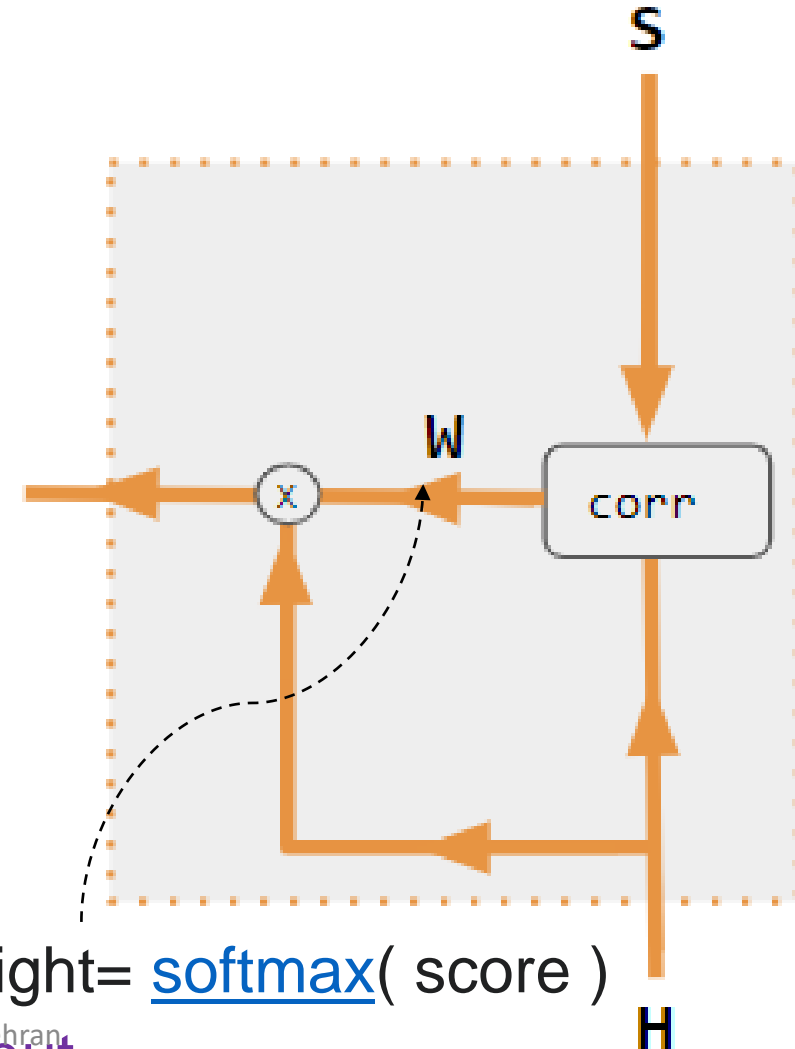
1. encoder-decoder dot product

s = decoder input to Attention
= Decoder hidden state

Both Encoder & Decoder are needed to calculate Attention.

c = context vector = $H * w$

$W * H$



1. Luong, Minh-Thang (2015-09-20). "Effective Approaches to Attention-based Neural Machine Translation". [arXiv:1508.04025v5](https://arxiv.org/abs/1508.04025v5) [cs.CL].

w = attention weight = softmax(score)

H = encoder output

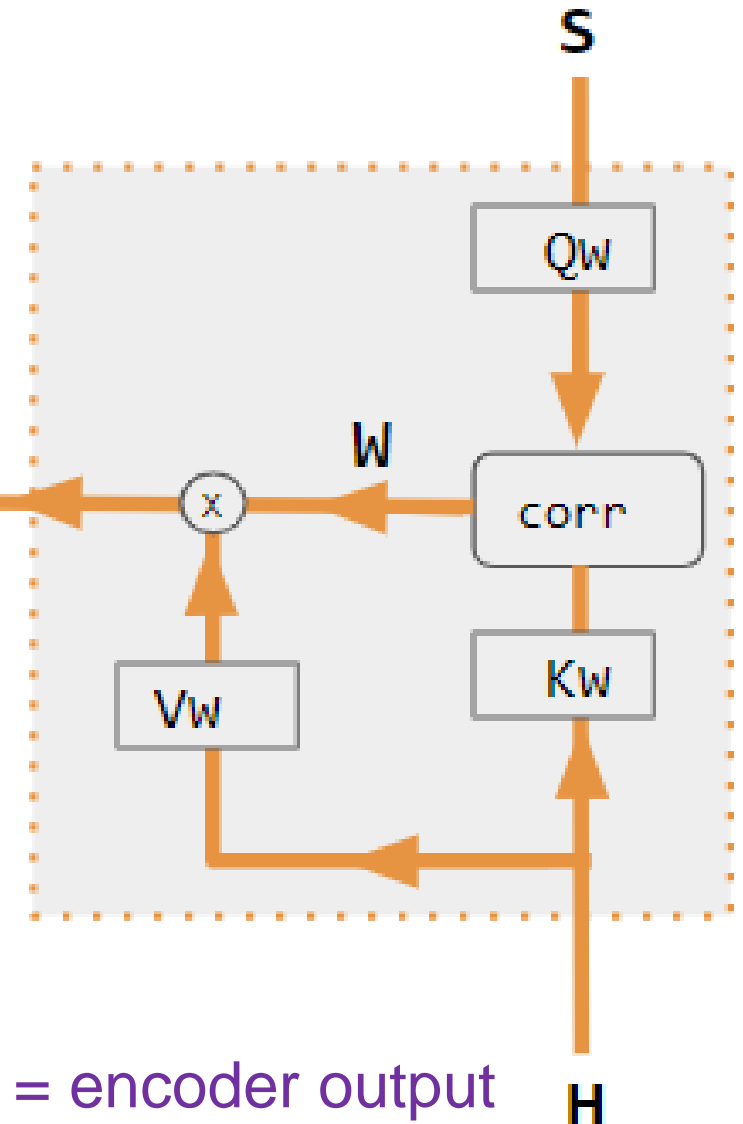
2. encoder-decoder QKV

s = decoder input to Attention

Both Encoder & Decoder are needed to calculate Attention.

c = context vector

$W * V_w * H$

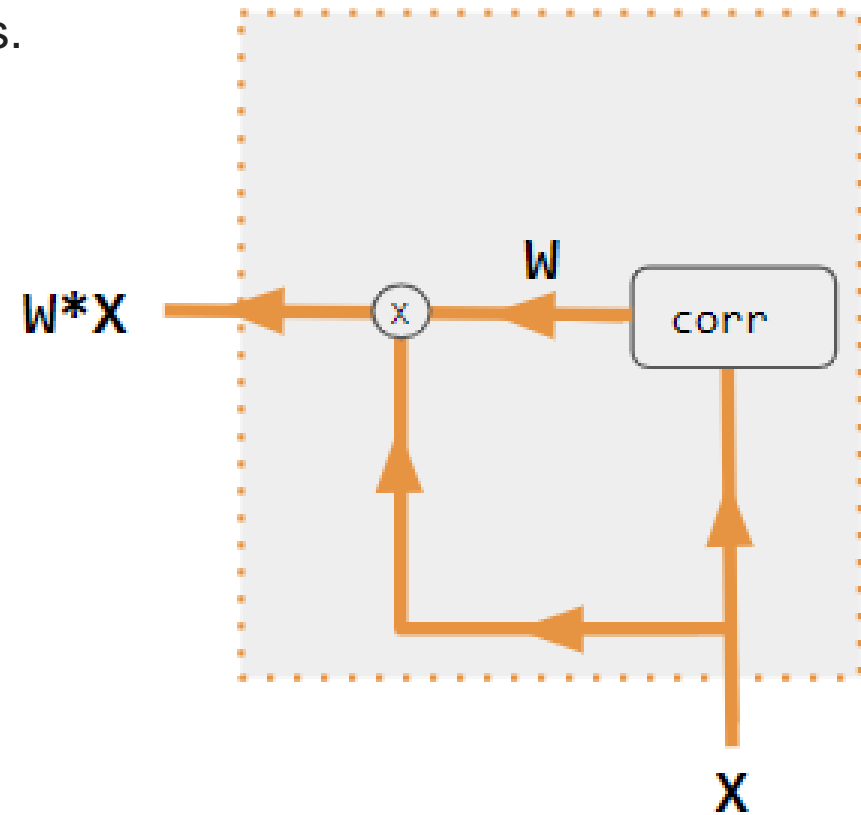


1. Neil Rhodes (2021). [CS 152 NN—27: Attention: Keys, Queries, & Values](#). Event occurs at 06:30. Retrieved 2021-12-22.

3. encoder-only dot product

Decoder is NOT used to calculate Attention. With only 1 input into corr, W is an auto-correlation of dot products.

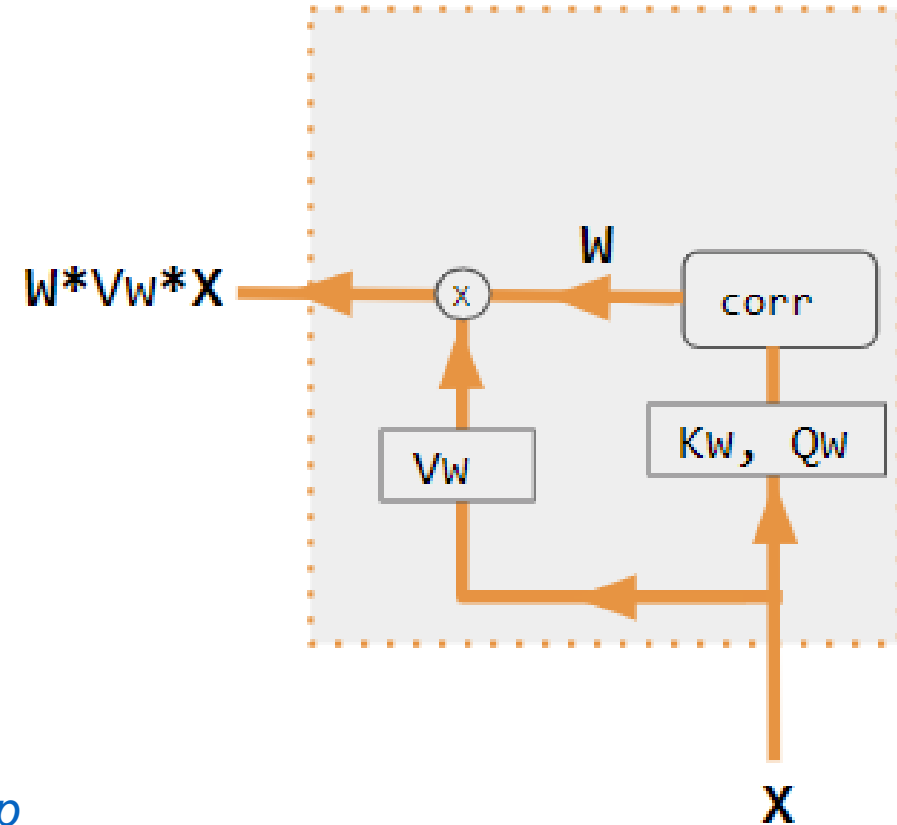
$$w_{ij} = x_i * x_j \text{ [10]}$$



1. Alfredo Canziani & Yann Lecun (2021). [NYU Deep Learning course, Spring 2020](#). Event occurs at 05:30. Retrieved 2021-12-22.

4. encoder-only QKV

Decoder is NOT used to calculate Attention.^[11]



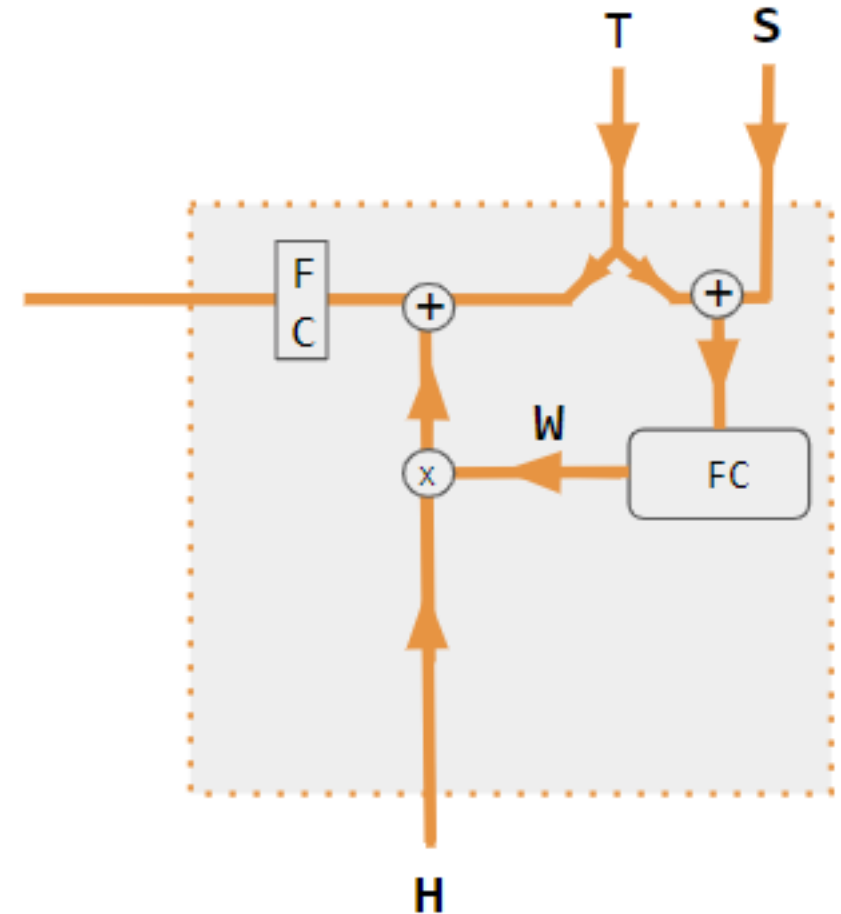
1. Alfredo Canziani & Yann Lecun (2021). [NYU Deep Learning course, Spring 2020](#). Event occurs at 20:15. Retrieved 2021-12-22.

5. Pytorch tutorial

A FC layer is used to calculate Attention instead of dot product correlation.

S = decoder hidden state, T = target word embedding.

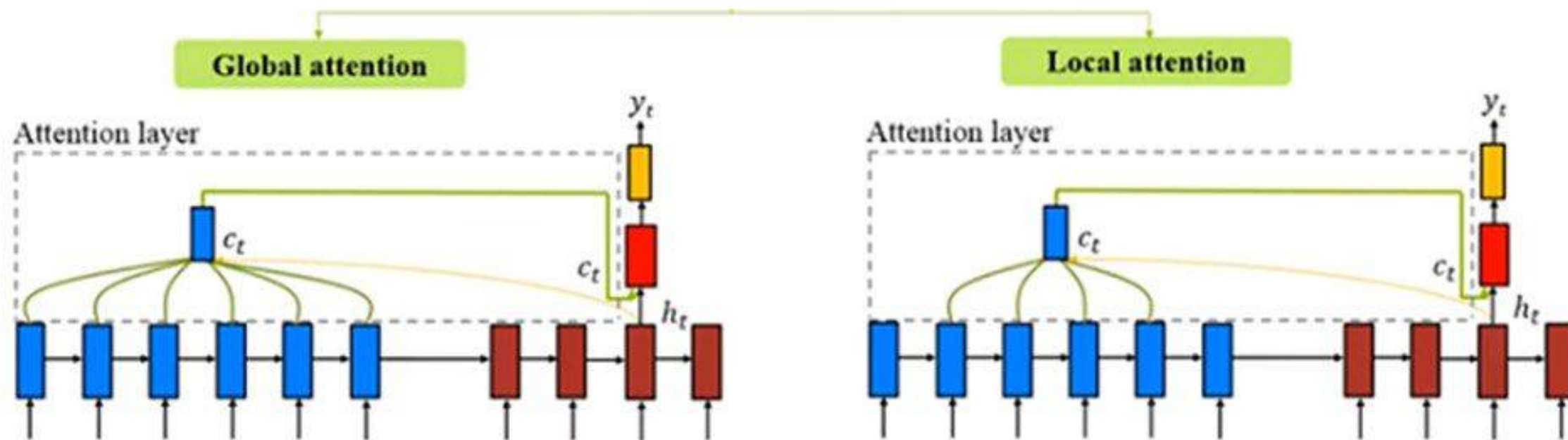
H = encoder hidden state, X = input word embedding



1. Robertson, Sean. ["NLP From Scratch: Translation With a Sequence To Sequence Network and Attention"](#). *pytorch.org*. Retrieved 2021-12-22.

Types of attention

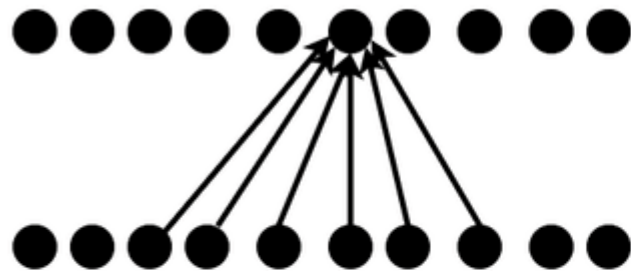
Depending on how many source states contribute while deriving the attention vector (α), there can be three types of attention mechanisms:



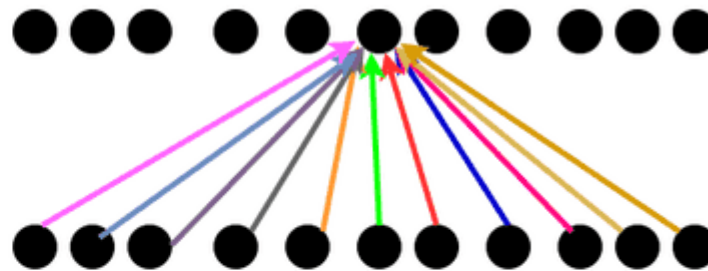
- **Global Attention:** When attention is placed on all source states. In global attention, we require as many weights as the source sentence length.
- **Local Attention:** When attention is placed on a few source states.
- **Hard Attention:** When attention is placed on only one source state.

You can check out my [Kaggle notebook](#) or [GitHub repo](#) to implement NMT with the attention mechanism using TensorFlow.

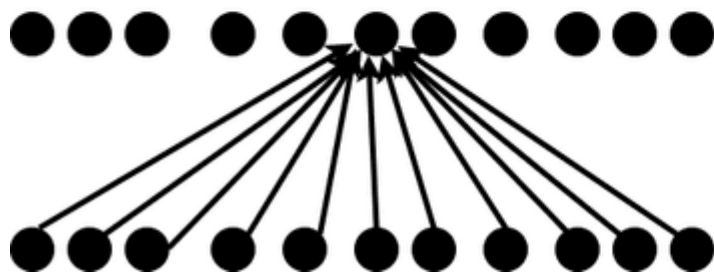
Convolution



Global attention



Fully Connected layer



Local attention

