Presentation

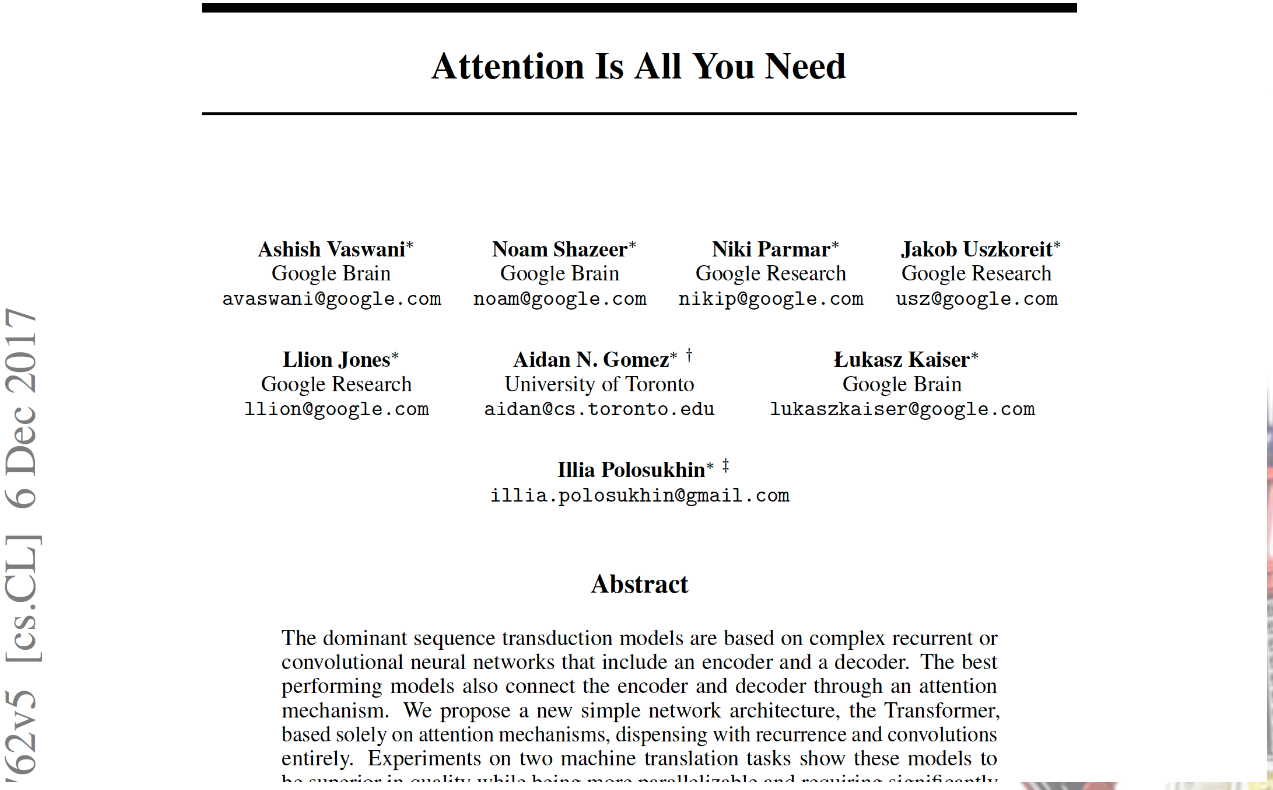
Mingyu (Jerry) Liu 21st Aug 2020

https://github.com/JerryLiuMY/self_attention_rnn

Outline

- Present high-level overview of the Transformer model
- Illustrate from end-to-end my data pipeline
- Clarify *TimeDistributed wrapper* mentioned last time

1. Introduction to Transformer



62v5 [cs.CL] 6 Dec 2017

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

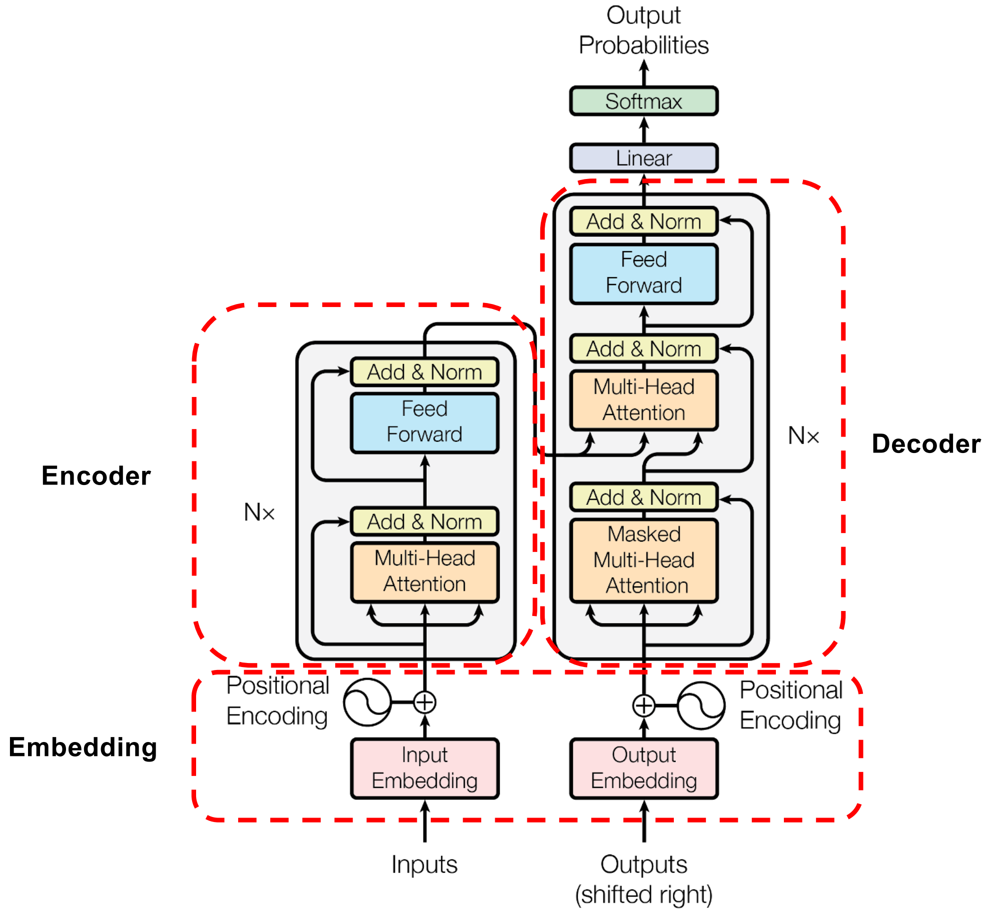
Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly

Transformer Architecture



Advantages

- Infinite memory with attention (next slide)
- Computationally efficient by avoiding recursion
- Computationally efficient with high hidden dim

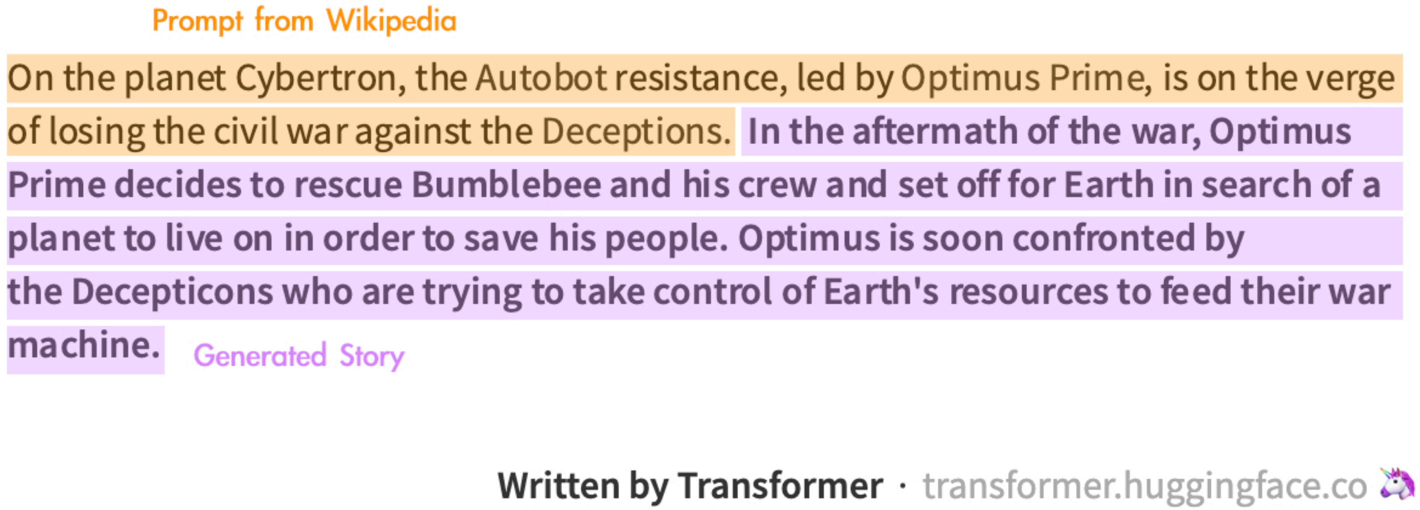
$$\text{LSMT: } \mathcal{O}(\text{seq\_len} \times \text{hidden\_dim}^2)$$

$$\text{Transformer: } \mathcal{O}(\text{seq\_len}^2 \times \text{hidden\_dim})$$

Disadvantage

- Can only process fixed length sequences
- Can be expensive to train with long sentences

Example: Text Generation



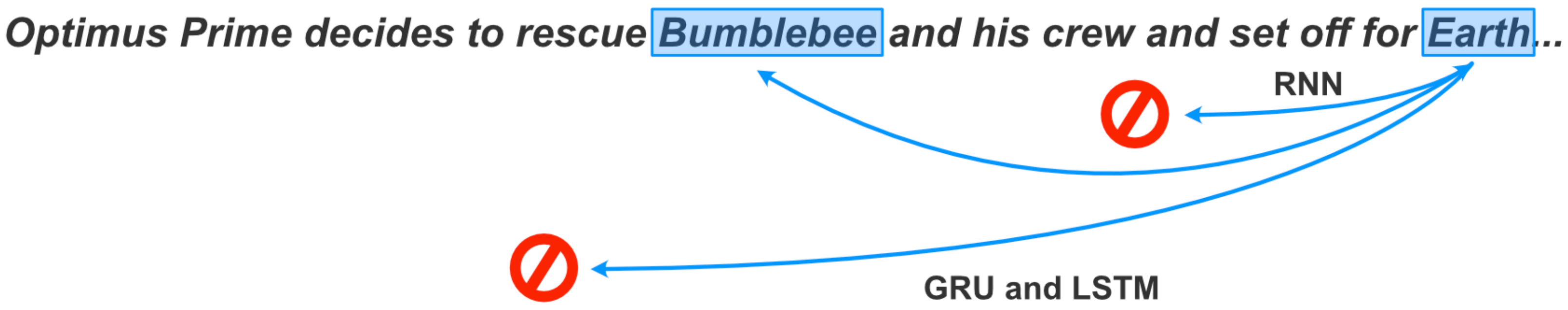
Prompt from Wikipedia

On the planet Cybertron, the Autobot resistance, led by Optimus Prime, is on the verge of losing the civil war against the Deceptions. **In the aftermath of the war, Optimus Prime decides to rescue Bumblebee and his crew and set off for Earth in search of a planet to live on in order to save his people. Optimus is soon confronted by the Decepticons who are trying to take control of Earth's resources to feed their war machine.**

Generated Story

Written by Transformer · transformer.huggingface.co 🦄

RNN, GRU and LSTM

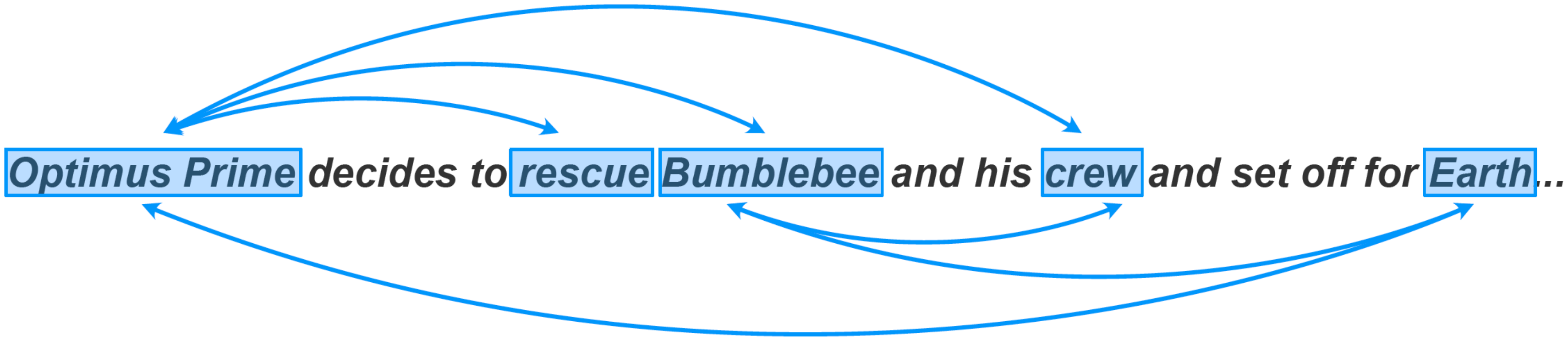


Optimus Prime decides to rescue Bumblebee and his crew and set off for Earth...

RNN

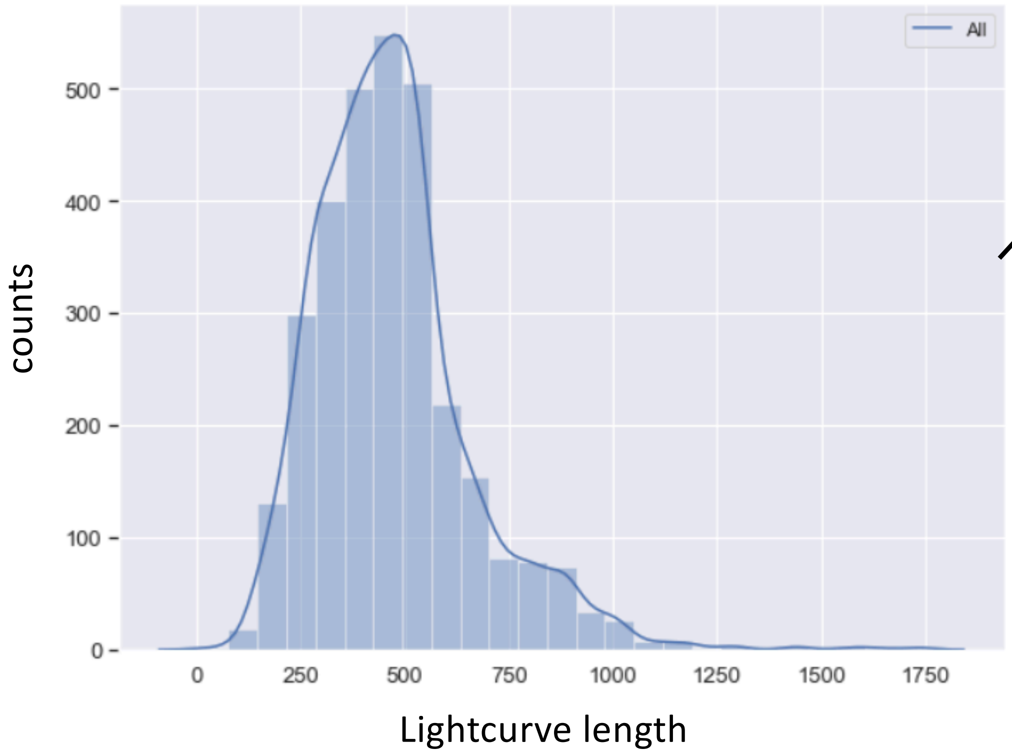
GRU and LSTM

Attention Mechanism



Optimus Prime decides to rescue Bumblebee and his crew and set off for Earth...

2. Data Pipeline Revisited



Exceptionally unequal len

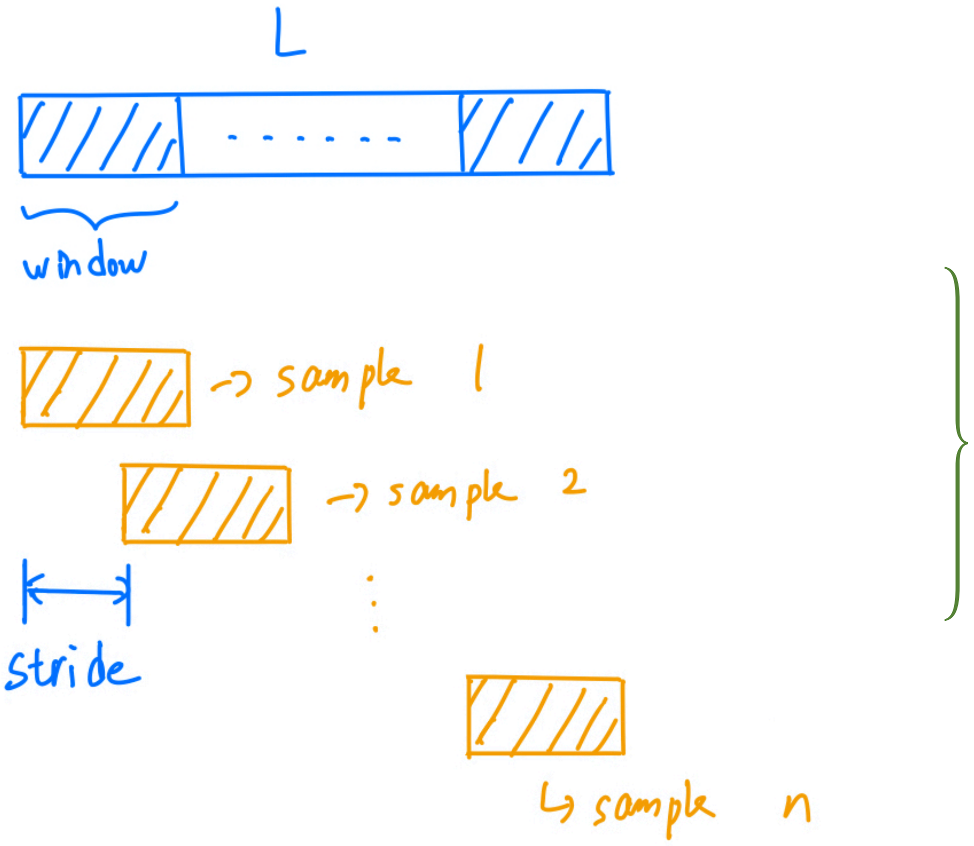
Common Solutions:

a) Pad short sequences

b) Truncate long sequences

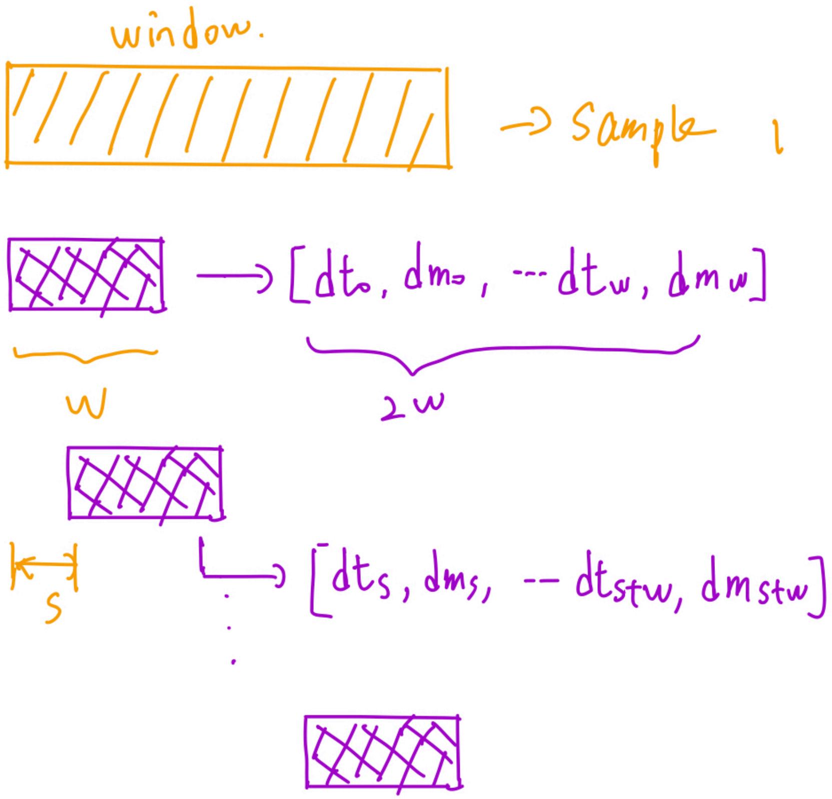
c) Single sample per iteration

2.1 Sub-sampling



$$\text{num\_samples} = \left\lfloor \frac{L - \text{window} + \text{stride}}{\text{stride}} \right\rfloor$$

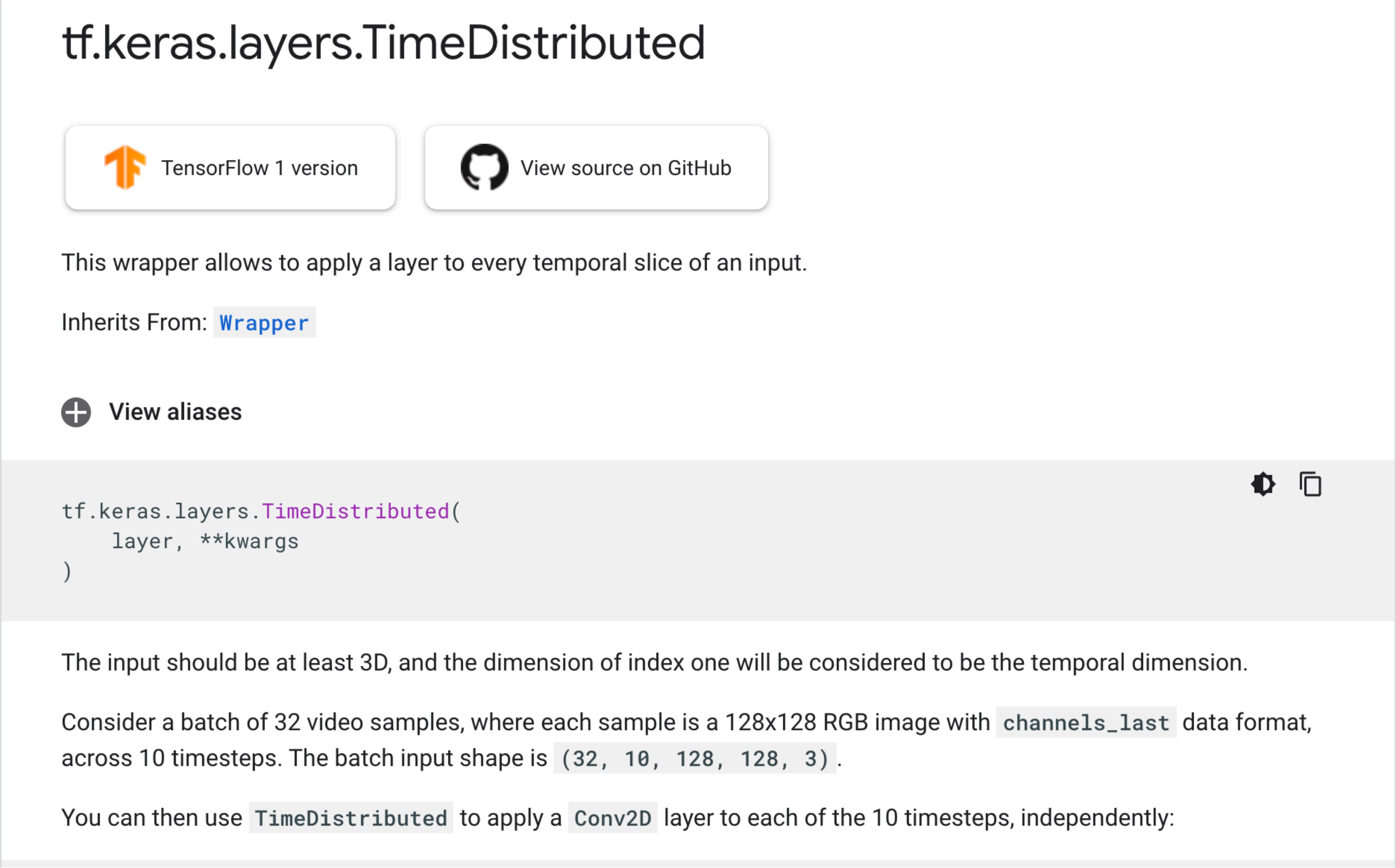
2.2 Feature Composition


window.
→ Sample 1
→ $[dt_0, dm_0, \cdots dt_w, dm_w]$
w
2w
s
→ $[dt_s, dm_s, \cdots dt_{s+w}, dm_{s+w}]$

$$x_i = \underbrace{\begin{bmatrix} dt_0 & dm_0 & \dots & dt_w & dm_w \\ dt_s & dm_s & \dots & dt_{w+s} & dm_{w+s} \\ & & \vdots & & \end{bmatrix}}_{2w} \left.\vphantom{\begin{bmatrix} a \\ b \\ c \end{bmatrix}}\right\} \text{num\_steps}$$

$$\text{num\_steps} = \left\lfloor \frac{\text{window} - \text{w} + \text{s}}{\text{s}} \right\rfloor$$

3. TimeDistributed



tf.keras.layers.TimeDistributed

This wrapper allows to apply a layer to every temporal slice of an input.

Inherits From: `Wrapper`

```
tf.keras.layers.TimeDistributed(
    layer, **kwargs
)
```

The input should be at least 3D, and the dimension of index one will be considered to be the temporal dimension.

Consider a batch of 32 video samples, where each sample is a 128x128 RGB image with `channels_last` data format, across 10 timesteps. The batch input shape is `(32, 10, 128, 128, 3)`.

You can then use `TimeDistributed` to apply a `Conv2D` layer to each of the 10 timesteps, independently:

Next Week

- Layer normalization and residual connection
- Details of each part of the Transformer model
- Experimental results