Attention Revisited

Laurel Koenig

Middle Tennessee State University

8/15/22

Fake

[1] [2] [3] [4] [5] [6]

Introduction

- **Generalization** is the ability to use past experiences to solve a problem
- People are good at this
- Deep learning can do some types but not all

Interpolation is a form of generalization that focuses on determining unknowns within a set

- Deep learning is good at this
- Spline functions

Extrapolation is a form of generalization that focuses on extending a set into the unknown

- In the past it was suggested that extrapolation wasn't possible through machine learning
- The problem is completely new

Introduction

- Working Memory (WM) is a biologically inspired solution
- WM models are extremely effective
- WM models require extensive training as well as increased memory and CPU requirements
- Recent studies have suggested that transformer architecture is capable of extrapolation [3]
- Other studies suggest that transformers break down when given more complex tasks [2]

Background: Artificial Neural Networks

- Artificial Neural Networks (ANNs) are inspired by the physical properties of the brain
- Composed of layers of nodes (artificial neurons)
- Interconnected with activation wieghts
- trained on examples of the task it should be learning to do

Background: Long Short-Term Memory

- Long Short-Term Memory (LSTM) is a particular type of Recurrent Neural Network (RNN) that was developed in response to the exploding and vanishing gradient problem
- RNNs maintain information past a single time step. That information impacts current time step.
- Vanishing gradient problem happens when the network weights stop changing before it's actually trained
- Exploding gradient problem happens when the weights are over updated and can't find an equilibrium
- LSTM networks use input, output, and forget gates to dimenish this issue

Background: Transformers

- Transformers are an encoder/decoder architecture reliant on attention mechanisms that were proposed in the 2017 paper "Attention Is All You Need".[3]
- They leverage self-attention mechanisms
- They process sequences by looking everything at once rather than one thing at a time
- Faster training and more parallelizable
- **Attention** can be thought about as mapping key-value pairs and a query to an output
- The keys, values, and query are all vectors
- A weighted sum of the values is used to compute the output
- To assign a weight to a value a compatibility function of the related key and query is used

Background: Encoder/Decoder Architecture

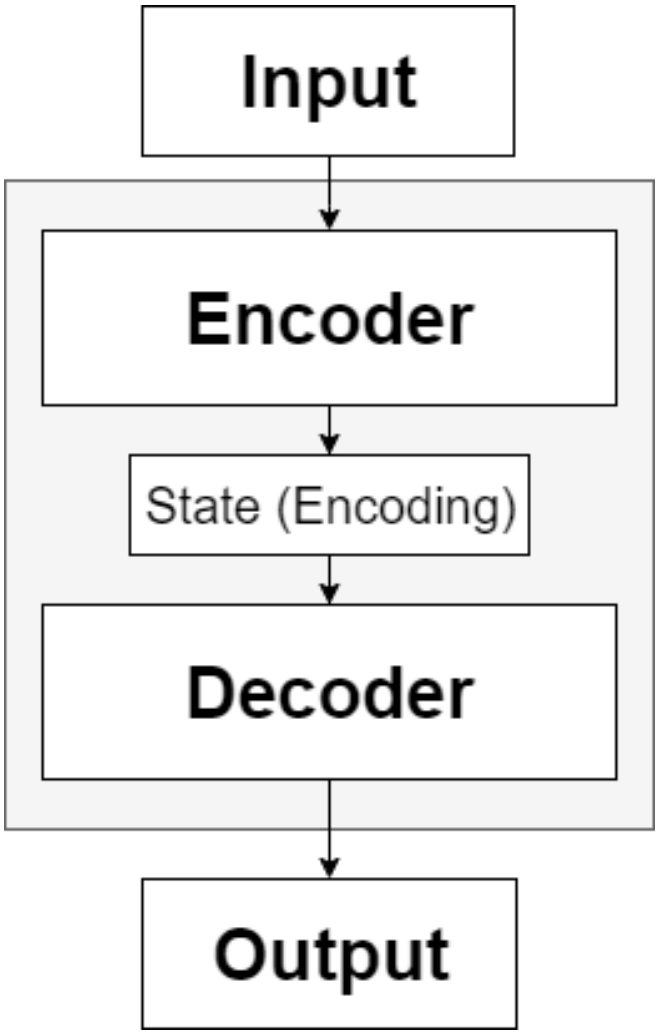

Figure: Encoder/Decoder Architecture

- The Encoder translated the input into something the decoder can understand
- The Decoder translated that encoding back into something useful to us
- They are trained and used together (not plug and play)

Background: Previous Results and Paper

- "Attention Is Not Enough" tested transformers in conjunction with LSTM architecture
- A WM model was used as the control to test the models against
- There are four different configurations and four different tests
- The repository was incomplete

Re-Implementation

- The duplicated portions of the code were moved to their own files
- The scripts that existed were rewritten as notebooks
- The missing model was then reconstructed
- Once all the models were working scripts were created from the notebooks to more easily run different tests

Re-Implementation: Model Details

- All of the models are built using a nested encoder/decoder structure
- ★ Note: For readability this will be written as Inner/Outer
- The outer models are pretrained to 100% accuracy
- All of the configurations are created with both coupled and uncoupled versions of the inner model
- All of the models use the same training and testing data
- All the models were made using Tendorflow/Keras (version 2.7.0)

Re-Implementation:Outer Models

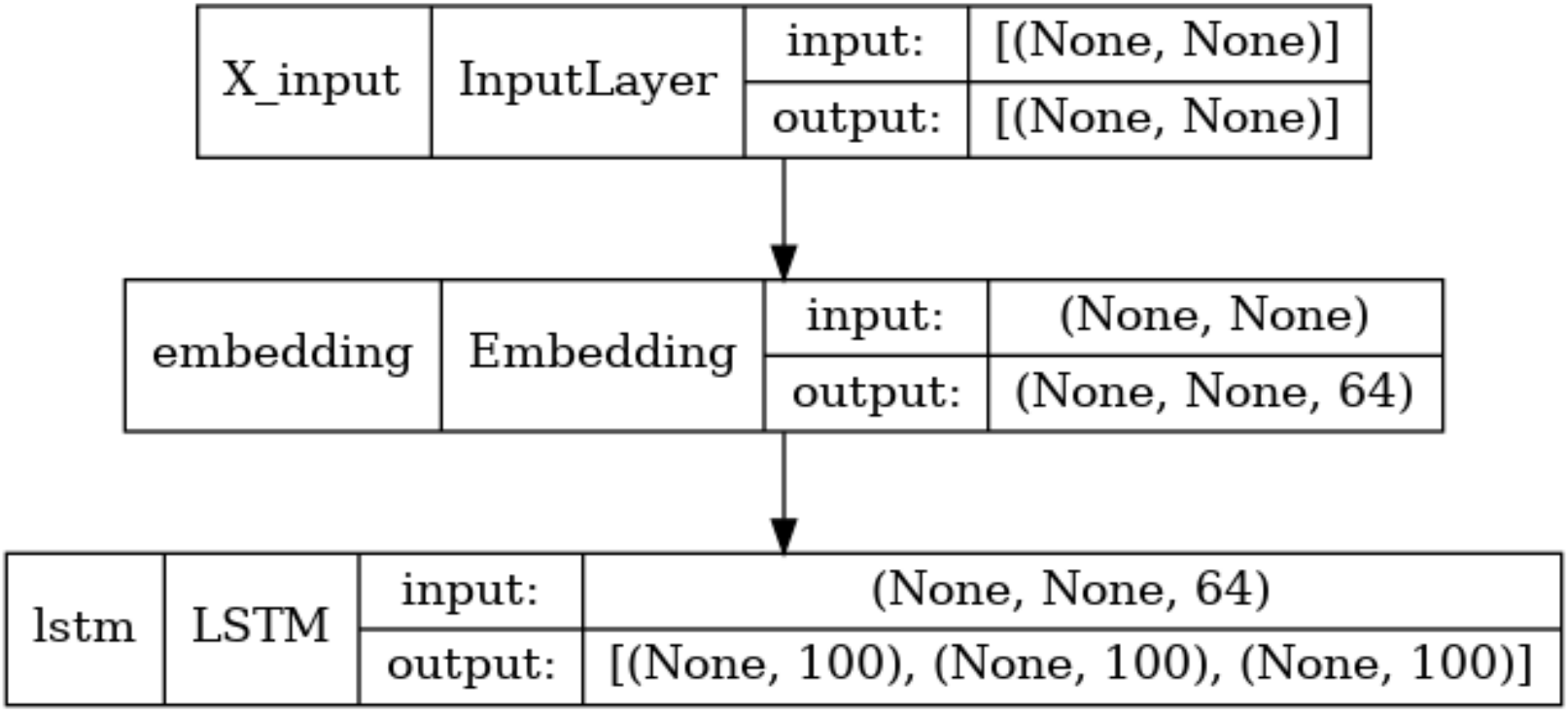

Figure: Outer LSTM Encoder

Re-Implementation:Outer Models

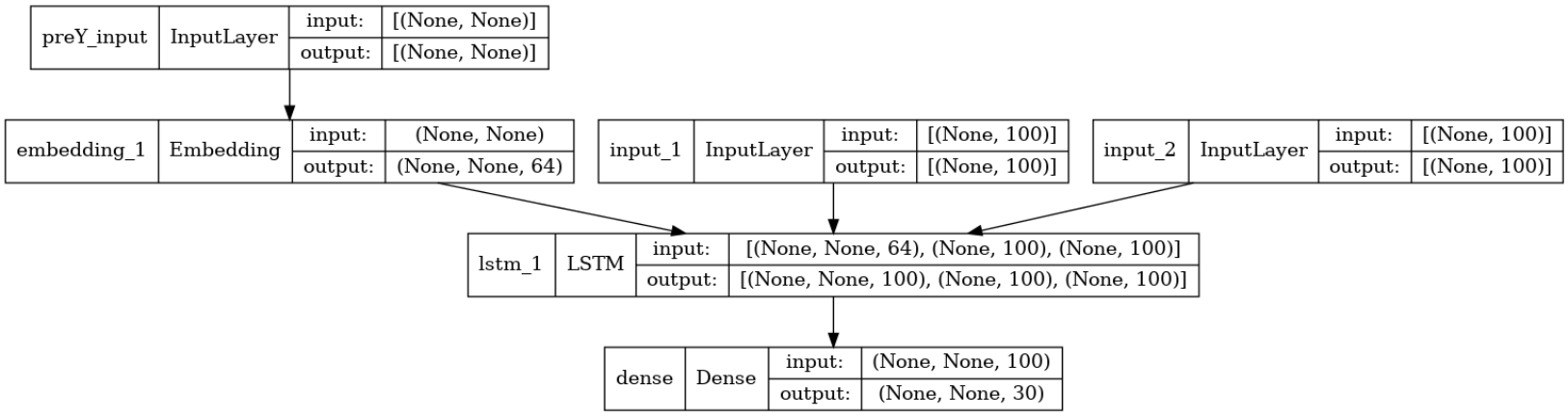

Figure: Outer LSTM Decoder

Re-Implementation:Outer Models

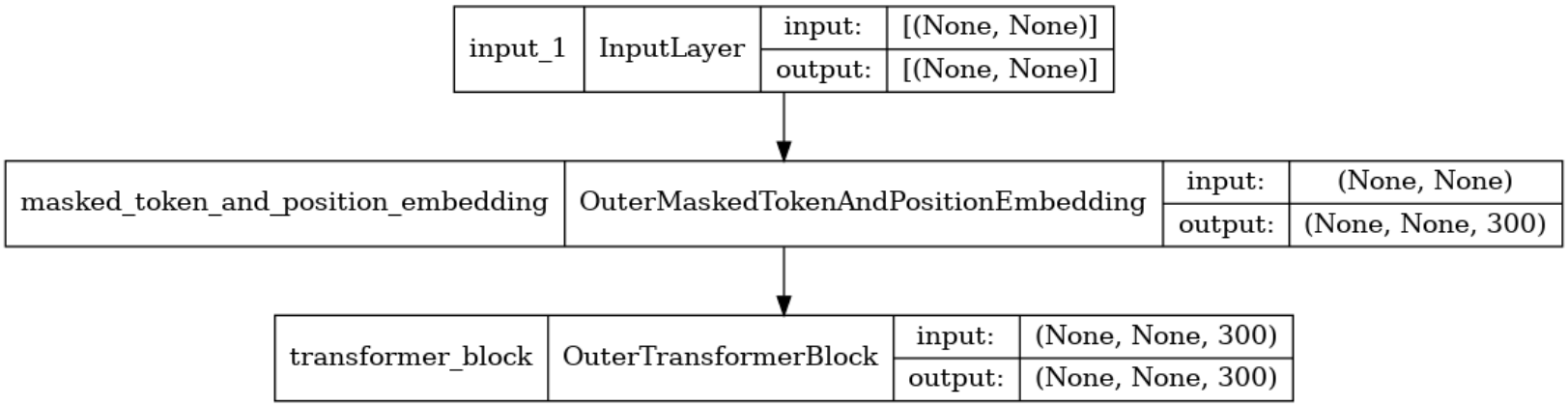

Figure: Outer Transformer Encoder

Re-Implementation:Outer Models

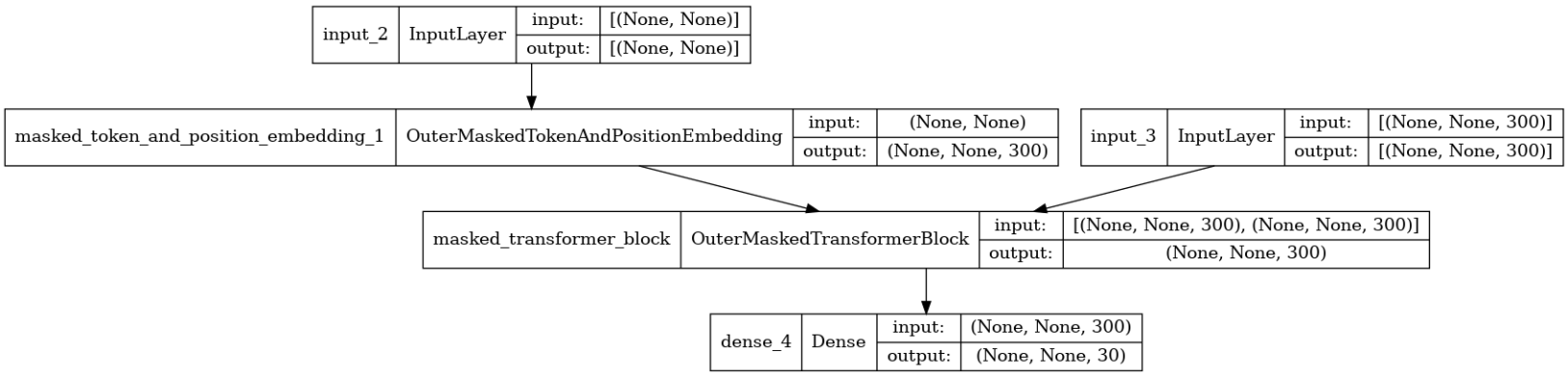

Figure: Outer Transformer Decoder

Re-Implementation: LSTM/LSTM

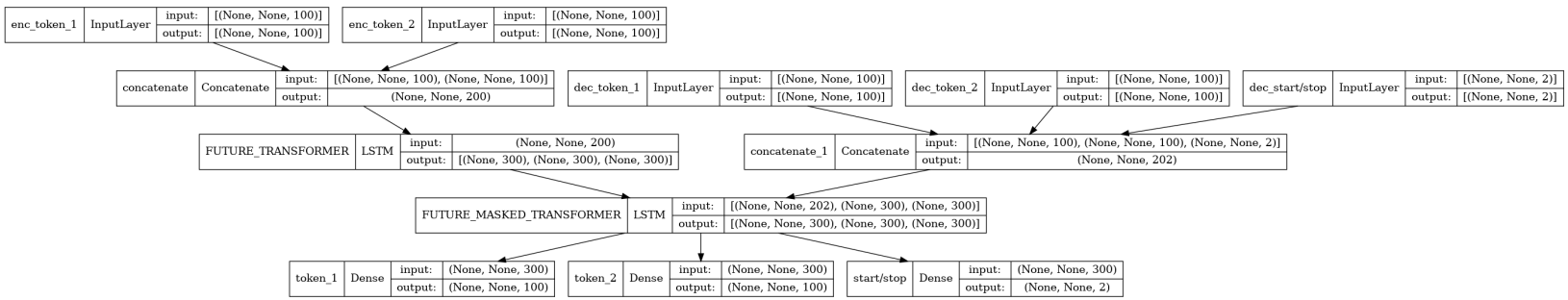

Figure: The Coupled Inner LSTM model for the LSTM/LSTM configuration

Re-Implementation:Transformer/Transformer

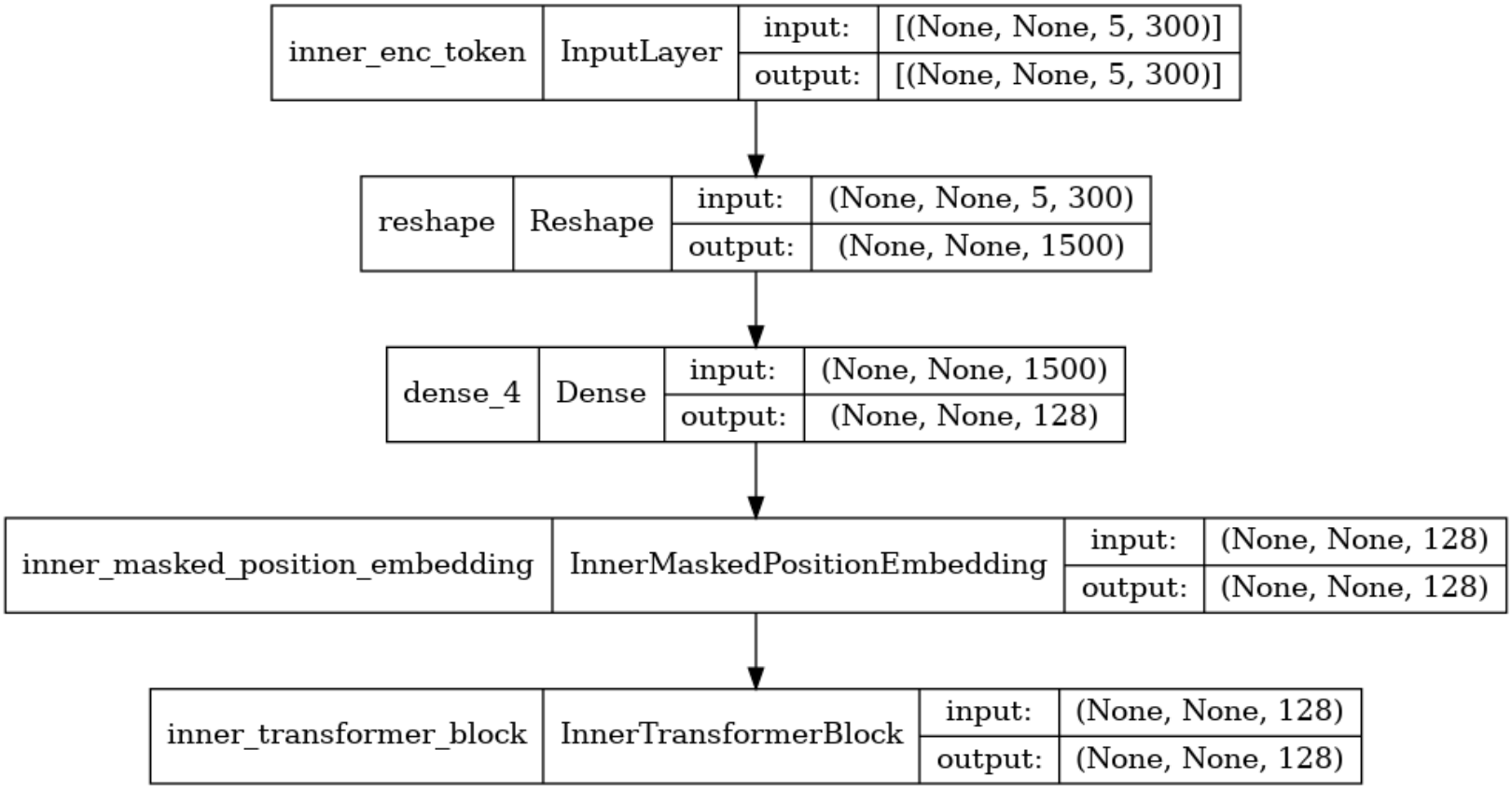

Figure: Inner Encoder for the Transformer/Transformer configuration

Re-Implementation: Transformer/Transformer

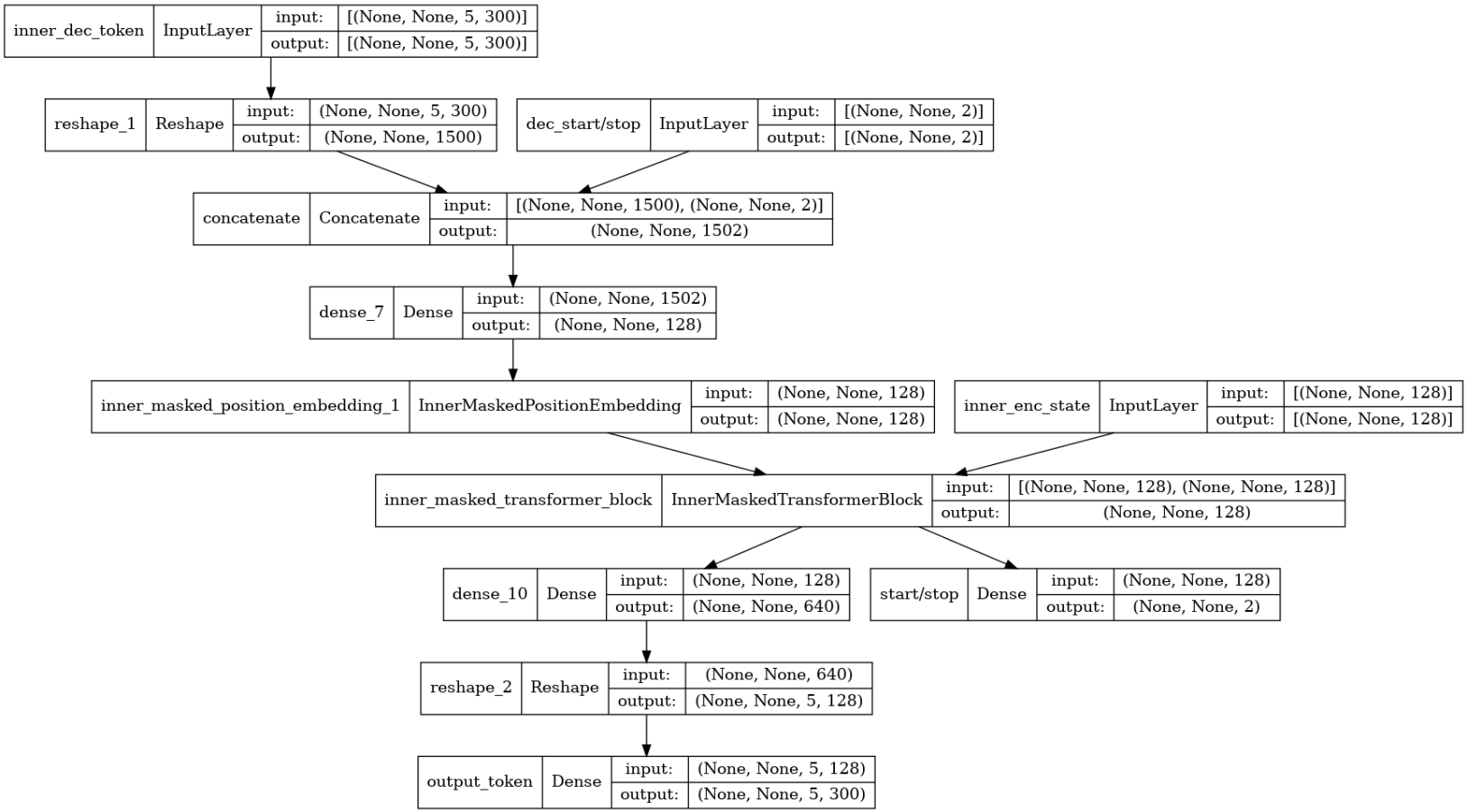

Figure: Inner Decoder for the Transformer/Transformer configuration

Re-Implementation: LSTM/Transformer

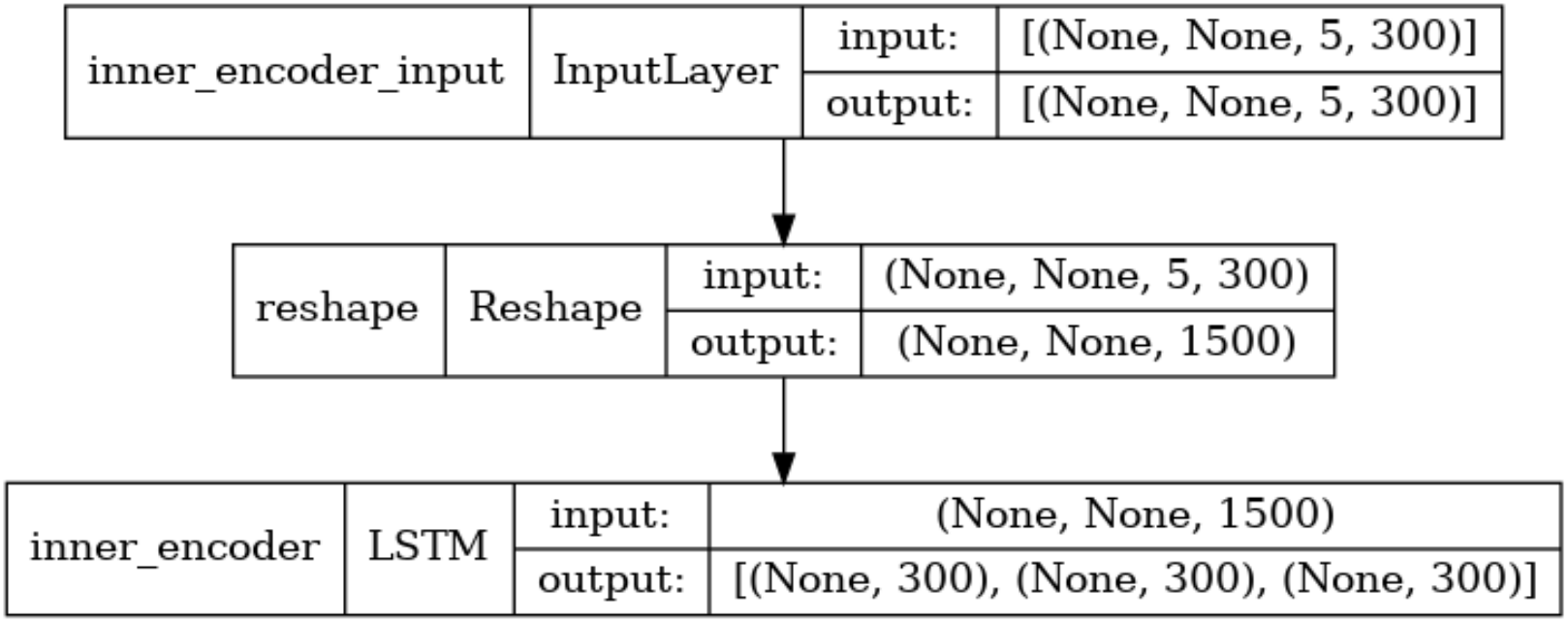

Figure: Inner Encoder for the LSTM/Transformer configuration

Re-Implementation: LSTM/Transformer

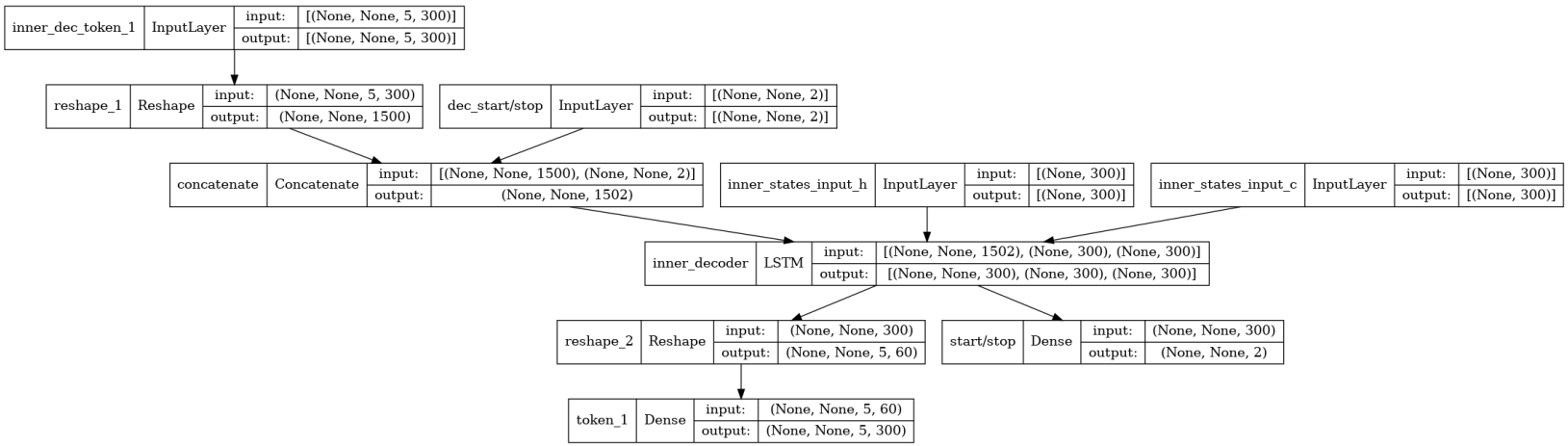

Figure: Inner Decoder for the LSTM/Transformer configuration

Re-Implementation: Transformer/LSTM

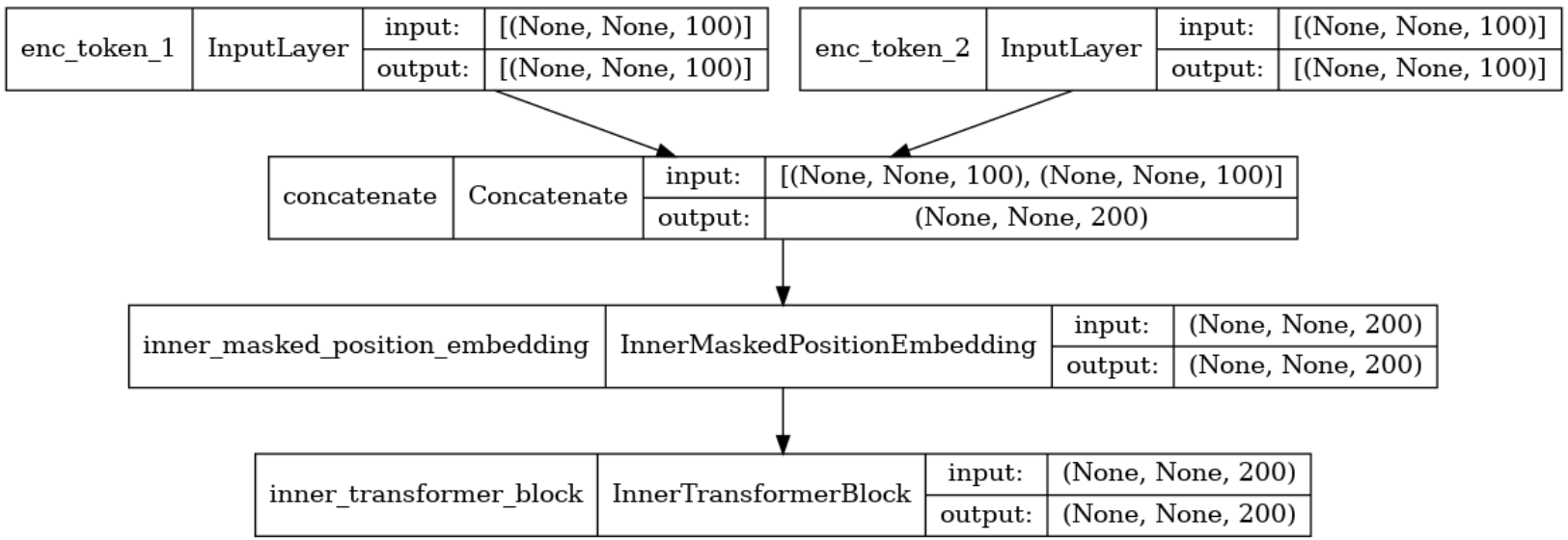

Figure: Inner Encoder for the Transformer/LSTM configuration

Re-Implementation: Transformer/LSTM

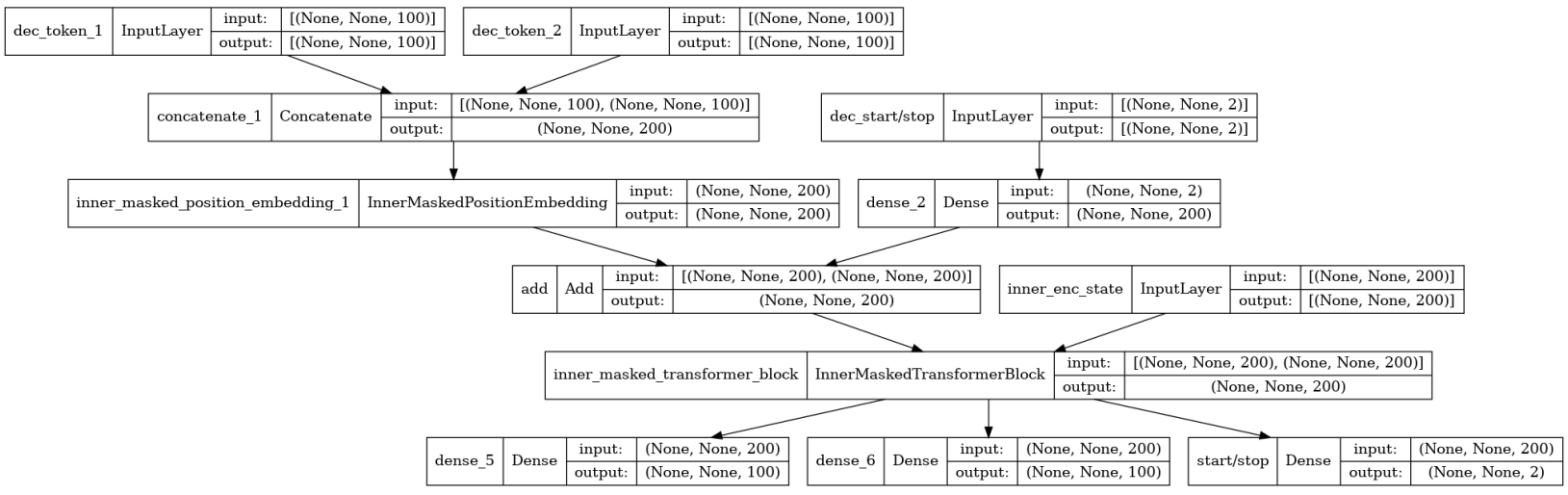

Figure: Inner Decoder for the Transformer/Transformer configuration

Testing

- The outer models were pretrained to 100% on a corpus of 1000 filler tokens
- The training and testing data for the full configurations was made from that corpus
- The filler tokens were five letters each
- The sequences were each composed of three filler tokens

Testing: The tests

Standard Generalization (SG) The training set presents the model with every filler being used in every role, as well as every filler being used in a sequence together. For example, if the fillers are "dog", "tooth", and "ball" and the roles are where they appear in a sequence then the training set could include "dog tooth ball", "tooth ball dog", and "ball dog tooth". The testing will have unique role-filler pairs. With our example a testing phrase could be "tooth dog ball".

Testing: The tests

Spurious Anticorrelation (SA) The training set includes all fillers being used in every role, but not all in the same sequence together. For example, if the fillers are "dog", "tooth", "ball", and "pig" and their roles are where they appear in the sequence then the training set could include, "dog tooth ball", and "tooth ball pig". The testing will have unique role-filler pairs that have not been in the same sequence together. With our example a testing phrase could be "dog ball pig" because "dog" and "pig" weren't put together in the training set. Note that this test should have sets where "dog", "tooth", "ball", and "pig" have a chance to be trained in every role.

Testing: The tests

Full Combinatorial (FC) The training set has fillers that are not used in every role. For example, if the fillers are "dog", "tooth", "ball", and "pig" and their roles are where they appear in the sequence the training set could include "pig tooth ball", "pig dog ball", and "pig tooth dog". The testing set will have fillers used in roles that are not in the training set. With our example a testing phrase could be "dog tooth ball". In this "dog" was tested in the first role but new filled that role in the training set.

Testing: The tests

Novel Filler (NF) The testing set includes fillers not in the training set at all. For example, some of the training sequences could be "dog tooth ball" "tooth ball dog", and "ball dog tooth". Following this example a testing sequence could be "pig dog tooth" as "pig" is not in the training set.

Results

	LSTM/LSTM		LSTM/Transformer		Transformer/LSTM		Transformer/Transformer	
Test	Word Accuracy	Letter Accuracy	Word Accuracy	Letter Accuracy	Word Accuracy	Letter Accuracy	Word Accuracy	Letter Accuracy
SG	98.333	99.555	73.333	92.777	97.666	99.0	100.0	100.0
SA	98.666	99.666	80.0	92.777	99.666	99.944	100.0	100.0
FC	5.333	28.444	20.0	55.0	52.666	65.555	100.0	100.0
NF	35.666	53.5	50.0	67.777	52.333	58.055	59.999	72.277

Figure: Results from one pass of training and testing

Discussion

Comparison

	LSTM/LSTM		LSTM/Transformer		Transformer/LSTM		Transformer/Transformer	
Test	Word Accuracy	Letter Accuracy	Word Accuracy	Letter Accuracy	Word Accuracy	Letter Accuracy	Word Accuracy	Letter Accuracy
SG	-0.1003	-0.0561	-2.0000	0.4440	-0.6340	-0.0833	0.0000	0.0000
Deviations	1.0651	0.2612	6.5168	1.9387	1.0899	0.7430	0.0000	0.0000
SA	-0.7340	-0.1451	4.6670	0.5550	0.7327	0.3940	0.0000	0.0000
Deviations	0.4899	0.1614	6.1262	2.0951	1.0520	0.5238	0.0000	0.0000
FC	-8.6003	-12.5893	3.3340	-0.6111	-32.1673	-26.1061	0.0667	0.0222
Deviations	4.1333	7.0561	2.2222	5.8352	9.1388	6.1480	0.2108	0.0703
NFs	-18.2673	-11.6611	-4.0010	-3.2785	-3.0003	-37.3894	2.4660	0.8604
Deviations	4.4642	4.0325	4.0976	3.7263	3.8151	4.2712	6.9143	4.6052

Figure: Comparison of the results from "Attention Is Not Enough" with the results from this study [2] The deviations were taken from "Attention Is Not Enough". The differences were calculated by subtracting the averaged results from the previous study from the one pass results from this one.

Discussion: Future Work

- Running the tests more times for more data
- Filling out the remainder of the missing code
- Investigation into varience in results compared to "Attention Is Not Enough"
- Look into other network architecture that may be more successful.

References

[1] Taylor W. Webb, Ishan Sinha, and Jonathan D. Cohen. "Emergent Symbols through Binding in External Memory". In: (Mar. 2021). Number: arXiv:2012.14601 arXiv:2012.14601 [cs]. DOI: 10.48550/arXiv.2012.14601. URL: http://arxiv.org/abs/2012.14601.

[2] In: (). URL: https://www.cs.mtsu.edu/~jphillips/papers/MillerNaderiMullinaxPhillips-CogSci-2022-preprint.pdf.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. "Attention Is All You Need". In: arXiv:1706.03762 (Dec. 2017). arXiv:1706.03762 [cs]. URL: http://arxiv.org/abs/1706.03762.

[4] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-term Memory". In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[5] Trenton Kriete, David C. Noelle, Jonathan D. Cohen, et al. "Indirection and symbol-like processing in the prefrontal cortex and basal ganglia". en. In: *Proceedings of the National Academy of Sciences* 110.41 (Oct. 2013), pp. 16390–16395. DOI: 10.1073/pnas.1303547110. URL: https://pnas.org/doi/full/10.1073/pnas.1303547110.

[6] Michael P. Jovanovich. "Biologically Inspired Task Abstraction and Generalization Models of Working Memory". en. PhD thesis. Middle Tennessee State University, Oct. 2017. URL: http://jewlscholar.mtsu.edu/xmlui/handle/mtsu/5561.