

Conditional Generative Transformers for Hand-Guided Animation Automation

by

Maxwell Clarke

A thesis

submitted to the Victoria University of Wellington

in partial fulfilment of the requirements

for the degree of

Master of Science

in Computer Science.

Victoria University of Wellington

November 2022

Abstract

In this thesis I develop neural networks and apply them to the problem of hand motion modelling for film production.

Firstly, I develop a novel training regime for transformers which allows auto-regressive sampling of high-dimensional data in an arbitrary order, and demonstrate this technique on pixel-by-pixel sampling of MNIST images.

Secondly, I develop a predictive model for hand-motion data, via unsupervised training on a motion-capture dataset. I compare fixed-order sampling to best order I found using the above technique.

Contents

1	Introduction	1
1.1	Motivation	4
1.1.1	Why focus on hand motion?	4
1.1.2	Why focus on transformers?	5
1.2	Previous Work	5
2	Background	7
2.1	Neural Networks and Deep Learning	7
2.1.1	Notation	7
2.1.2	Tasks	11
2.2	Auto-regressive models	15
2.3	Animation	15
2.4	Angle representations	15
3	Understanding Transformers	17
3.1	The Attention Operation	18
3.1.1	Mathematical Definition	18
3.1.2	Permutation-invariance with respect to K and V . . .	20
3.1.3	Permutation-equivariance with respect to Q	20

3.1.4	Dynamic length inputs	21
3.1.5	Parallel computation	21
3.2	Transformer models	22
3.2.1	Masked Sequence Modeling: Encoder-only models .	26
3.2.2	Sequence Prediction: Decoder-only models	26
3.2.3	Encoder-decoder models	27
3.2.4	Unified attention models	31
3.2.5	Comparison of architectures	32
3.3	Pretraining tasks	32
4	Sampling Sequences In Any Order	35
4.1	Autoregressive models	36
4.1.1	Distributions over high-dimensional data	36
4.2	Changing The Transformer Input	38
4.3	Tasks	39
4.3.1	Arbitrary order decoder-only transformer using in- put triples	39
4.3.2	Queries	40
4.4	Hypothesis	41
4.5	Method	42
4.5.1	42
4.5.2	Baseline: Fixed-order sequence prediction	42
4.6	Results	42
5	Angles, Joints and Hands	43
5.1	Parameterizing Hand Configurations	43
5.1.1	Euler Angles	45

<i>CONTENTS</i>	<i>v</i>
5.1.2 Axis-Angle	46
5.2 Loss Functions for learning angles	46
6 Hand Motion Model	49
6.1 Predicting the next frame	49
6.2 Learning a probabilistic model	49
7 Conclusions	51

List of Figures

1.1	The relationship between the different fields in this thesis . .	2
1.2	Where my work sits.	3
2.1	Ontology of loss functions.	12
2.2	Ontology of dataset types	13
2.3	Fixed vs variable input shape	14
3.1	Typical residual block in transformer	23
3.2	Self-attention	25
3.3	Cross-attention	25
3.4	Self-attention with causal masking	27
3.5	Partial self-attention	28
3.6	Transformer model	29
3.7	Unified attention	30
3.8	Attention Masks	31
3.9	Masked Language Model Pretraining	33
5.1	Joints of the hand	44

List of Tables

Chapter 1

Introduction

In this thesis I discuss my learnings and experiments from the past year, which are at the intersection of two areas: Deep Learning and Computer Graphics. There are three main areas of focus:

1. I experiment with *transformer* models, and understand them in depth (Chapter 3).
2. I experiment with using Transformer models for Bayesian inference, essentially as Gaussian Processes (Chapter 4).
3. I create a proof of concept machine learning application for *hand motion prediction*, useful for film and video game production (Chapters 5 and 6).

Although much of my work is summarizing others' research and presenting my learnings, I have two main novel contributions:

1. I present experiments with *dynamically-ordered* auto-regressive sampling, which utilises the *permutation-invariance* of the attention oper-

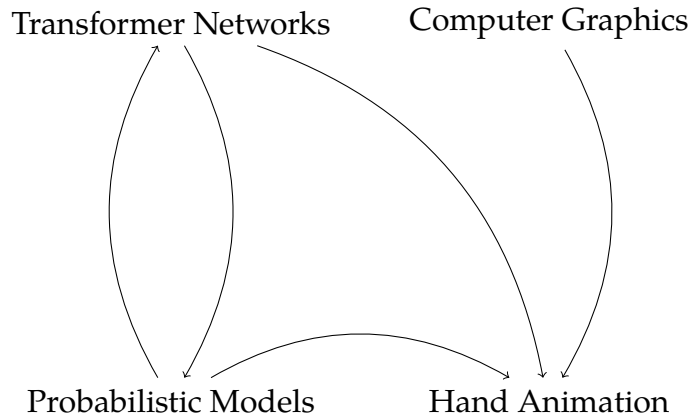


Figure 1.1: How learnings and experiments from different fields contribute to each other in this thesis.

ation in transformer models.

2. I present a proof-of-concept transformer-based generative model for hand motion prediction, which can be used to predict hand motion at arbitrary target frames, and to predict the joints of a hand in any order within that frame.

These contributions involved training neural networks (see Figure 1.2), in particular transformers, on two datasets - MNIST, and a motion capture dataset of hand motion. The results of these experiments are presented in Chapter 4 and Chapter 6 respectively.

In the rest of this chapter, I will discuss the motivation for this thesis, and previous work that relates to it.

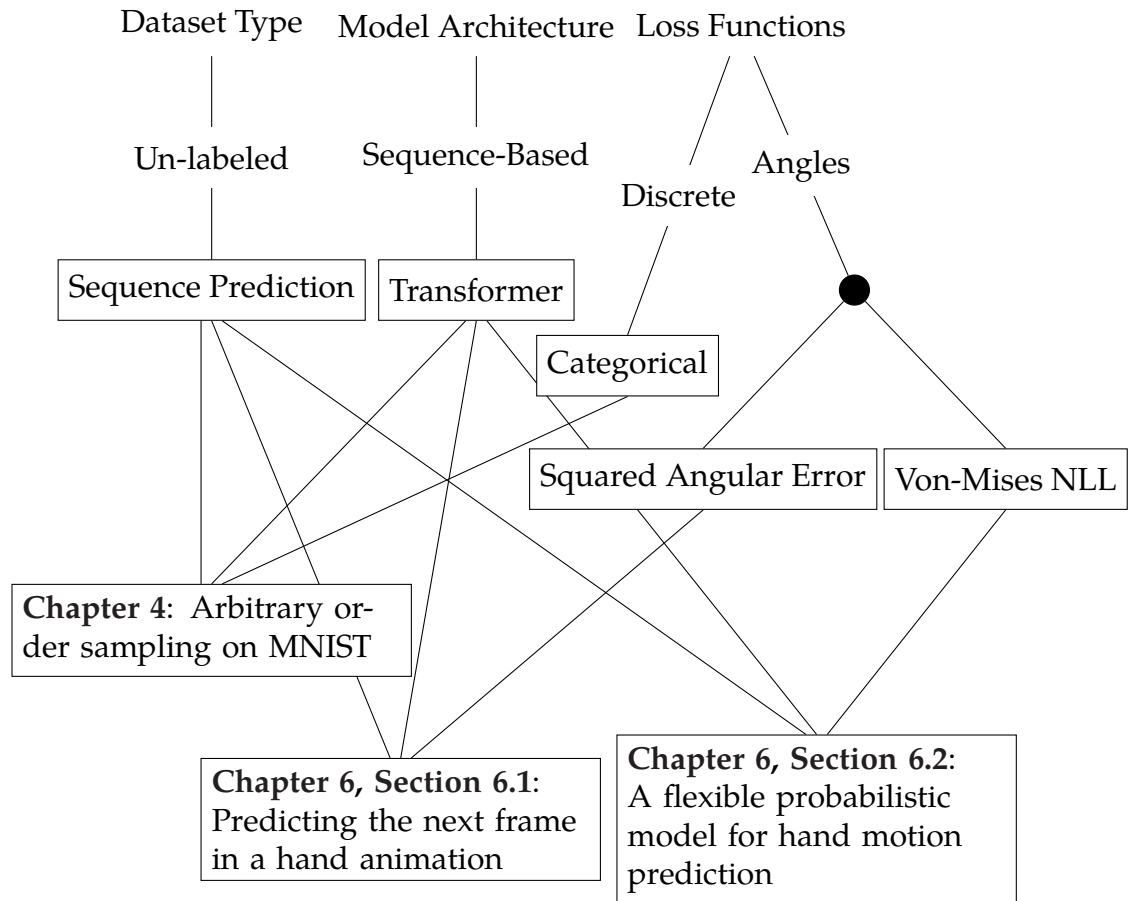


Figure 1.2: The later work in this thesis sits focuses on learning un-labeled sequence data with transformers, in two different domains.

In Chapter 4, I train a transformer-based probabilistic model which can be used as a gaussian process for predicting pixels on the MNIST dataset.

In Chapter 6, I train a transformer-based model on the ManipNet hand motion dataset. Section 6.1 focuses on a deterministic model, while Section 6.2 focuses on a probabilistic model.

1.1 Motivation

The problem domain I focused on – hand motion prediction – is a sequence prediction problem, which are the general kind of task that transformer models are used on. To this end, I hoped to find that these two motivations would feed back into each other as I worked – the application providing direction for the more general / theoretical research, and the insights gained from the more general research contributing back to better solutions for the problem domain.

1.1.1 Why focus on hand motion?

Whenever a moving virtual character appears in an animated film or a video game, someone had to spend the time to specify the angles of all the joints. Fortunately is not necessary to lay out every single frame, because animation tools used for both games and film production make use of interpolation techniques between key-frames, but artists must still specify many joints, over many key-frames, over many different kinds of animation.

In order to fully animate a character, an animator needs to appropriately animate all moving parts of that character. We can divide these into three groups, each involving a similar amount of work:

- Facial animation – animating the muscles of the face when a character talks or otherwise makes facial expressions.
- Hand animation – animating the fingers and wrists when a character makes gestures or manipulates objects.

- Body animation (also simply called character animation) – animating the rest of the body, e.g. the legs, arms, neck and spine when a character walks, dances, etc.

Of these, the face and body are often the most noticeable, and most of the time receive the most attention from the audience. However, the hands are also very important, and are often the most difficult to animate.

todo: Examples of problematic hand animation.

1.1.2 Why focus on transformers?

1.2 Previous Work

Chapter 2

Background

2.1 Neural Networks and Deep Learning

Since Krizhevsky et al.'s AlexNet [[alexnet](#)] in 2012, many problems are increasingly being solved best by Artificial Neural Networks (NN). Any particular NN is not designed but discovered by gradient descent, where the parameters of the NN are iteratively improved with respect to a loss function, and a dataset.

Many variants exist, and in this section I firstly introduce and clarify my notation, and then give an overview of the different aspects of neural networks to contextualize my work.

2.1.1 Notation

Since we are often using tensors of rank 3 or higher in neural network implementations, it is useful to have some notation that clarifies how functions are applied across these dimensions.

First, a simple example using activation functions. The ReLU activation function is defined as:

$$\begin{aligned}\phi_{\text{relu}} : \mathbb{R} &\rightarrow \mathbb{R} \\ \phi_{\text{relu}}(x) &\stackrel{\text{def}}{=} \max(0, x)\end{aligned}\tag{2.1}$$

It is customary to use the symbol ϕ for activation functions. Activation functions are typically scalar functions, but are applied independently across all components of a tensor. To represent such, I will use the following notation, for example, applying the ReLU activation function to a vector \mathbf{x} :

$$\mathbf{x}'_i = \phi_{\text{relu}}(\mathbf{x}_i)$$

In the above equation, the subscript i shows that the function is applied *independently* to each component of the vector \mathbf{x} .

This notation also works for multiple dimensions, including when an operation is not applied independently across some dimension. For example, the following is how I will show the *softmax* function, which is a vector valued function, applied independently across the rows of a $B \times N$ matrix \mathbf{X} (which is used in the definition of the attention operation in Chapter 3).

The softmax function is defined as:

$$\begin{aligned}\sigma : \mathbb{R}^N &\rightarrow \mathbb{R}^N \\ \sigma(\mathbf{x}_n) &\stackrel{\text{def}}{=} \frac{e^{\mathbf{x}_n}}{\sum_{n'} e^{\mathbf{x}_{n'}}}\end{aligned}\tag{2.2}$$

We can apply the the above function to a matrix $\mathbf{X} \in \mathbb{R}^{B \times N}$ independently

over B as follows:

$$\begin{aligned}\mathbf{X}'_{b,n} &= \sigma(\mathbf{X}_b)_n \\ &= \frac{e^{\mathbf{X}_{b,n}}}{\sum_{n'} e^{\mathbf{X}_{b,n'}}}\end{aligned}$$

The term \mathbf{X}_b refers to the b -th row of X as a vector, as is common in tensor algebra software such as NumPy [7] or TensorFlow [9]. Here however the order of the subscript indices b and n is ignored – the indices index their respectively-named dimensions. This that we abandon the distinction between row- and column-vectors. I will thus be explicit when applying matrix and vector products.

I will now define some simple neural networks as examples for clarifying any later notation.

I will typically use N for the input dimensionality of a network, and D for the *embedding* (or *hidden* / *latent*) dimensionality. Let $\mathbf{x} \in \mathbb{R}^N$ be some input data embedded into an N -dimensional vector space. Let $W \in \mathbb{R}^{N \times D}$ be a matrix of learned weights, and let $\phi: \mathbb{R} \rightarrow \mathbb{R}$ be some non-linear function. Then, the computation done by one layer of a simple fully-connected neural network is represented as follows.

$$\begin{aligned}f_{\text{mlp}}: \mathbb{R}^N &\rightarrow \mathbb{R}^D \\ f_{\text{mlp}}(\mathbf{x}) &\stackrel{\text{def}}{=} \phi(W\mathbf{x}) + \mathbf{b}\end{aligned}\tag{2.3}$$

$$W \in \mathbb{R}^{N \times D}, \quad \mathbf{b} \in \mathbb{R}^D$$

W is the weight matrix, and \mathbf{b} is the bias vector, which together are the parameter set for this simple model. The output of the neural network is a

D -dimensional vector.

A simple classifier network would be defined as follows, for N dimensional data classified into C classes, with L hidden layers:

$$\begin{aligned}
 f_0: \mathbb{R}^N &\rightarrow \mathbb{R}^D \\
 f_0(\mathbf{x}) &= \phi(W_0\mathbf{x}) + \mathbf{b} \quad W_0 \in \mathbb{R}^{N \times D} \quad \mathbf{b}_0 \in \mathbb{R}^D \\
 f_\ell: \mathbb{R}^D &\rightarrow \mathbb{R}^D \quad \forall \ell \in 1, \dots, L \\
 f_\ell(\mathbf{x}) &= \phi(W_\ell f_{\ell-1}(\mathbf{x})) + \mathbf{b}_\ell \quad W_\ell \in \mathbb{R}^{D \times D} \quad \mathbf{b}_\ell \in \mathbb{R}^D \\
 f_L: \mathbb{R}^D &\rightarrow \mathbb{R}^C \\
 f_L(\mathbf{x}) &= \sigma(W_L f_{L-1}(\mathbf{x}) + \mathbf{b}_L) \quad W_L \in \mathbb{R}^{D \times C} \quad \mathbf{b}_L \in \mathbb{R}^C \\
 f_{\text{classifier}}: \mathbb{R}^N &\rightarrow \mathbb{R}^C \\
 f_{\text{classifier}} &= f_L \circ f_{L-1} \circ \dots \circ f_0 \quad \theta = \{W_0, \dots, W_L, \mathbf{b}_0, \dots, \mathbf{b}_L\}
 \end{aligned} \tag{2.4}$$

The parameters of the network are $\theta = \{W_0, \dots, W_L, \mathbf{b}_0, \dots, \mathbf{b}_{L-1}\}$. The output of the network is a C -dimensional vector, where each component is the probability that the input belongs to that class. This model would be trained with a categorical cross-entropy loss function – which I will discuss in the next section.

2.1.2 Tasks

Neural networks are applied to a wide variety of tasks, which lead to a number of different choices for the architecture and loss function. In this section, I will help to contextualize my later work by giving a brief overview of the ways that different neural networks and training setups differ.

On the following pages, I give some simple ontologies of the different considerations that combine to define a particular task and architecture, in particular:

1. Training objective / loss function (Figure 2.1)
2. Data dimensionality / length (Figure 2.3)
3. Dataset format (Figure 2.2)

The choice of training objective affects the settings in which a model can be used, which theoretical properties we get from it, and more. The structure of the data affects the type of model that can be used, and the format of the dataset affects what tasks we can learn from it.

The simplest kind of training objective is regression. When we train a model with a regression objective it learns to predict the expected value of the output. Regression is characterized by using an error function as the loss, for example *mean-squared-error*:

$$L_{\text{MSE}}: \mathbb{R}^{N \times D} \times \mathbb{R}^{N \times D} \rightarrow \mathbb{R}$$

$$L_{\text{MSE}}(y, \hat{y})_{ni} \stackrel{\text{def}}{=} \frac{1}{N} \sum_n \left[\sum_i (y_{ni} - \hat{y}_{ni})^2 \right] \quad (2.5)$$

This function sums the error over the *feature* dimension D and averages

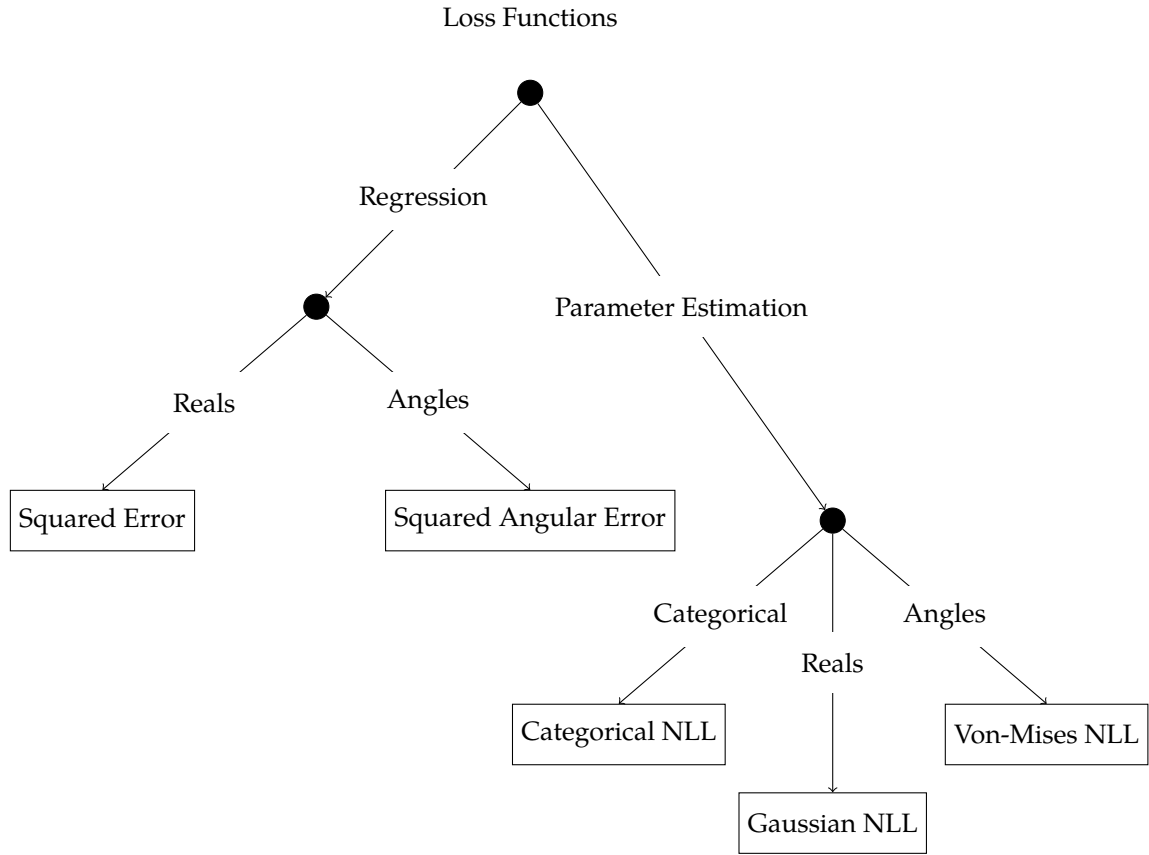


Figure 2.1: We can split loss functions into two categories.

In the former, the loss function has the form of an error function. When minimizing this function, the model learns to output the expected value of the posterior $E[p(y | x)]$ of the output y given the input x . This is called a regression or maximum-a-posteriori (MAP) task.

In the latter, the loss function has the form of a negative-log-likelihood (NLL) function. The model outputs the parameters of a probability distribution, and the loss function is the negative log-likelihood of the data under that distribution. This includes the case of categorical NLL (also called categorical cross-entropy), where the model outputs a probability distribution over a discrete set of classes.

Models trained with a NLL loss learn to output an explicit posterior distribution $p(y | x)$, given a fixed functional form for p , such as a Gaussian, mixture of Gaussians, Categorical, Von-Mises, etc. Depending on the task, and output format, different functional forms for p may be appropriate.

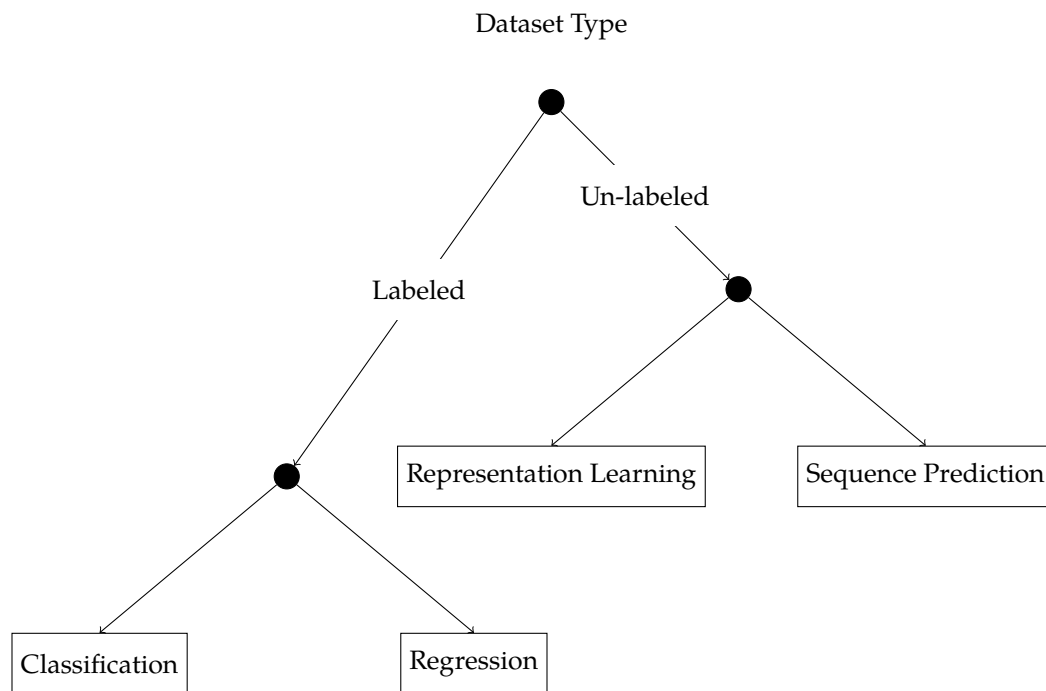


Figure 2.2: Basic ontology of dataset types.

When learning on unlabeled data, the goal is to learn a representation of the data that is useful for some downstream task.

When learning on data that is explicitly labeled – the goal is to learn a model that performs well on the task directly.

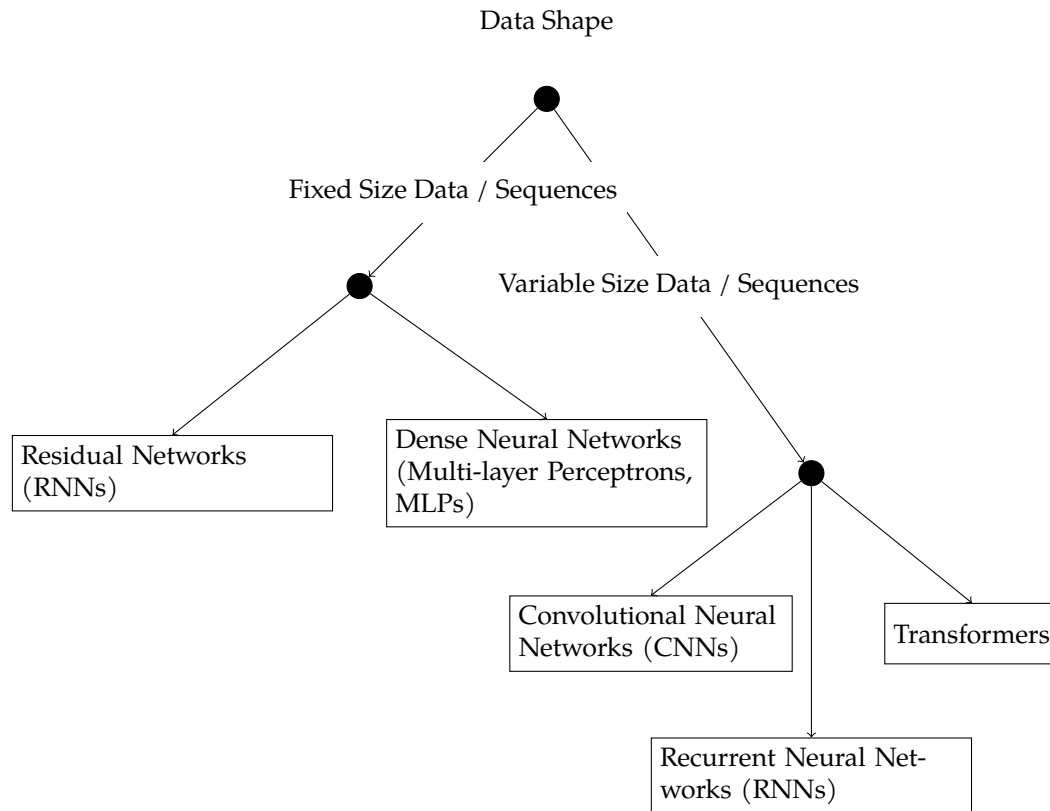


Figure 2.3: Neural network variants which support variable input shape.

Due to their construction, MLPs and ResNets are restricted to a fixed input shape, and so can only be trained and used on data that is of a fixed size, such as tabular data, or data that has been processed into a fixed size by re-sampling, chunking etc.

RNNs, CNNs and Transformers can accept variable length data, each with their own tradeoffs. They are typically more suitable for data that is naturally of variable size/length, such as text, audio or images.

the error over the *batch* dimension N ¹. Training a model by minimising this loss function, is equivalent to maximising the likelihood of a Gaussian distribution. Given input x , the model output $y = f(x)$ can be interpreted as $E[p(y|x)]$.

2.2 Auto-regressive models

todo: Formulation of auto-regressive models and sampling

todo: Things we can do with an auto-regressive model

2.3 Animation

2.4 Angle representations

todo: Quaternions and unitary / complex matrices

¹Averaging has no effect on the optimization, it is simply that dividing by the batch and/or sequence length means that the loss value remains in the same range independent of the batch size or sequence length.

Chapter 3

Understanding Transformers

Many of the recent amazing results in deep learning have been achieved with a class of neural networks called *transformers*, which were introduced and named in "Attention Is All You Need", Vaswani et al. 2017 [13]. The distinguishing feature of these models is using one or more *attention* layers to enable information propagation between elements in sequence data. Since 2017, a large number of transformer variants have been developed, and in this chapter I seek to understand this class of models at a broader level. I will discuss:

- The unique properties of the attention operation, which include working with sequences of any length without changing the weights, being able to be computed in parallel across the sequence during training, and being invariant to the order of the inputs.
- The different variants of transformers, namely encoder-only, decoder-only, and encoder-decoder models.
- The variety of different tasks that these model are trained on and used

for, building up to the next chapter where I introduce an extremely generic gaussian-process-like task.

Firstly, I will discuss the attention operation that under-pins transformers.

3.1 The Attention Operation

Attention is a biologically-inspired mechanism that allows a model to receive inputs from distant parts of the input data, weighted by the *attention weight* given to those inputs, which is typically computed from the data itself. This has proven extremely useful for diverse tasks including machine translation, image generation, and more. In this section I will describe the attention operator.

Attention has a number of useful properties which come from its mathematical construction, such as permutation-invariance in the inputs.

3.1.1 Mathematical Definition

An attention operation is of the following form, using short summation notation, where σ is the *softmax* operator (see 2.2), and A is the pre-softmax attention logits.

$$f_{\text{attn}}: \mathbb{R}^{M \times D} \times \mathbb{R}^{N \times D} \times \mathbb{R}^{N \times V} \rightarrow \mathbb{R}^{M \times V}$$

$$f_{\text{attn}}(Q, K, V)_{mv} \stackrel{\text{def}}{=} \sum_n \left[\sigma \left(\sum_d Q_{md} K_{nd} \right)_{mn} V_{nv} \right] \quad (3.1)$$

$$Q \in \mathbb{R}^{M \times D}, K \in \mathbb{R}^{N \times D}, V \in \mathbb{R}^{N \times V}$$

$$M, N, D, V \in \mathbb{N}$$

The innermost multiplication of Q and K is simply the inner product (dot product) between vectors Q_m and K_n . This however is not inherent. Instead of the inner product, we can substitute any kernel function. (Although this is not usually done because the inner product is the most natural choice, and is efficient to compute)

For clarity, the expanded form of the attention computation, resulting in the unnormalized attention weights A , for an arbitrary kernel function k , is as follows:

$$A_{mn} = k(Q_m, K_n) = \begin{bmatrix} k(Q_1, K_1) & k(Q_1, K_2) & \cdots & k(Q_1, K_N) \\ k(Q_2, K_1) & k(Q_2, K_2) & \cdots & k(Q_2, K_N) \\ \vdots & \vdots & \ddots & \vdots \\ k(Q_M, K_1) & k(Q_M, K_2) & \cdots & k(Q_M, K_N) \end{bmatrix}$$

or when $k(a, b) = a \cdot b$, then

$$\begin{aligned} &= \begin{bmatrix} Q_{1,1} & Q_{1,2} & \cdots & Q_D \\ Q_{2,1} & Q_{2,2} & \cdots & Q_D \\ \vdots & \vdots & \ddots & \vdots \\ Q_{M,1} & Q_{M,2} & \cdots & Q_D \end{bmatrix} \begin{bmatrix} K_{1,1} & K_{1,2} & \cdots & K_D \\ K_{2,1} & K_{2,2} & \cdots & K_D \\ \vdots & \vdots & \ddots & \vdots \\ K_{N,1} & K_{N,2} & \cdots & K_D \end{bmatrix}^T \\ &= QK^T. \end{aligned}$$

We can see that the attention weights A have shape $M \times N$. This is the primary drawback of the attention operation. M and N are typically both large, and it takes $O(MN)$ space and time to compute. Despite this

drawback, the attention operation has proven extremely useful in a variety of tasks. There are many different ways to address this but I will not discuss them.

3.1.2 Permutation-invariance with respect to K and V

The first interesting property of attention is that it is permutation-invariant with respect to the key and value inputs. This property is more or less useful depending on the task. For example, in the case of graphs, or sets of heterogeneous values, there may not be a natural ordering in which to process the inputs. In this case, we do not have to introduce any artificial ordering. (However, when *sampling* outputs, we typically still need to decide on some order. I will discuss this in more detail in Chapter 4).

This property is due to the construction of the attention operator. We can see that the output O_m corresponding to a query vector Q_m is independent of the order of the key and value vectors K_n and V_n , because the summation across n is commutative:

$$O_m = \sum_n V_n \sigma(A_m)_n \quad (3.2)$$

3.1.3 Permutation-equivariance with respect to Q

Relatedly, attention is also permutation-*equivariant* with respect to the query inputs. Equivariance means that the value of the output O_m is dependent on the value of the query vector Q_m , but independent of the order of all other query vectors $Q_{m'}, m' \neq m$. This property is due to the fact that softmax operation is equivariant to the order of its inputs, which we can see

from the construction in Equation (2.2).

This property of attention stands in contrast to the two main other methods used to process sequence data, convolution (CNNs) and recurrence (RNNs). Neither of these operations are invariant (or equivariant) with respect to their inputs.

3.1.4 Dynamic length inputs

The second (and most useful) property of attention is that it can be used to process inputs of dynamic length. We can again see why this is the case from Equation (3.2). The softmax operation normalizes the attention weights, which causes the resulting summation of vectors V_n to be a convex combination. The resulting output O_m will therefore sit within the convex hull of the vectors V_n . This means that the output O_m will be a “valid” output regardless of the length of the input sequence K_v .

3.1.5 Parallel computation

The third property of attention is that it can be computed in parallel across the inputs sequence during training. At all steps of the attention computation except the softmax operation, there are no dependencies between neighbouring elements of the tensors. This stands in contrast to RNNs, where the outputs for one position in the sequence depend on the previous outputs.

The fact that attention can be computed in parallel is a very useful property, since while the operation requires $O(MN)$ time and space, we can utilize bigger GPU hardware the real-time cost to $O(1)$ (assuming sufficient

memory capacity, compute capability, and also GPU memory bandwidth, which is often the limiting factor [12].) The attention logits A_{mn} can be computed entirely in parallel. The softmax operation depends on all previous attention logits across the key-value dimension N , which requires cross-talk between GPU units but does not prevent parallel computation. The final computation for the outputs O_m can also be computed in parallel.

So attention is a operation with a number of useful properties. Now we will see how it is used to build a variety of models capable of solving a variety of tasks with sequence data.

3.2 Transformer models

An attention operation model does not itself make a neural network. It is simply a building block that can be used to construct a neural network.

A transformer model puts attention operations together with MLP blocks similarly to in residual networks (ResNets). Without MLP blocks as non-linearities, the outputs of an attention operation are linear functions of the inputs, since the softmax is only used to compute coefficients when summing the values. This means that the outputs of an attention operation are linear combinations of the inputs. This is useful as a building block but we would like to be able to learn non-linear functions of the inputs. MLP blocks are used to introduce non-linearities into the model.

In Figure 3.1 we can see a diagram of a typical residual block in a transformer, combining an attention operation, a feed-forward layer, and a residual connection as in a ResNet.

Attention is computed from the three matrices (or sequences of vec-

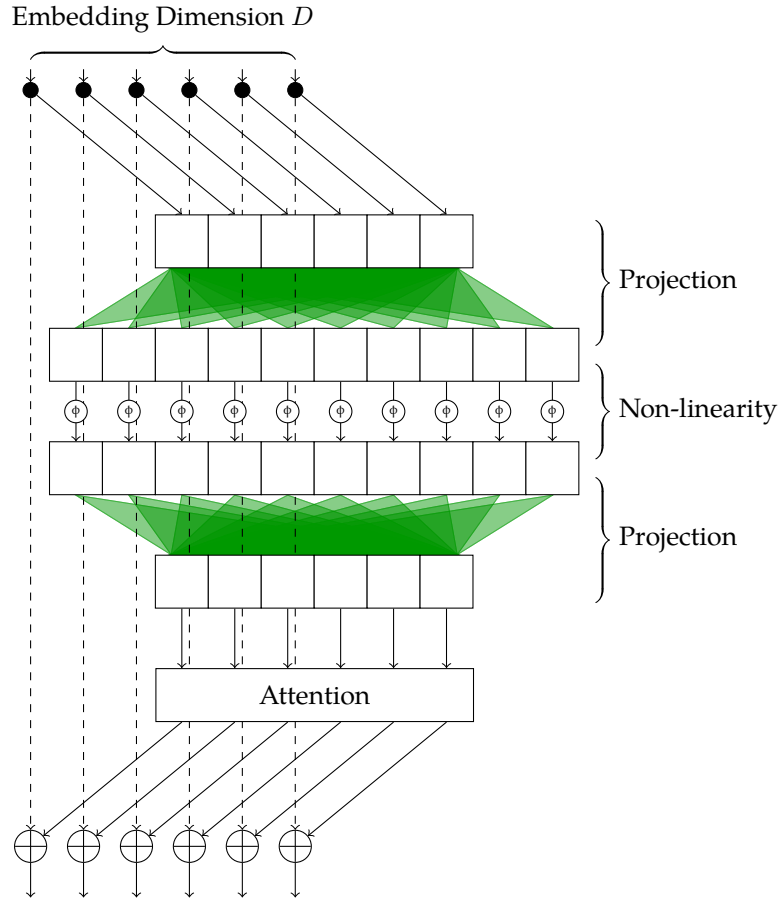


Figure 3.1: Typical residual block in transformer (with the D dimension expanded, rather than the N or M dimension as usual.).

In the MLP block, the sequence of residual latents are independently projected into a higher-dimensional space, where a non-linearity is applied, and then projected back down to the original dimension. The output of the MLP block is then projected into Q , K , and V spaces, and the attention operation is applied. Finally, the results are added to the residual stream.

tors) Q , K and V . In a neural network, these are each typically derived in some fashion from the inputs to the network. The most common way to do this is to use a learned linear transformation, which is simply a matrix multiplication followed by a bias term, for example

$$\begin{aligned} Q &= W_Q \mathbf{X} + \mathbf{b}_Q \\ K &= W_K \mathbf{X} + \mathbf{b}_K \\ V &= W_V \mathbf{X} + \mathbf{b}_V \end{aligned} \tag{3.3}$$

If we derive all three matrices from the same input \mathbf{X} , then $M = N$ and the attention operation is called a *self-attention* operation. A diagram of this is shown in Figure 3.2. The blue shaded areas show the receptive field used when computing each output vector x'_i .

When Q and K are derived from separate sequences of feature/embedding vectors, then in general $M \neq N$ and this is called *cross-attention*. A diagram of this is shown in ??.

Since the introduction of transformers it is common to use *multi-head* attention, which allows for multiple *heads* which each perform an attention operation in parallel with smaller key dimensionality $D_{\text{head}} = \frac{D}{n_{\text{heads}}}$.

The defining feature of a transformer model is that it has “attention” layers. However, there is not just one way to assemble these layers, and there is not just one way to train these models.

We can broadly split the transformer architecture variants into four categories: *encoder-only* models, *decoder-only* models, unified attention models, and *encoder-decoder* models.

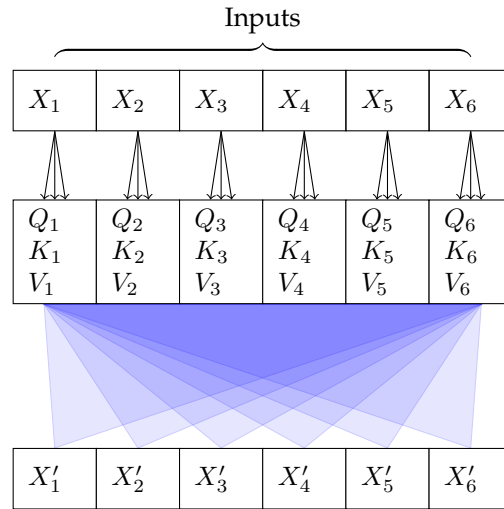


Figure 3.2: Full self-attention (bi-directional attention), as used in transformer encoders. The blue shaded regions show which inputs are used to compute each output. In full self-attention, all inputs are used to compute all outputs.

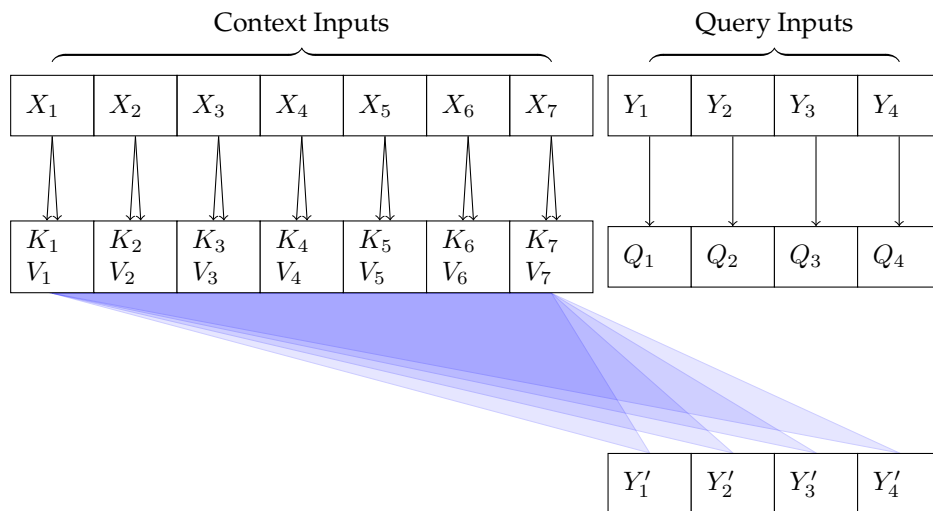


Figure 3.3: Cross-attention, as used in encoder-decoder models.

3.2.1 Masked Sequence Modeling: Encoder-only models

Arguably the simplest attention-based model architecture is encoder-only transformers. When used in natural-language-processing (NLP) they are known as bi-directional language models, because they allow information to flow in both directions. Examples are the BERT [4] language model family, Wav2Vec [1] for speech, and SimMIM [14] image model.

These models are trained on a sequence reconstruction task, called Masked Language Modeling (MLM), Masked Image Modeling (MIM), or more generally Masked Sequence Modeling, which I will discuss more in Section 3.3.

These models are typically used for sequence understanding tasks and classification tasks, however they can also be used for generating sequences. The limitation of these kinds of models is that their pretraining task is not efficient for training these models to generate sequences. For generating sequences, an encoder model in conjunction with a *decoder* model (see ?? 3.2.3), or simply a decoder-only model.

3.2.2 Sequence Prediction: Decoder-only models

A diagram of the decoder-only architecture, is shown below in ?. The distinguishing feature of a *decoder* as opposed to an encoder is that its attention layers are all causally-masked self-attention layers as in Figure 3.4. These models are used for sequence prediction/generation, and trained via self-supervised learning.

Some examples of where we see this architecture in use are:

- OpenAI’s GPT-series [10, 2] language models.

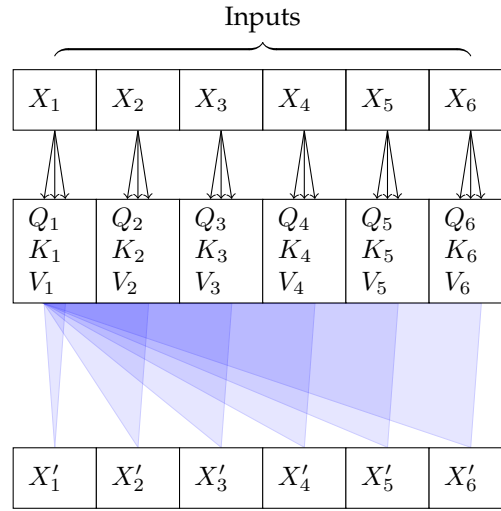


Figure 3.4: Self-attention with causal masking, (uni-directional attention) as used in transformer decoders during training. The blue shaded regions show which inputs are used to compute each output. In causal masking, only inputs to the left of the current output are used to compute the current output.

- Latent code prediction (the “prior”) in VQ-GAN [6]

3.2.3 Encoder-decoder models

When the transformer was introduced in [13], the first architecture proposed was an encoder-decoder architecture. This is a model which has both an encoder and a decoder. The encoder is used to encode a sequence of *conditioning* or *context* inputs, and the decoder is used to generate the output sequence. The encoder and decoder are connected by cross-attention layers (see Figure 3.3), which allow the decoder to attend to the encoded context sequence.

These are the most flexible class of model, because they allow predicting some outputs, or sequences of outputs, *given* some inputs.

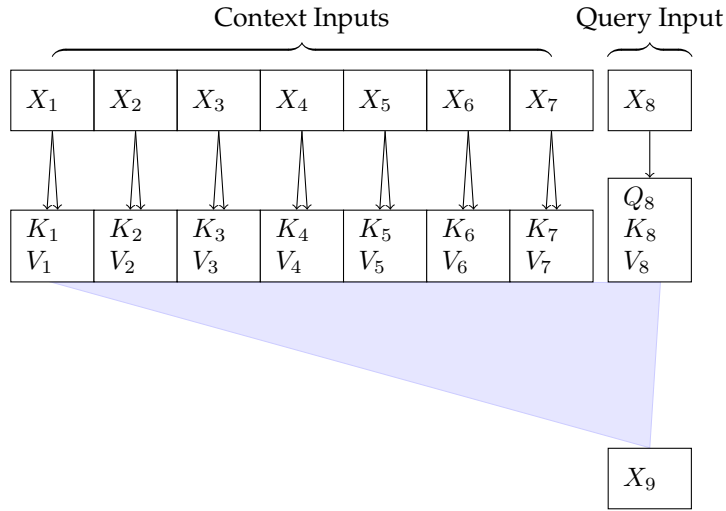


Figure 3.5: Partial self-attention, as used during incremental autoregressive inference (in models with a decoder).

Examples of this are the original transformer architecture [13], the BART [8] model, and more recently Google’s Parti multi-modal text-to-image model [15].

In [13], they train an encoder-decoder architecture for text translation. Their architecture takes one sequence of text tokens (if you are wondering what text tokens are, I will discuss input formats in the next section) in language A as conditioning input, then auto-regressively samples a target sequence in a second language B. The two languages may have very different word orderings or numbers of words to each other, but the cross-attention operation introduces no bias towards aligned word orderings or even word counts.

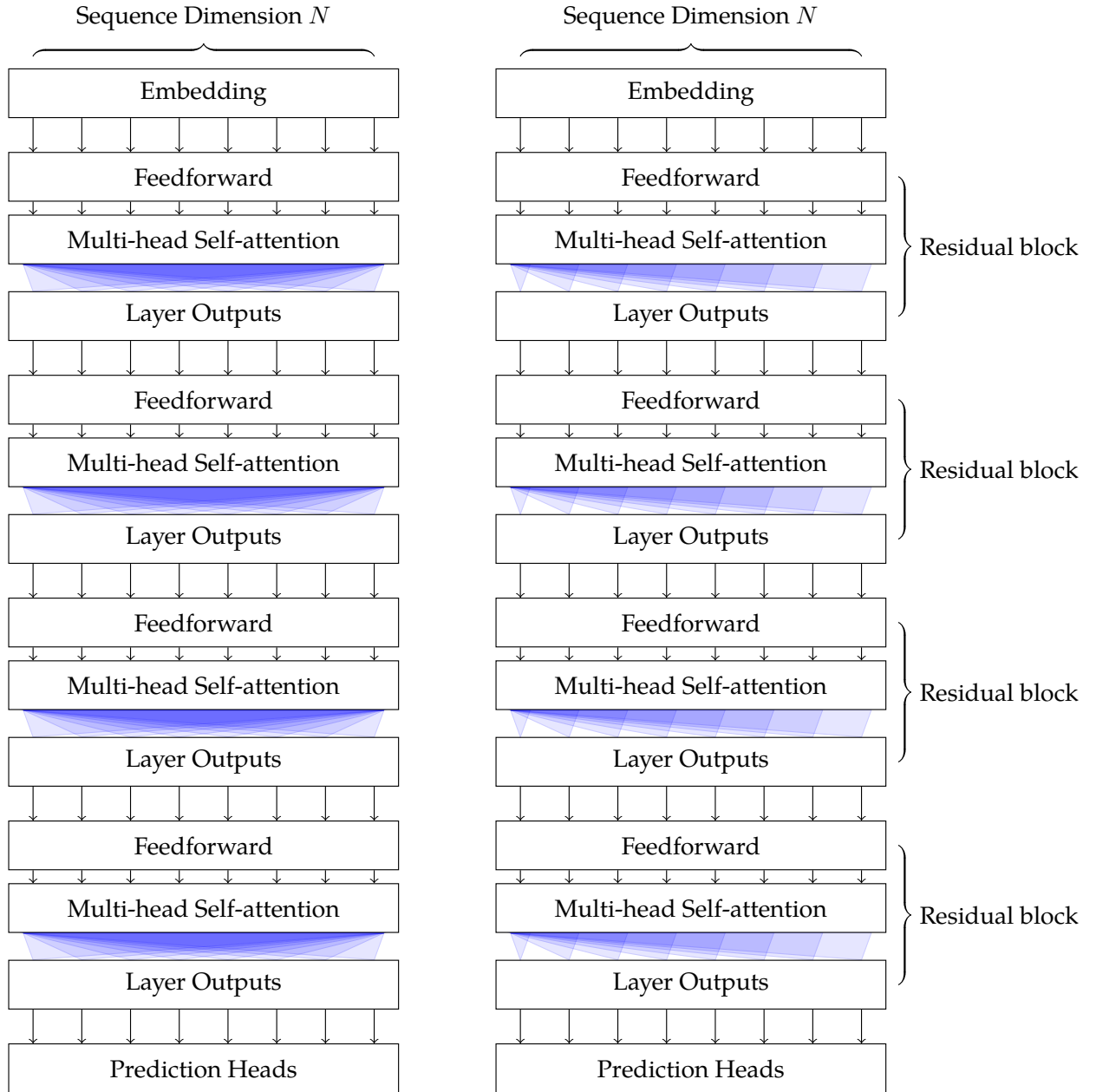


Figure 3.6: Left: Encoder-only model. Right: Decoder-only model. Residual connections have been omitted for brevity.

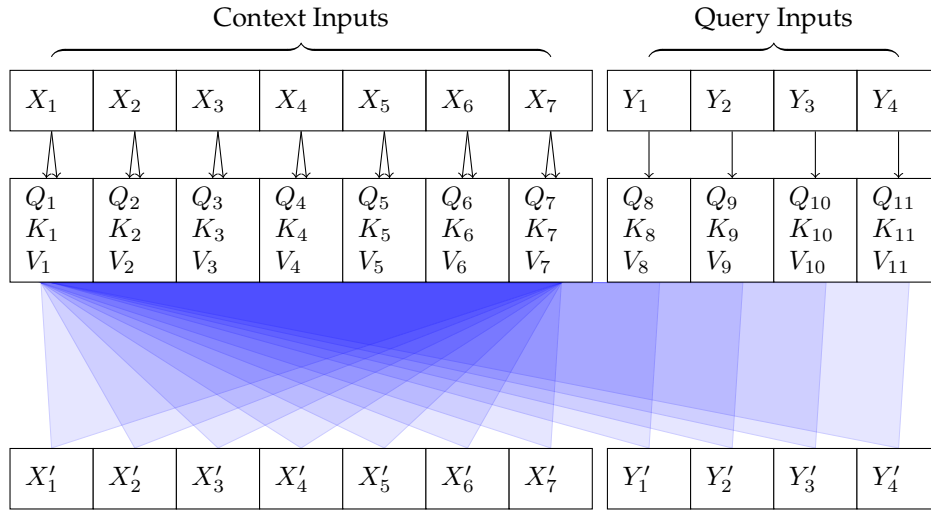


Figure 3.7: Unified self- and cross-attention, as in [5].

Bi- and uni-directional attention can be performed with the same attention layers with careful masking, allowing the same model to be trained on any mixture of pre-training tasks.

The “encoder” outputs X' are computed with full self-attention, and the “decoder” outputs Y' are computed with full attention with respect to X and causal attention with respect to Y .

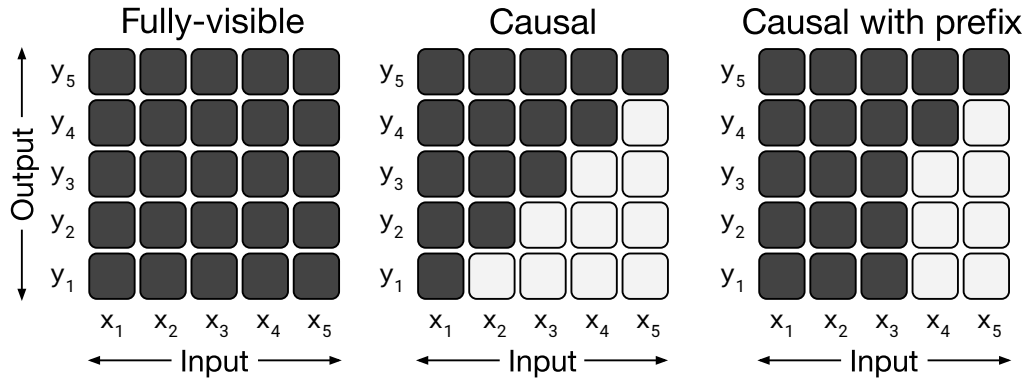


Figure 3.8: Masking in unified attention layers [5, 11]. The bi- and uni-directional attention can be performed within a single attention layer.

3.2.4 Unified attention models

As we see in 3.7, there are actually two ways we might include information from the encoded sequence into the decoded sequence. In the first, which is used in most encoder/decoder models including the original, all the layers of the encoder are computed fully, and the final encoded sequence is provided to the decoder at all layers via cross-attention. In the second, the encoder is separated from the decoder only by correct application of masking in the attention layers, and is only computed up to the layer which is being decoded at.

The different layouts that can be used are shown in Figure 3.8.

The first method is more efficient because the attention matrices of the encoder and the decoder are split, using $O(N^2 + M^2)$ memory rather than $O(N^2M^2)$. However, the second method is conceptually simpler and is better when the trained model will be used for fine-tuning/transfer learning [11], because it uses the decoder to attend to the encoded sequence at any layer, and not just the final layer. This is useful for tasks such as image

captioning, where the decoder may want to attend to the encoded image at multiple layers, and not just the final layer.

Transformer models trained on a mix of pre-training tasks using this architecture are known as T5 models (Text-To-Text Transfer Transformers) [11].

3.2.5 Comparison of architectures

Decoder-only models

3.3 Pretraining tasks

A *pre-training task* is a self-supervised task used to train a transformer. There are two main pre-training methods for transformer models which I have already mentioned: *Masked Sequence Modeling* and *Auto-regressive Sequence Modeling* (ASM).

In masked sequence modeling (MSM), the model is trained to reconstruct the original sequence, but with some of the inputs masked out. An example of this for a language model is shown in Figure 3.9.

The inputs to the transformer models are typically sequences of discrete tokens, which each have a corresponding entry in a learned codebook of latent vectors. Masking out an input in this context means adding an additional token and learned embedding “<mask>” to the codebook, which is used to represent the masked out inputs. As we can see in the figure, the masked positions still have embeddings of the positional information, which is necessary because a transformer model has no implicit order in-

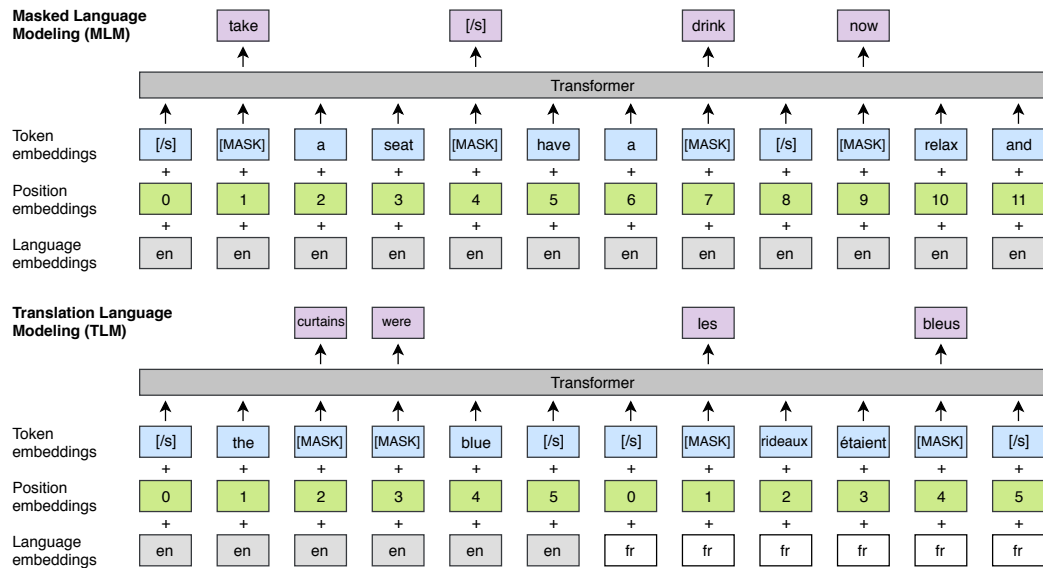


Figure 3.9: Cross-lingual masked language model pretraining, from [3].

Top: The standard MLM pretraining task. A sequence of text tokens are given as input, some are masked and the model is trained to predict the masked tokens.

Bottom: The cross-lingual MLM pretraining task. A sequence of text tokens from two different languages are given as input, some tokens are masked and the model is trained to predict the masked tokens. The model is trained on a mixture of monolingual and cross-lingual examples.

formation available to it.

The second main pre-training task is auto-regressive sequence modeling, where the model is trained to predict the next token in a sequence, given the previous tokens. For this pre-training task the model must have a causal mask applied to its attention layers, so that it can only attend to previous tokens, as in Figure 3.4.

In the next chapter (Chapter 4) I experiment with a new pre-training task, which is a variation on auto-regressive sequence modeling, which I call *arbitrary-order autoregressive pretraining*, which I will later apply to hand motion prediction.

Chapter 4

Sampling Sequences In Any Order

In this chapter I investigate learning auto-regressive transformers to generate pixels of the MNIST dataset. By changing the model architecture and training procedure, I show that we can learn to generate pixels in any order, including dynamically choosing the order of the pixels while sampling. In effect, we learn a model that can be used as a Gaussian process on this dataset.

In particular, I compare the following sampling orders with my model:

- Left-to-right, top-to-bottom
- Random
- Highest-entropy-first
- Lowest-entropy-first

Contrary to my hypothesis that a lowest-entropy-first sampling order would result in the best samples, I find that this biases the samples towards images with large amounts of empty background, such as images of 1s. Correspondingly, a highest-entropy-first sampling order biases the samples towards images of 8s and 9s.

4.1 Autoregressive models

When modeling data, it is often useful to have a *probabilistic* model of the data, rather than a deterministic model. This allows us to quantify uncertainty about the model outputs, sample from the model, and avoid problems with multi-modal distributions.

However, when the data we are modeling is high-dimensional, probabilistic models often become computationally expensive or intractable. For example, if we are modeling a sequence of N observations, and each observation is a D -dimensional vector, then even a simple gaussian distribution over the sequence has $DN + (DN)^2$ parameters (DN means, plus a $DN \times DN$ covariance matrix). In addition, it is not natural to model sequences of arbitrary length with this approach.

4.1.1 Distributions over high-dimensional data

There are a few main approaches to address this problem, which I show below:

- Discretization
- Independence assumptions

- Auto-regressive factorization

We can discretize the data, and model it as a sequence of discrete variables. For example, we can use a clustering algorithm to learn a series of K points within our $(DN)^2$ high-dimensional space, and then form a discrete distribution over these points. A discrete distribution over K points has $K - 1$ free parameters, so this approach reduces the number of parameters from $(DN)^2$ to $K - 1$, and makes sampling and inference more efficient. However the discretization changes the domain of the data, which may not always be useful.

Independence assumptions are another way to reduce the number of parameters. For example, we can assume that the observations are independent, and model them as a product of N independent distributions. For a Gaussian, this means fixing parts of the co-variance matrix, often having only a single variance parameter. A single variance parameter reduces the number of parameters from $(DN)^2$ to DN , but it also makes the model less expressive. For sequence data, this assumption is almost never valid along the sequence dimension, so this approach is not very useful.

When we use an auto-regressive model to predict sequences, we usually choose some fixed order. For data with a temporal dimension, this is usually first-to-last, which is usually natural because the real process that generated the data had a causal structure in the temporal dimension.

However, a simple ordering may not always be the best. For example, when predicting data with spatial dimensions such as pixels in images, the best ordering may not be obvious. Also, some data may have very long sequences and latency requirements (such as character animation data), where it is expensive to generate the data in order, and we would instead

like to generate only particular parts of the sequence.

4.2 Changing The Transformer Input

Transformers have their input provided as (position, content) pairs, or more generally, as some kind of token embedding which is unique among the input set, such as special “BEGIN” or “CLASS” tokens.

As we saw in chapter 3 both the attention layers and feed-forward layers are invariant to permutations of the input sequence. Additionally, there is no requirement that the input set be contiguous in the sequence dimension - there can be (potentially large) gaps, with no change to the structure of the model (however, the model must be trained for the particular problem still).

As a result of being invariant to permutations, and working with non-contiguous sequences, we can present many different kinds of sequence prediction tasks to a transformer that we could not easily present to other models.

- A Recurrent Neural Network (RNN) can be given sequences with gaps, but is not invariant to the order of the “previous token” conditioning data, which must be incorporated first in some particular order.
- A convolutional or dense neural network applied to the input including the sequence dims (ie. not “pointwise”) is neither invariant to the order, nor can be applied to sequences with gaps.

In the next section I describe some tasks we might want to perform with these unique abilities of a transformer.

4.3 Tasks

In this section I discuss various tasks that utilize the ability of a transformer to predict sequences in any order.

4.3.1 Arbitrary order decoder-only transformer using input triples

Let us examine first decoder-only transformers. these predict the next input from the previous input, conditioned on the rest of the sequence via their attention layers.

An input examples in the training data for these models is typically a set as follows:

$$\{\dots, (i_{n-1}, x_{n-1}, i_n), (i_n, x_n, i_{n+1}), (i_{n+1}, x_{n+1}, i_{n+2}), \dots\}$$

Where i represents the position of a token, and x represents the value of a token.

However, if we instead construct an input sequence in the following way, we can train the model such that given any conditioning sequence (previous tokens), we can ask the model to predict a specific next token.

We do this by providing the input as (input position, input value, target position] triples instead of [input position, input value] pairs (in which the target is implicit)

todo: Make and include a figure for this. Should show difference between input as triples [input position, input value, target position] and pairs [input position, input value]

todo: I haven't found any works that do this - but I still think there probably are some out there

An example input in the training data is a set as follows:

$$\{\dots, (i_{n-1}, x_{n-1}, i_n), (i_n, x_n, i_{n+1}), (i_{n+1}, x_{n+1}, i_{n+2}), \dots\}$$

todo: Improve the above formatting

If our sequence is presented in contiguous-forward-only ordering, i_n is always paired with i_{n+1} and we do not introduce any new information. However, we can create a sequence with an arbitrary order during training, so the model learns to utilize this information.

Then, at inference time we can choose any position i_{n+1} the model should predict next, by constructing the following triple (i_n, x_n, i_{n+1}) and appending it to the rest of the previous tokens.

4.3.2 Queries

todo: Write the introduction to this architecture variant in Chapter 3

If we now examine the case of pure-query-decoder transformers, which I introduced in ???. These are encoder-decoder transformers and are trained

with a different task:

$$\begin{aligned}\text{input} &= (\{ \dots, (i_{n-1}, x_{n-1}), (i_n, x_n) \}, i_{n+1}) \\ \text{output} &= (x_{n+1})\end{aligned}$$

todo: Improve the above formatting

The previous tokens are provided as a set of pairs to the encoder. Importantly, the decoder is run independently for every input/output pair.

4.4 Hypothesis

Using the above two methods “triples” and “pure-query-decoder” models, I investigate a hypothesis about selecting a better sampling order on a toy dataset.

The hypothesis is as follows.

Assume that using the above methods, we can choose a dynamic ordering in which to sample a sequence at inference time. In particular, one of the ways we can do this is by evaluating the *entropy* of all candidate positions, then sampling from the one with either the lowest or the highest entropy.

I hypothesize that when auto-regressively sampling pixels to produce MNIST images, using a “lowest-entropy-first” ordering, will produce visually better results than a “highest-entropy-first” ordering.

4.5 Method

todo: Method: Details of experiments.

4.5.1

4.5.2 Baseline: Fixed-order sequence prediction

todo: Subsection on training forward-only task

todo: Subsection on Training with arbitrary-order methods, but in forward-only mode (sanity check - should have similar, perhaps somewhat worse results)

todo: Subsection on Ablation / Hyper parameter tuning

todo: Subsection on Training (input position, input, target position)

todo: Subsection on Training query-decoder

4.6 Results

todo: Section MNIST Results / Comparison

Chapter 5

Angles, Joints and Hands

Before I discuss my experiments training a model on hand motion data in the next chapter, I will first introduce some background related to modeling human hands, so we can understand many of the implementation choices I made.

5.1 Parameterizing Hand Configurations

To a first approximation, the human hand has 23 degrees of freedom.

As we can see in Figure 5.1,

Considering that different combinations of joint angles can have very different meanings, animators have a lot of work to do when bringing a digital character to life. Small mistakes can cause a character to look unnatural, but in many scenes getting the hands perfect goes unnoticed.

It is not so straightforward to assign an angle to each of those 23 degrees of freedom. There are a variety of different parameterizations of the joints we might choose from, and additionally we have the option of placing con-

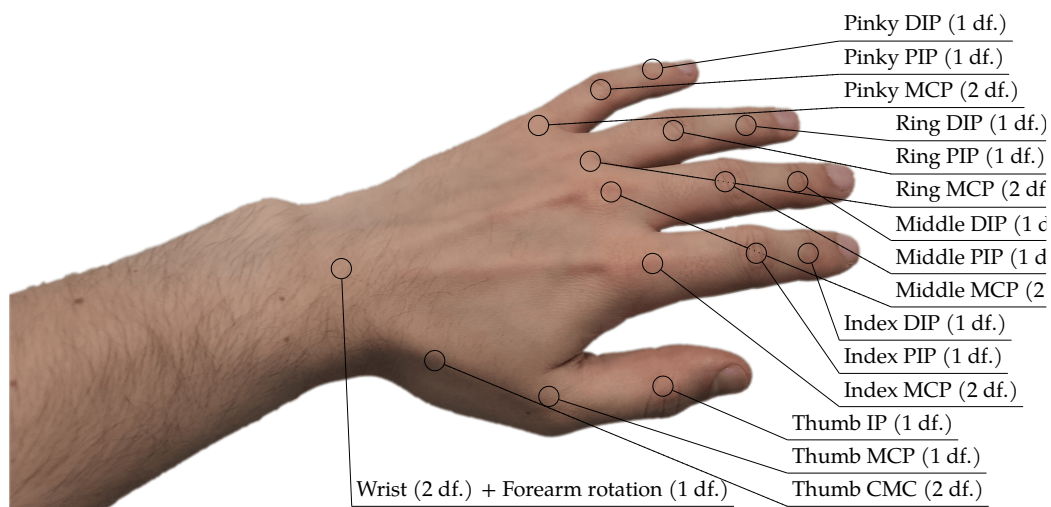


Figure 5.1: Each hand has 16 joints - three per digit, plus the wrist.

The wrist and the first joint on each digit (the metacarpophalangeal (MCP) joints) can rotate on two axes, and so each have 2 degrees of freedom. The rest (the proximal- and distal-interphalangeal (PIP & DIP) joints) can only rotate on one axis, and have 1.

This naive counting gives 22 degrees of freedom. In addition, we usually consider rotation of the forearm (about the longitudinal axis) to be part of the hand, modeling it as a third degree of freedom of the wrist, which brings the total to 23 degrees of freedom.

straints on the range of values that the angles can take. In this section we will discuss some of the most common parameterizations, and the pros and cons of each.

5.1.1 Euler Angles

Euler angles are the most common parameterization for human joints. They are also the simplest to understand. Each joint is assigned three angles, one for each axis of rotation. The axes are usually chosen to be the x , y , and z axes of the coordinate system, but they can be chosen to be any three axes that are orthogonal to each other. For example, the axes could be chosen to be the axes of the joint itself, or the axes relative to the parent joint. The order in which the rotations are applied is also important. The most common order is to apply the rotations in the order z, y, x , but they can be applied in any order.

This parameterization has the topology $S^1 \times S^1 \times S^1 \cong T^3$, where S^1 is the circle of angles. However, the space of rotations $SO(3)$ itself has topology S^3 , so the Euler angle parameterization cannot be perfect. In particular, there must be singularities.

The principle advantage of the Euler angle parameterization is that it is easy to understand and implement. The principle disadvantages are due to the singularities:

1. there may be multiple sets of Euler angles that correspond to the same rotation
2. as a result, we cannot easily interpolate between two configurations that are close together but have different Euler angles

For example, if we rotate a joint by 90° about the x axis, and then by 90° about the y axis, we will get the same rotation as if we had rotated by 90° about the y axis, and then by 90° about the x axis.

todo: Euler angle figure

If we tried to interpolate between these two configurations, we would get a rotation that is not the same as either of the two configurations. This occurs because the Euler angle parameterization is not a smooth function.

5.1.2 Axis-Angle

An alternative parameterization is to use an axis-angle representation. In this representation, each joint is assigned a unit vector and an angle. The unit vector specifies the axis of rotation, and the angle specifies the amount of rotation. The axis-angle representation has the topology $S^2 \times S^1$, where S^2 is the sphere of unit vectors. Because topology still does not match the space of rotations $SO(3)$, there are still singularities. However

5.2 Loss Functions for learning angles

In Chapter 2 I introduced some loss functions such as the Mean Squared Error (MSE), which predicts the posterior expectation $\mathbb{E}(y|x)$. In this section I will discuss some other loss functions, that are commonly used for learning data which sits on .

A variant of this is the *angular* mean-squared-error, which I will use

later on in Chapter 6.

$$L_{\theta\text{-MSE}}: \mathbb{R}^{N \times D} \times \mathbb{R}^{N \times D} \rightarrow \mathbb{R}$$

$$L_{\theta\text{-MSE}}(y, \hat{y})_{ni} \stackrel{\text{def}}{=} \frac{1}{N} \sum_n \left[\sum_i (\sin y_{ni} - \sin \hat{y}_{ni})^2 + (\cos y_{ni} - \cos \hat{y}_{ni})^2 \right] \quad (5.1)$$

Here y and \hat{y} are vectors of angles, and the error is defined in terms of the squared arc length between their corresponding components. This loss function is equivalent to maximising the likelihood of a von Mises distribution, the derivation for which can be found in Chapter 5. This is useful when we want to model angles, for example when we want to predict the orientation of a hand, which we will do in Chapter 6.

todo: Derivation of θ -MSE minimizing von-mises distribution

Chapter 6

Hand Motion Model

I experimented with a variety of different model architectures,

Loss Value mean across whole dataset	0.07275154
Total number of frames in dataset	467372
Total length of dataset	4h 19m 39s

6.1 Predicting the next frame

6.2 Learning a probabilistic model

Chapter 7

Conclusions

Bibliography

- [1] Alexei Baevski et al. “wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations”. In: *CoRR* abs/2006.11477 (2020). arXiv: 2006.11477. URL: <https://arxiv.org/abs/2006.11477>.
- [2] Tom B. Brown et al. “Language Models are Few-Shot Learners”. In: *CoRR* abs/2005.14165 (2020). arXiv: 2005.14165. URL: <https://arxiv.org/abs/2005.14165>.
- [3] Alexis Conneau and Guillaume Lample. “Cross-lingual language model pretraining”. In: *Advances in neural information processing systems* 32 (2019). arXiv: 1901.07291. URL: <http://arxiv.org/abs/1901.07291>.
- [4] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: 2018. URL: <https://arxiv.org/abs/1810.04805>.
- [5] Li Dong et al. “Unified language model pre-training for natural language understanding and generation”. In: *Advances in Neural Information Processing Systems* 32 (2019). arXiv: 1905.03197. URL: <http://arxiv.org/abs/1905.03197>.

- [6] Patrick Esser, Robin Rombach, and Bjorn Ommer. “Taming transformers for high-resolution image synthesis”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 12873–12883. URL: http://openaccess.thecvf.com/content/CVPR2021/papers/Esser_Taming_Transformers_for_High-Resolution_Image_Synthesis_CVPR_2021_paper.pdf.
- [7] Charles R. Harris et al. “Array programming with NumPy”. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2). URL: <https://doi.org/10.1038/s41586-020-2649-2>.
- [8] Mike Lewis et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 7871–7880. arXiv: [1910.13461](https://arxiv.org/abs/1910.13461). URL: <http://arxiv.org/abs/1910.13461>.
- [9] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from [tensorflow.org](https://www.tensorflow.org). 2015. URL: <https://www.tensorflow.org/>.
- [10] Alec Radford et al. “Language Models are Unsupervised Multitask Learners”. In: (2019). URL: <https://github.com/openai/gpt-2>.
- [11] Colin Raffel et al. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: <http://jmlr.org/papers/v21/20-074.html>.

- [12] Noam Shazeer. “Fast transformer decoding: One write-head is all you need”. In: *arXiv preprint arXiv:1911.02150* (2019). arXiv: 1911 . 02150. URL: <http://arxiv.org/abs/1911.02150>.
- [13] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017). arXiv: 1706 . 03762. URL: <http://arxiv.org/abs/1706.03762>.
- [14] Zhenda Xie et al. “SimMIM: A Simple Framework for Masked Image Modeling”. In: *CoRR* abs/2111.09886 (2021). arXiv: 2111 . 09886. URL: <https://arxiv.org/abs/2111.09886>.
- [15] Jiahui Yu et al. *Scaling Autoregressive Models for Content-Rich Text-to-Image Generation*. 2022. DOI: 10 . 48550 / ARXIV . 2206 . 10789. URL: <https://arxiv.org/abs/2206.10789>.