Deep learning methods for style extraction and transfer

Omar Mohammed

Contents

List of Figures

List of Tables

Part I

Introduction and Problem Description

Chapter 1

Introduction

Contents

In one sentence, my PhD focus on the extraction, characterization and *transfer* of *styles* or *personas*, isolated from a task, using deep neural networks.

1.1 What is a style?

Style is generically defined in Merriam-Webster dictionary (figure 1.1). In general, it is defined as *the manner of doing things* (Gallaher, 1992). To get more sense of what style is, it is better to give some examples first, to get an idea about what we are dealing with.

- When we say the word "seriously?". Depending on the manner we say it, it will carry different meaning (sarcastic or surprise for example). One word, two different manners to say it.
- Handwriting: You can write down the same the letter (the task), but with multiple typefaces (the style) (figure 1.2).

style noun

\ ˈstī(-ə)l \

Definition of *style* (Entry 1 of 3)

1 : DESIGNATION, TITLE

2 **a** : a distinctive manner of expression (as in writing or speech)
// writes with more attention to *style* than to content
// the flowery *style* of 18th century prose

b : a distinctive manner or custom of behaving or conducting oneself
// the formal *style* of the court
// his *style* is abrasive
also : a particular mode of living
// in high *style*

c : a particular manner or technique by which something is done, created, or performed
// a unique *style* of horseback riding
// the classical *style* of dance

3 **a** : STYLUS

b : GNOMON sense 1b

c : the filiform usually elongated part of the pistil bearing a stigma at its apex
— see FLOWER ILLUSTRATION

d : a slender elongated process (such as a bristle) on an animal

4 : a distinctive quality, form, or type of something
// a new dress *style*
// the Greek *style* of architecture

5 **a** : the state of being popular : FASHION
// clothes that are always in *style*

b : fashionable elegance

c : beauty, grace, or ease of manner or technique
// an awkward moment she handled with *style*

6 : a convention with respect to spelling, punctuation, capitalization, and typographic arrangement and display followed in writing or printing

Figure 1.1: Definition of style in Merriam-Webster dictionary

- A movie setting: the script is provided to the actor. There are, however, many ways for the actor to perform what is written in the script, in order to convey different messages/experiences to the audience.
- Clothing and fashion: different groups of people have different general style lines – depending on the region, ethnicity ... etc –. Within each group, we can see people having diverse styles within this general defining style.

Figure 1.2: Multiple styles (different typefaces) for the same letters.

In these two examples, we see a basic structure:

- A fixed part: the word to be said, or the letter identity. We will call this the *content*.
- A variable part: the manner we say the word, or the manner we write the letter. We will call this the *style*.
- Together, a style and a content forms a *task*.

The mention of styles is quite a lot in the literature in multiple domains, for example:

- **Speech synthesis**: (Tachibana et al., 2004) defines speaking style in a high-level manner, in terms of emotions expressed by the speaker, like 'joy', 'sad', or the interpolation between them, when reading a text. (Wang et al., 2018) looks at the speech style in a more detailed manner, considering different aspects of speech prosody, like the paralinguistic information, intonation and stress.
- **Car driving**: there are multiple ways to categorize the different driving styles. It can be based on the safety aspect (Johnson and Trivedi, 2011), the aggressiveness

of the maneuver (Dörr et al., 2014; Xu et al., 2015b), the impact on fuel consumption (Manzoni et al., 2010; Neubauer and Wood, 2013). Many other identification basis for driving styles are summarized nicely in (Martinez et al., 2017).

- **Handwriting**: handwriting can be offline (the final image of the letter) or online (recording the movement of the pen/drawing tool). Depending on which one considered, the style profile can change. Figure 1.2 is an example of different offline styles (the typefaces). But when we consider the online aspect of the drawing, we can see different aspects, like in figure 1.3, where we see that the same drawing can be in clockwise or counterclockwise direction (we will expand more on this point in chapter 4.2.4).

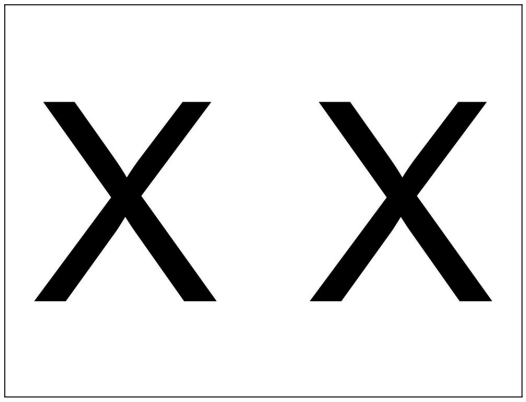
One thing we would like to highlight here: that styles are not thing we all agree on. It can be hierarchical as well. In the clothing example earlier, people are affected by the style group they belong to, but within this cluster, people have diverse individual styles. The same happens in handwriting: education and culture affects the style cluster people belong to, but we can observe a wide range of individual styles in each cluster.

Another thing to highlight is that there is no one definition for styles. It depends on the aspect of interest that we want to observe and study – in car driving, safety and fuel consumption are two areas of interest, leading to different type of styles –. This leads to an important characteristic of *styles*, that it is an ill-defined concept. We know that *styles* are rich in information and important in communication between humans. They are needed in order to convey meaning. As noted in (Taylor, 2009) – in the context of speech synthesis –, a proper rendering of styles affects the overall perception. However, we can not completely remove the ambiguity in this definition.

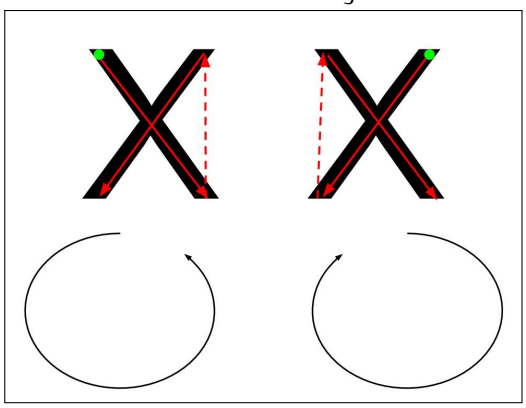
1.2 What is the objective of this project?

Our long-term objective is to enable our humanoid iCub robot Nina 1.4, to have a exhibit personalized behavior suitable for the person interacting with it. This will enhance the user experience, and will allow for a more natural interaction with the robot. It is shown that a robot which exhibit a personalized behavior is more likable acceptable by humans, and perceived as an intelligent entity (Churamani et al., 2017). Humans have different preferences when interacting with each other, or interacting with the robot, and taking them into account does improve the quality of interaction, and the potential of success for the task (Kashi and Levy-Tzedek, 2018).

At the moment, we successfully used machine learning approaches in order to build models of human-robot interaction (Bailly et al., 2018; Mihoub et al., 2016; Nguyen et al., 2017). However, when using these models to generate behaviors, this



1 Offline drawing



2 Online drawing

Figure 1.3: Example for style in case of online handwriting. Although both examples looks the same when we look at it in the offline mode (the final drawing), they are quite different when we consider the online aspect (the dynamics of the pen when drawing them). Although not illustrated here, it is important to note that some aspects of the online drawing dynamics can be deduced from the offline drawing – in other words, the dynamics can affect the end result – (Diaz et al., 2017). In this case, the starting point and the direction of drawing (clockwise or counterclockwise) are different. The solid line indicate a stroke, and the dotted line indicate an air stroke (the transition of the pen in the air between two strokes). The green dot is the starting point.

behavior usually represents an average over the learned behaviors (which is expected). The goal is to learn models of styles, and use it to bias the models of interaction that we have, in order to generate more personalized behaviors. We want the robot to adapt to the human partner on different levels, and not just act in a reactive manner to the human actions. It has been shown that a suitable cognitive model for human-robot interaction takes into account long term styles – for example, the person age, gender and if he/she is shy or not –, and short term decision making – in talking with the human for example – (Bailly et al., 2010; Thórisson, 2002). The ability to identify, extract and use the human style traits will enable biasing the robot model to adapt for the human partner.

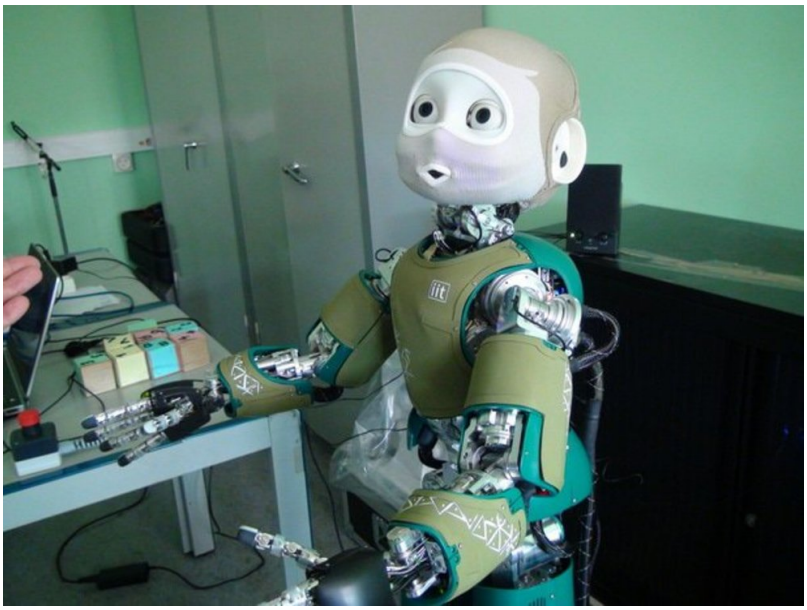


Figure 1.4: The iCub humanoid robot (Nina).

1.3 Why Handwriting?

As mentioned earlier, the final objective of this project is to extract and transfer styles in the context of human-robot interaction. What this has to do with handwriting?

The usage of deep learning in HRI is still in its infancy, mainly because of the lack of datasets. The issue is not collecting the data (there are many small open-source datasets available online), but having a unified set of objective and platforms for HRI, which is not an easy challenge. We envision that this problem will be resolved, as there is a growing interest in the community to address it.

Thus come handwriting. We use it as a proxy platform to understand, build and test different approaches to use deep learning. It has many advantages to make it a good proxy, including:

- Availability of datasets: several datasets already exist, with large quantities of data.
- Diverse of tasks and styles: there are many tasks (letters) in handwriting, ranging from simple (like letter 'C') to complex (like letter 'E'), and the writers exhibit a diverse set of styles on the different tasks, making handwriting a good candidate to explore the problem of styles.
- Several style aspects are accessible to investigate visually, making it more accessible for in-depth analysis, and getting insight on how the model behaves.
- Several datasets provide information about the writers, like the age, handiness, gender and the origin. This data is interesting in some aspects of the style problem.
- We have a clear idea about the content of each task (the identify of the letter or the shape). Usually, the task is presented to us with the content and the style mixed together. How to disentangle the content from the style is an open question, and the fact the styles is an ill-defined problem makes it more ambiguous. With the assumption that we know the content, we can focus our effort on the styles problem[1].

While the subjects are interacting with the environment (the pen, tablet ... etc) – which is observed and recorded –, there are no human interaction aspects in this domain of data, which is a disadvantage

1.4 What is transfer learning? and why do we need it?

We will discuss transfer learning in more detail in chapter 5, but for now, we want to motivate having this as one of the PhD objectives.

Transfer of knowledge deals with the problem of leveraging the knowledge learned from one task, to accelerate/improve the learning of a new task. This is a skill human do naturally, for example, if you learn Mathematics, and you want to learn physics, you can easily leverage the knowledge of Mathematics to bootstrap your learning of physics. This, intuitive as it seems, is not straightforward for machine learning

[1] We will argue later that the task identity is not necessarily a good representation for the task content.

models. A change in the distribution of the input to the model leads to significant degradation in the performance of the model (Shimodaira, 2000).

Transfer learning is thus a field of machine learning, concerned with developing algorithms and procedures, to enable the transfer of knowledge between different tasks. So many techniques are available for transfer learning, but there is always a common assumption, that the transfer has the potential of success if the tasks are related (i.e., if there is common knowledge between the tasks), otherwise, a transfer learning can at best lead to no improvement, or even reduce the performance of the new model (Weiss et al., 2016).

Why do we need transfer learning? We do not always have the availability of large datasets on the tasks that we want. In many cases, the acquisition and/or the annotation of large dataset can be prohibitively expensive. A simple relevant example here is human-robot interaction: collecting large dataset on each scenario we want to learn is simply unfeasible, due to the huge time it takes, and the limitations/cost of the robot hardware. Also, no data augmentation techniques does exist in the literature for HRI, thus synthesizing extra data is not possible. Thus, we need to be able to transfer the knowledge from the task where we have large amount of data, to a relevant task where we do not have this advantage (figure 1.5).

What we want to achieve in this thesis is to transfer the styles between different tasks. We hypothesize that, when the tasks are relevant (but different tasks, with different contents), that humans share leverage styles between the different tasks. In case of handwriting for example, we can reuse the strokes that we learn in the uppercase letters in order to learn the lowercase letters and digits. We will test this hypothesis in details in chapter 5.

1.5 If we want transfer learning, why extracting styles?

It is important not to forget our mission while being consumed by the shiny light of machine learning. It is easy to fall into the trap of "getting the extra 0.1% accuracy", figure 1.6. Extracting styles enables us to have an idea of what the model actually learned (i.e., it makes the model more interpretable). But why do we need interpretability? Many reasons:

- Debugging: Neural networks are notoriously known for being black box models. Extracting styles gives us some indication on what the system learned, and whether it learned relevant information.
- Discovery of new things: in the era of big data, trying to reason on the data directly is no longer feasible. Instead, a good approach is to reason on a model

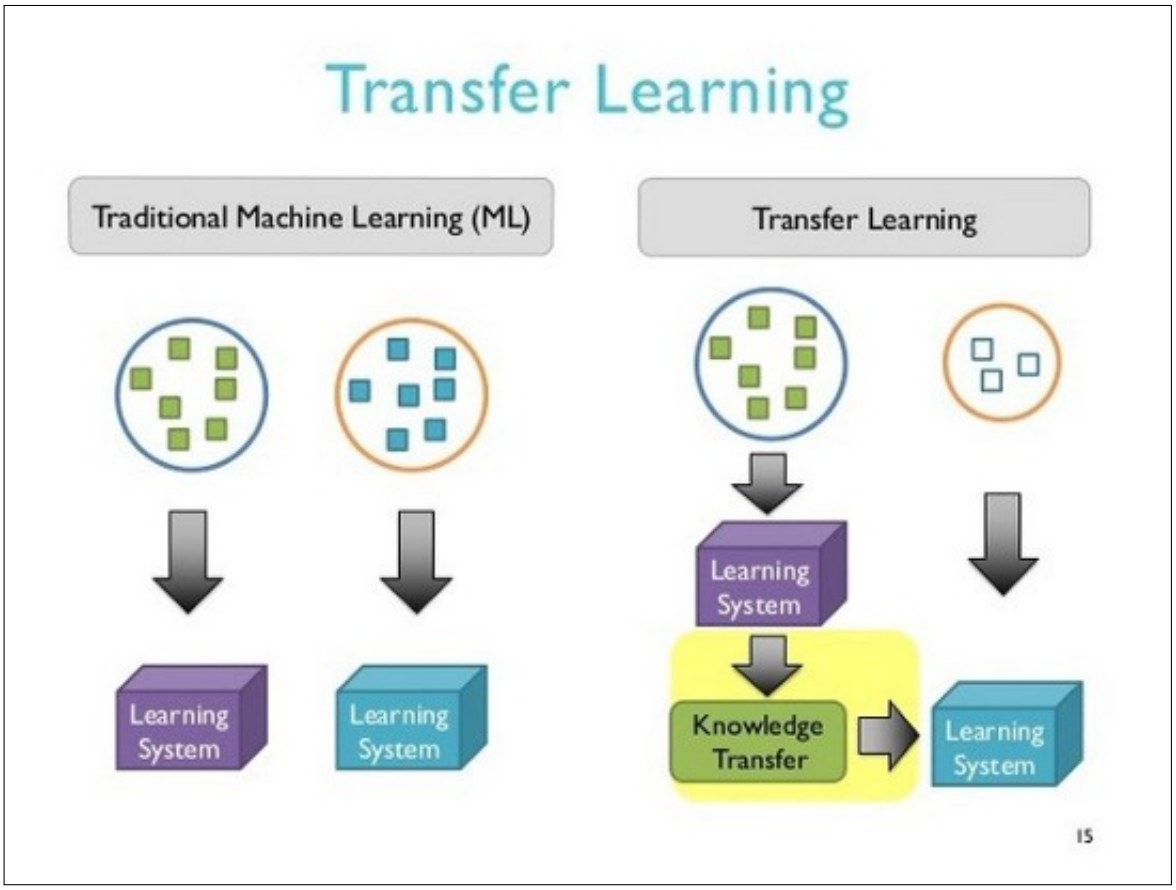


Figure 1.5: Illustration for a practical case of using transfer learning, where we have insufficient data on the new task, and we want to leverage the knowledge learned from another relevant task in order to learn the new task[2].

that fits these data (i.e., the model compresses the data into fewer dimensions). We would like to use this model to reason and discover new things about the data. After all, **the goal of science** is to gain knowledge about the world, and an understanding of how it works. One thing we are interested in is to illuminate the workspace of interaction: what are the limits of possible actions, expected reactions and end results. These kind of questions will benefit a lot understanding what the model actually learned.

- Understanding why: in some cases (e.g., in case of unexpected events), it is human curiosity to understand the reasoning behind the different decisions. We would like to get insight on why the model led to that particular decision.

- Developing safety measures: in many applications (like when dealing with the robots), it is important to be sure that the robot is a 100% safe for the human. Understanding what the model learned can give us insight on the shortcomings of the model, allowing us to fix or improve.

- Social acceptance: humans always tries to attribute beliefs, intentions, personality traits and desires to different objects (Heider and Simmel, 1944). An interpretable machine will reinforce these sensations in humans.

- Improve social interactions: when the robot can explain itself and its perception of the world, it creates a common understanding with the human. This allows the humans to build a mental model for what the robot is actually trying to do, thus, building trust between humans and the robot.

For further details, we strongly recommend the book (Molnar, 2019) on the topic of machine learning interpretability.

1.6 Contributions of this PhD

In this manuscript, we discuss the different contribution of this PhD, addressing different aspects of the styles:

1. Propose a manner of thinking about styles: traditionally, a lot of work has been done in order to manually extract and annotate styles, and deal with styles in terms of predictability (using the extracting styles in a regression/classification problem). In this PhD, we propose to implicitly evaluate styles by observing the generative aspects of the model (letting the model generates behaviors, and trying to evaluate the distance between these behaviors and the target/ground truth ones). This is discussed in chapter 3.

Figure 1.6: Source: `https://xkcd.com/`

2. Propose and build benchmarks and evaluation metrics, and ground those metrics, in order to compare and evaluate future style extraction methods. This is discussed in chapter 3.

3. Propose a generic framework to study styles, using a conditioned-autoencoder. We evaluate this framework in its basic form against the benchmarks. We further validate this framework by extracting verbose styles from it, including ones that are not known from the literature. This is discussed in chapter 4.

4. Last, we address the problem of style transfer. We show how to use our proposed framework in order to transfer styles. We perform extensive experiments in a lot of combinations of tasks, on two different datasets. We also enhance and expand on the evaluation metrics we use in order to quantify the quality of transfer. This is discussed in chapter 5.

1.7 Thesis outlines

We start in chapter 2 by explaining the datasets used in this PhD, and explore their different characteristics, and the preprocessing performed on them. In chapter 3, we discuss first the different aspect of deep learning that we are using in this PhD. We then discuss the different benchmarks and evaluation metrics we propose, and how can we ground them.

Once this step is done, it paves the way to discuss our proposed framework to study styles in chapter 4. We start first by presenting the relevant literature, then move on to the experimental part, where we show the performance of the proposed framework relative to the benchmarks. We conclude this chapter by a section on style extraction, where we shade some light on what the model actually learned.

With the elements in place, we can explore the topic of style transfer, in chapter 5. As usual, we study with literature over transfer learning, followed by the proposed experiments in order address our hypotheses. We perform a wide range of experiments, to solidify our conclusions about style transfer. We also discuss another possibility to interpret what the model actually learned in section 5.4.

We conclude the manuscript by a general discussion over this thesis (chapter 6), the difficulties it faced, and the possible future research directions based on the results. We believe that this thesis answered multiple questions, but it also created more questions and research interests for further investigation. We then make a short summary and conclusions about the work done in this thesis in chapter 7.

Chapter 2

Datasets

Contents

In order to test the different hypotheses and ideas presented in the introduction chapter, we need to select suitable domain to carry out the experiments, which determined by the data. The choice of the data is a selection between different trade-offs: relevance (i.e., it contains the relevant information to perform the study) VS readiness of the data (i.e., is it already available? and how expensive it is to acquire the data?), simplicity VS realism (simple data has less noise and less irrelevant patterns, but the ability to handle realistic data provides stronger support for the hypothesis), and the amount of data available.

It is traditional in the machine learning community to use two or more datasets to address the questions. This way, it is possible to detect issues like having a very specific method that work only in a limited context[1], and to avoid the effect of unknown contributing variables.

We settled on the domain of handwriting and drawing. In this chapter we present two datasets: online English letters handwriting dataset, *IRONOFF*, and online sketch drawing dataset, *Quick Draw!*. We present general information and exploratory

[1] This is not wrong in itself, it depends on the objective of the work. In our case, we want to show an indication that our methods can generalize.

statistics about both datasets, and discuss the categories/tasks in each dataset, and argue why both datasets are suitable for this study.

Points addressed in this chapter

- Present *IRONOFF* handwriting dataset.
- Present *QuickDraw!* sketch drawing dataset.
- Motivate the suitability of these datasets for this study.

2.1 Online Handwriting – *IRONOFF*

IRONOFF (Viard-Gaudin et al., 1999) is a cursive handwriting dataset provides us with isolated letters, thus allowing us to focus on the problem of styles with a reasonable complexity, and gives us the advantage that the content of the task is well known beforehand (i.e, the identity of the letter). Other cursive handwriting datasets do exist, like *IAM Handwriting Database* (Marti and Bunke, 1999). However, they use whole sentences/paragraphs, instead of individual letter, thus making the problem more complicated.

Basic information about *IRONOFF* dataset as a whole:

- Around 700 writers in total. We use the 412 writers who have written isolated letters.
- 10,685 isolated lower case letters, 10,679 isolated upper case letters, 4,086 isolated digits and 410 euro signs.
- The gender, handiness, age and nationality of the writers.
- For each writer/task (letter or digit) example, we have that example's image - with size of 167x214 pixels, and a resolution of 300 dpi -, pen movement timed sequence comprising continuous X, Y and pen pressure, and also discrete pen state. This data is sampled every 10ms at maximum, on a Wacom UltraPad A4, but the actual sampling results is not uniform[2]. Figure 2.1 shows an example for format of the the provided data.

We explored the information available about the writers in the dataset (see figure 2.2), we can see that almost all the participants are of French nationality, and majority of them are less than 30 years old. The data is almost balanced between males and females, but largely unbalanced between left and right handed people.

We then explored the handwriting examples from multiple points of view. In figure 2.3, we can see the frequency of strokes for different tasks. We can see that letters like **C** and **L** needs one stroke only, while **I** and **E** requires the most number of strokes in order to draw properly. By combining this with observation about the drawing time for different tasks (see figure 2.4) and the pausing time (see figure 2.5), we can have a good indication about the complexity of each task relative to the other tasks. The more strokes the task has, the more drawing and pausing time is needed, and the more complex the task is.

One challenging issue with this dataset however is that we have only one example for each writer-letter combination. This makes the task more difficult, because it is hard to extract a writer style using very few items (the 26 letters/writer in this case).

[2]To be treated in the preprocessing step.

```
.DATE           08 07 1998

.COMMENT        Declarations
.X_DIM             3600
.Y_DIM             3600
.X_POINTS_PER_INCH 300
.Y_POINTS_PER_INCH 300
.POINTS_PER_SECOND 100
.COORD          X Y P T

.COMMENT        Data
.WRITER_ID      sk18091962
.COUNTRY        Germany
.HAND           R
.AGE            35
.SEX            M
.SEGMENT CHARACTER ? ? "a"
.PEN_DOWN
88 104 2 0
89 103 12 10
90 101 24 21
90 100 30 31
90 99 39 40
91 97 46 50
92 95 53 62
92 89 57 74
91 86 61 85
90 83 64 97
88 80 67 109
85 78 69 121
80 80 71 132
74 88 70 144
69 101 69 156
64 114 68 167
62 120 68 179
61 127 68 191
62 131 68 203
68 132 67 214
72 130 67 224
81 123 68 234
89 116 68 244
93 113 70 256
94 111 73 272
94 112 75 285
93 119 76 297
92 123 76 308
92 130 76 320
93 136 74 331
96 138 66 340
99 142 15 350
.PEN_UP
```

Figure 2.1: An example from *IRONOFF* (letter 'a'). To the left, we have the image of the letter (i.e., offline-handwriting). To the right is an example of the format for the online-handwriting. We have the writer's ID, origin, handiness, age and gender. The sequence of pen movement to draw the letter are then given: pen state (PEN_DOWN, PEN_UP), X, Y coordinates, pen pressure, and time.

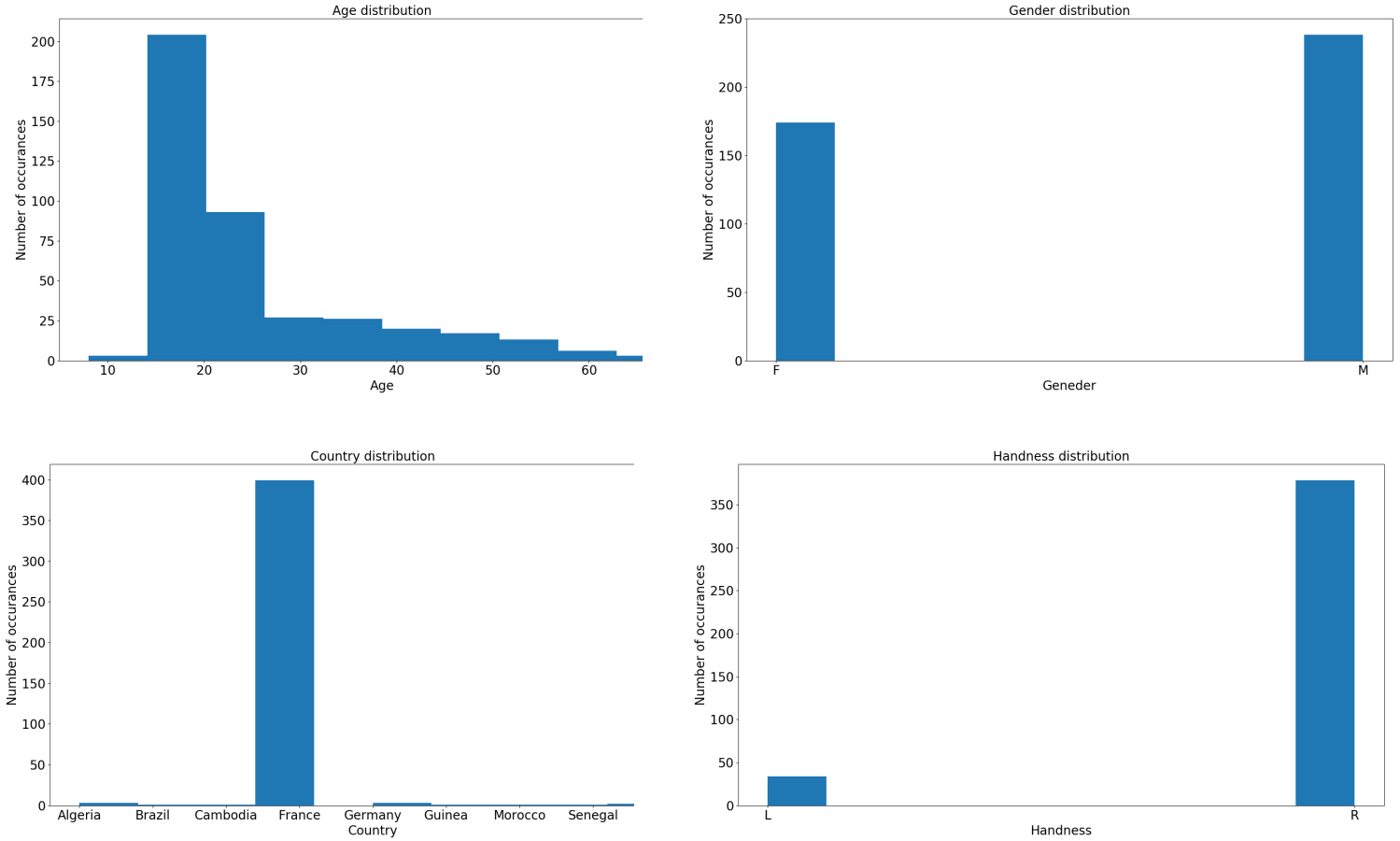


Figure 2.2: Summary statistics about the writers in *IRONOFF*: the age, gender, country and handiness.

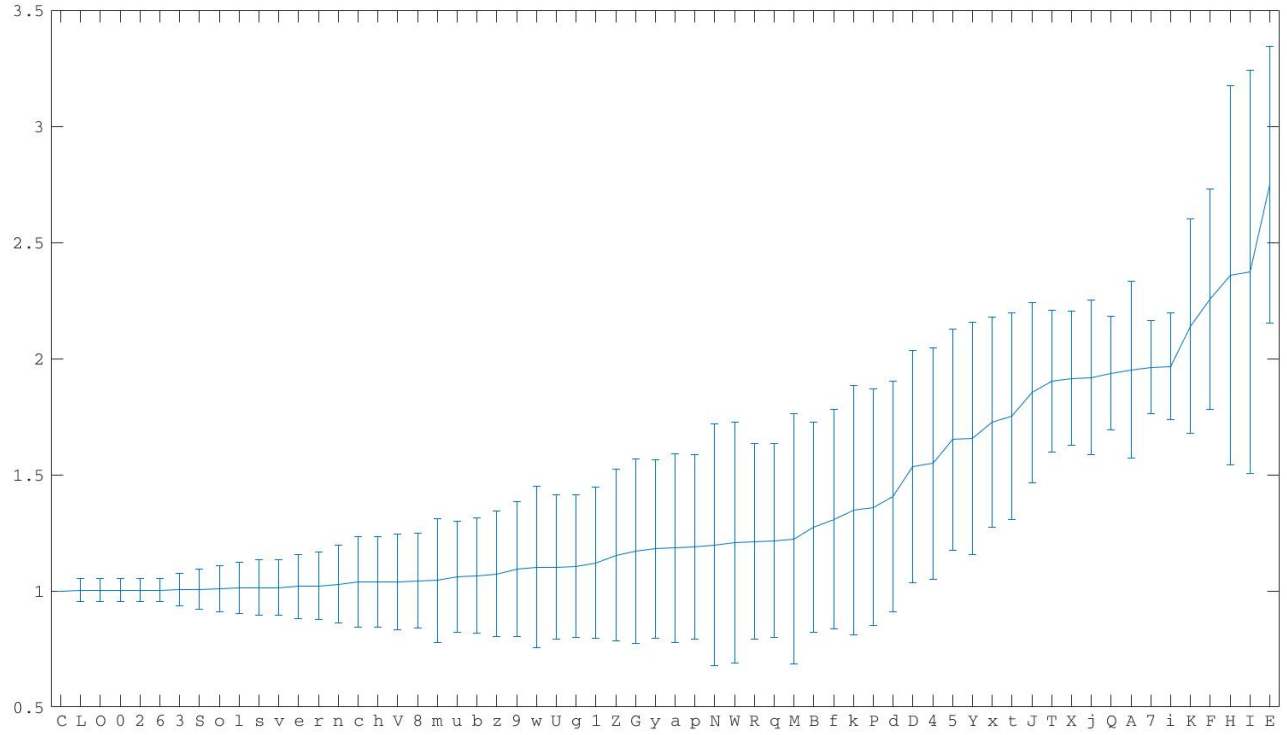


Figure 2.3: Summary statistics strokes for all categories in *IRONOFF* dataset (uppercase and lowercase letters, and digits), starting from the simplest to the more complex.

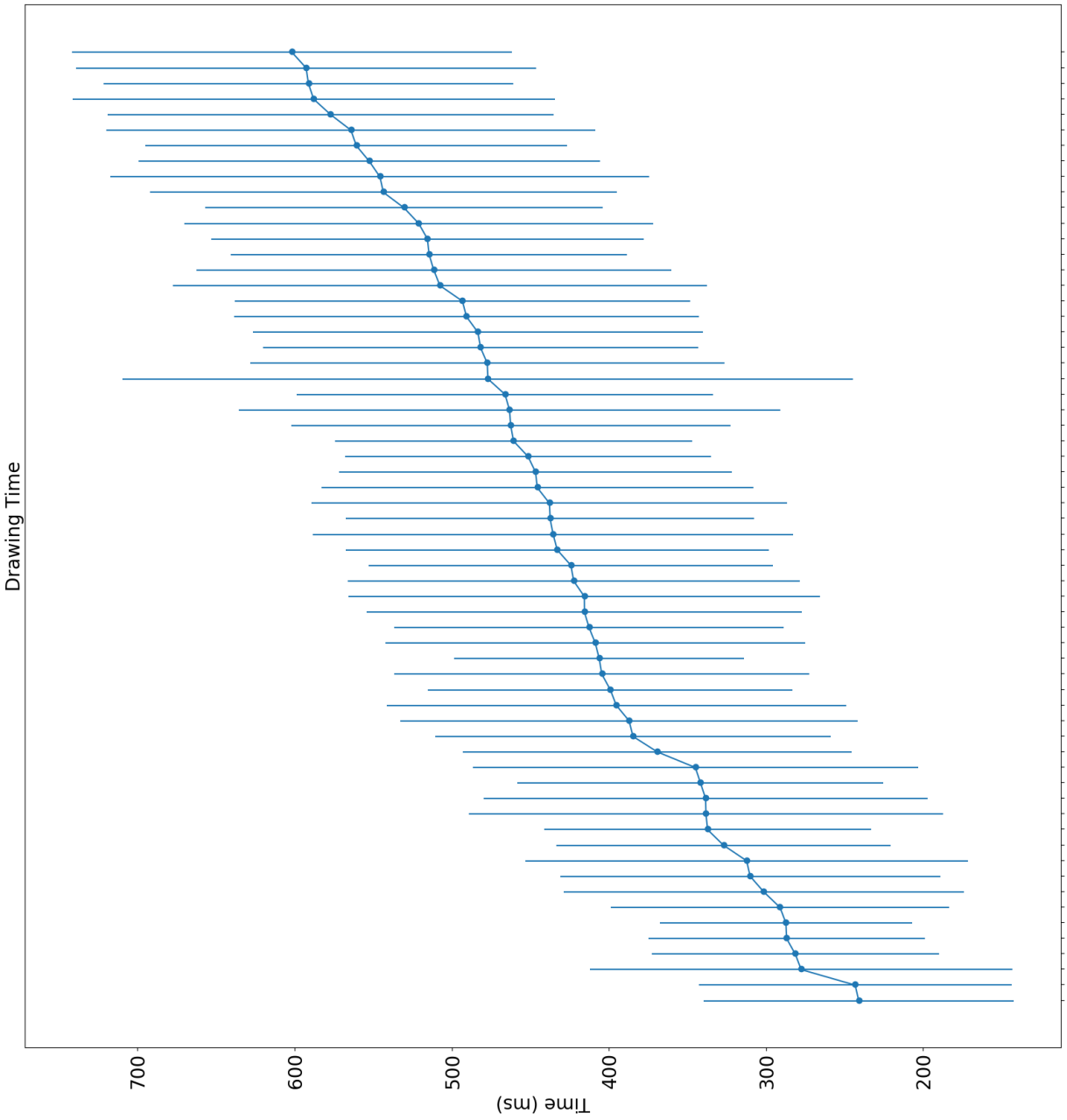


Figure 2.4: Drawing time for all categories in *IRONOFF* dataset, arranged from the smallest to the largest.

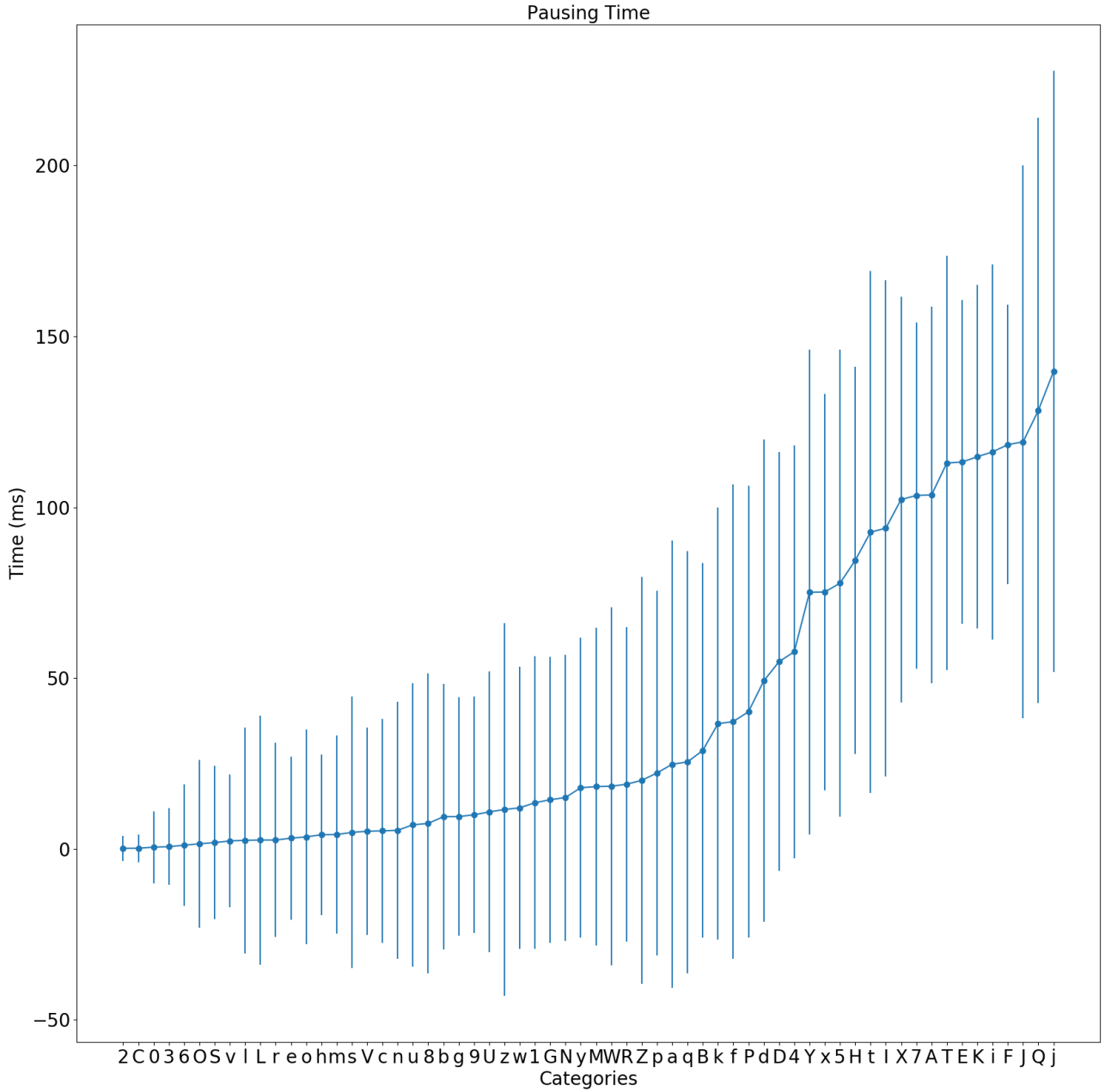


Figure 2.5: Pausing time for all categories in *IRONOFF* dataset, arranged from the smallest to the largest.

2.2 Sketch Drawing – *QuickDraw!*

Around 50 million drawings have been collected by players of the game *Quick, Draw!* (Jonas Jongejan and Team, 2017), where players are asked to draw one of 345 categories, and a neural network tries to classify the drawings into the right categories. With more data collected and labeled, the network gets better (it learns from the flagged errors). The collected dataset is available online for free (Google, 2017). Example of the shapes collected are in figure 2.6.

Each sample in the dataset contains the following data:

- key_id: a unique identifier for this sample.
- word: the category the player was asked to draw.
- recognized: if the neural network in the game did recognize the drawing as part of this category.
- timestamp: to mark the creation time of the drawing.
- countrycode: the location of the player when the drawing was made.
- drawing: an array containing the X, Y trajectories, and the time T for each point in the trajectory. Points belonging to each stroke are grouped together.

In order to focus on the styles aspect, we decided to focus on 5 categories: circle, triangle, square, hexagon and octagon. Our reasoning is that the more complex the task gets (cats for example), it is harder to have a subjective opinion about the styles, and hard to give insights about the results. This is not a limitation on the methods we are proposing though.

We sampled 20K samples from each task to perform exploratory analysis on them. In terms of strokes (see figure 2.9), we can see that there is a large variance surrounding the mean of each category. This trend continues when we look at the drawing time (see figure 2.11) and the pausing time (see figure 2.10). In the sample we analyzed, there are players from around 160 countries, figure 2.12. This is one indication to the increasing complexity *QuickDraw!* dataset presents in compare of *IRONOFF* dataset. **Put the country distribution in the dataset.**

Not all examples are recognizable during the game though. In many cases, the players do not draw the required shape, or draw something quite complicated. The results of the recognition can be seen in figure 2.7, with the relation between number of strokes and length of the drawing. We can conclude that, for the selected categories, the more complex the drawing is (more length or more strokes), the less likely it is to become a recognizable drawing.

This dataset is considerably more challenging than *IRONOFF*, for several reasons:

Figure 2.6: Examples from different categories in *QuickDraw!* dataset. Source of the images are (Google, 2017)

- Even though the players are asked to perform a particular task (draw a circle for example), in several cases, there is no clear resemblance between the drawing and task (e.g., when drawing an octagon, a lot of the recognized drawings do not really resemble an octagon).
- The players used a mouse in order to perform the drawing[3]. This sometimes lead to weird behaviors in terms of speed of movement (too slow, too fast), and the number of strokes (players sometimes tend to simplify complex shape, by drawing the whole shape in one stroke, and sometimes they spend too much time to draw it well, with too many strokes).
  This is unlike handwriting, where the writers usually tend to follow some rules (Séraphin-Thibon et al., 2019), which is not mostly the case in this dataset.
- Thus, some extra parts of pre-processing will need to be added in order to reduce these effects of the mouse, and make the data more closer to handwriting behavior.
- As mentioned earlier, the variance for each of the selected categories in this dataset is considerable higher than *IRONOFF*.

Since these drawings are done using the mouse, an interesting aspect for the recognizable images is the simplicity of strokes used – easier for player – (see figure 2.9). If a pen is used in the drawings, this particular behavior would not observed. For example, in case of hexagon and octagon, one can expect a higher density on the 6 and 8 strokes respectively, and much less on 1 stroke. Our observation is that it is easier with the mouse to draw the whole shape with one (or few) strokes only. This has two consequences:

- It is difficult to generate the strokes: One/few strokes means that a direct stroke representation is quite sparse (if the length of the shape sequence is 200 time steps, then among all the zeros (199 zeros, representing the current stroke), there is a single 1 value (representing the end of the stroke). This problem has also been noted in the work done in (Ha and Eck, 2017), and it is a challenging task to tackle.
- Unlike in *IRONOFF* dataset, where the strokes is a contributing feature in identifying the letter and the rules of drawing it, the strokes are not expected to play the same rule in *QuickDraw* – unless further processing is done –.

For each class, we randomly – without replacement – sampled only from the recognized drawing, traces with less than 200 time step long. 2K samples (total is 10K samples). 1K is used for test, 900 for validation, and the rest is the training data.

[3]Although there is no reporting about the tools used in the drawing, it is a valid assumption to assume that the mouse is the main tool.

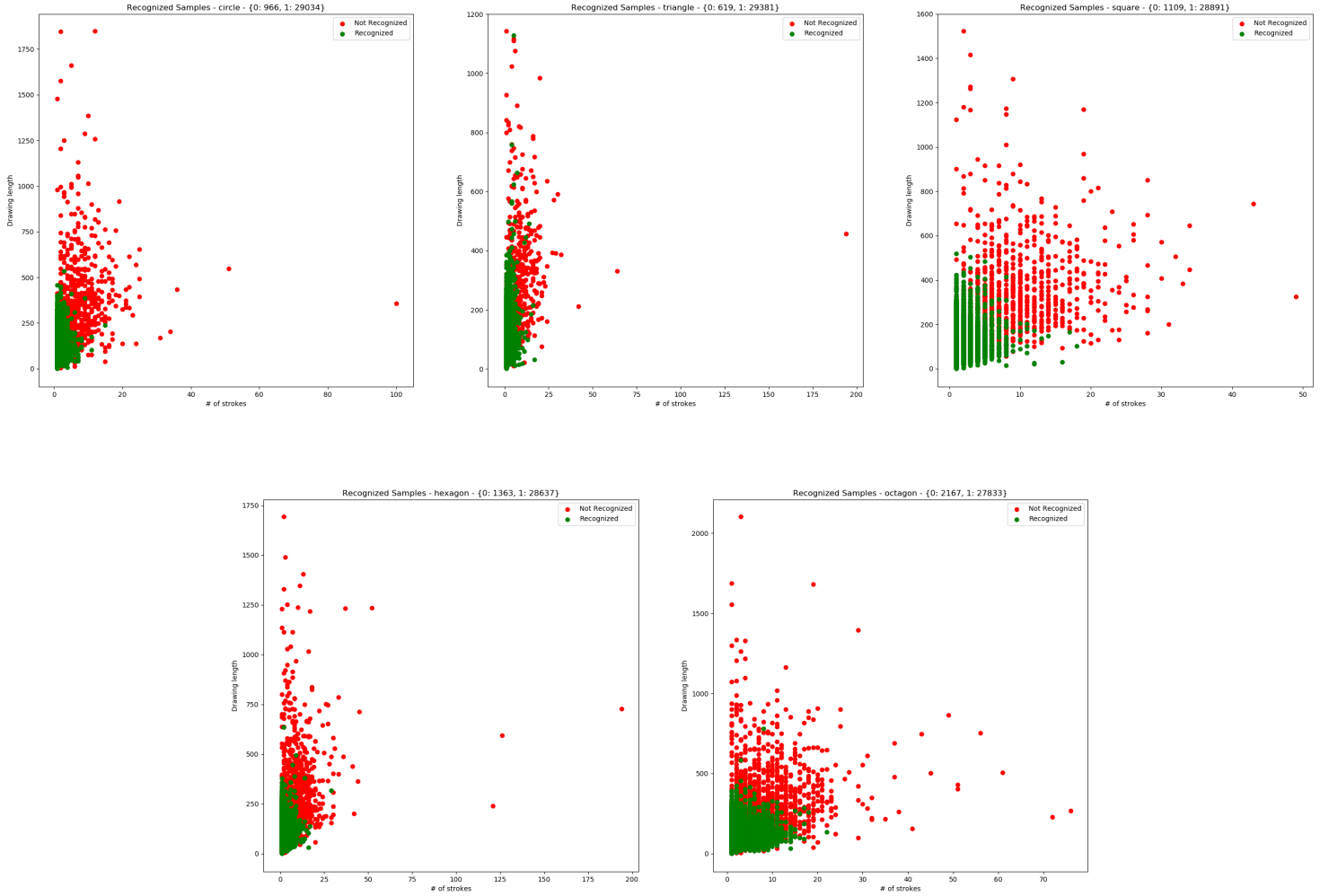


Figure 2.7: The distribution of the strokes in *QuickDraw!* for the recognized shapes, for each category.

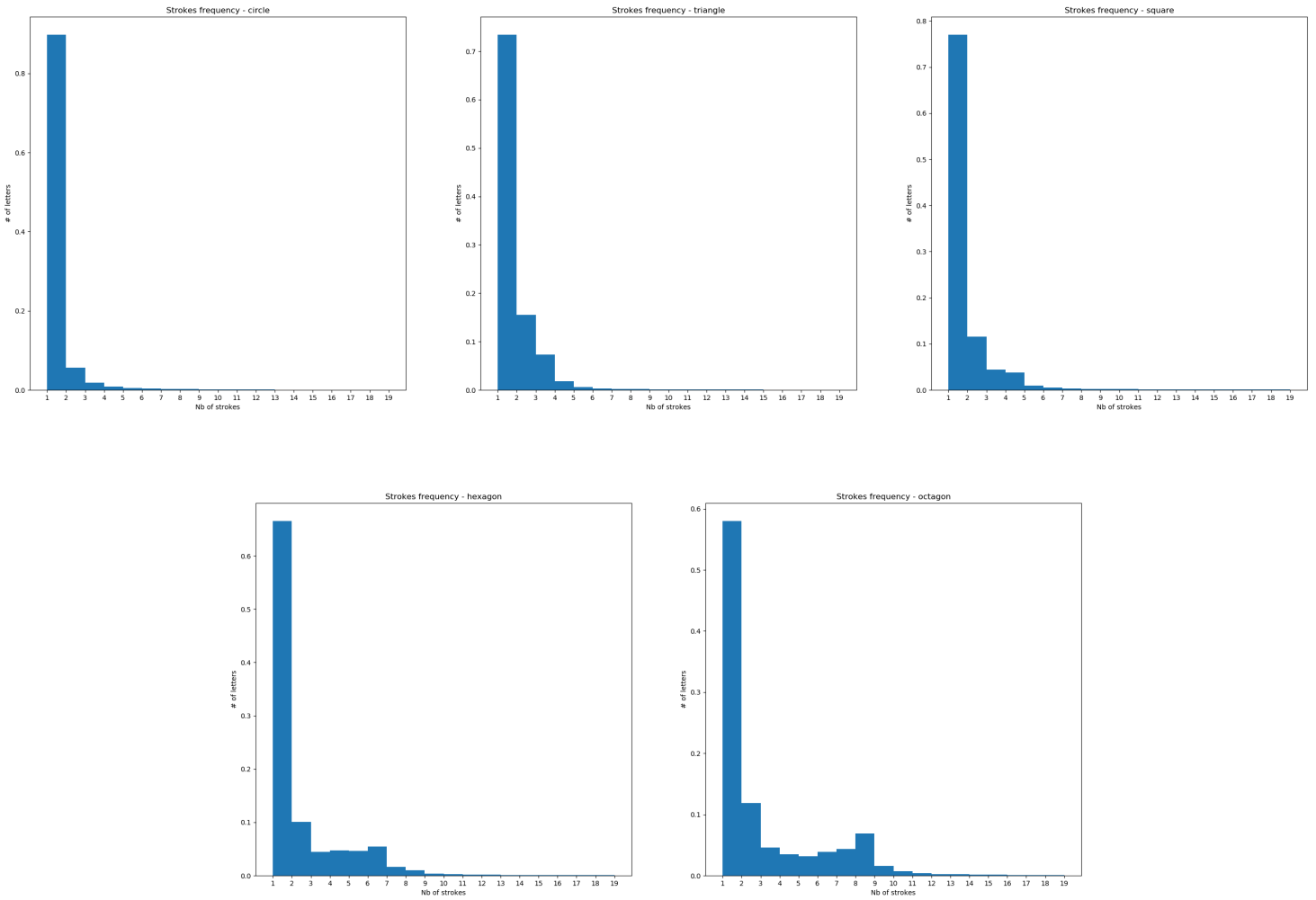


Figure 2.8: The recognized VS non-recognized drawings in *QuickDraw!* in each of the selected categories.

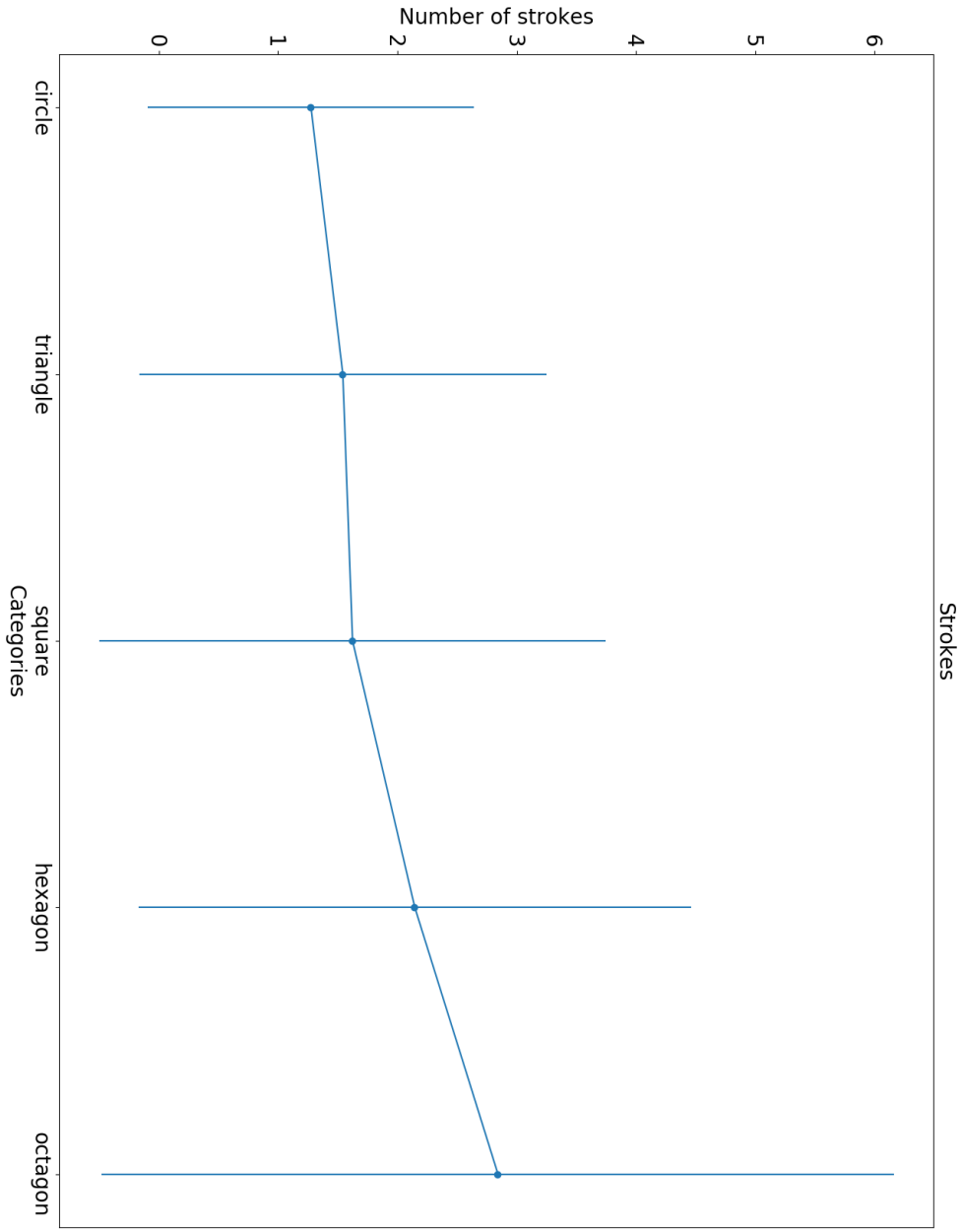


Figure 2.9: QuickDraw! strokes statistics for each of the selected categories.

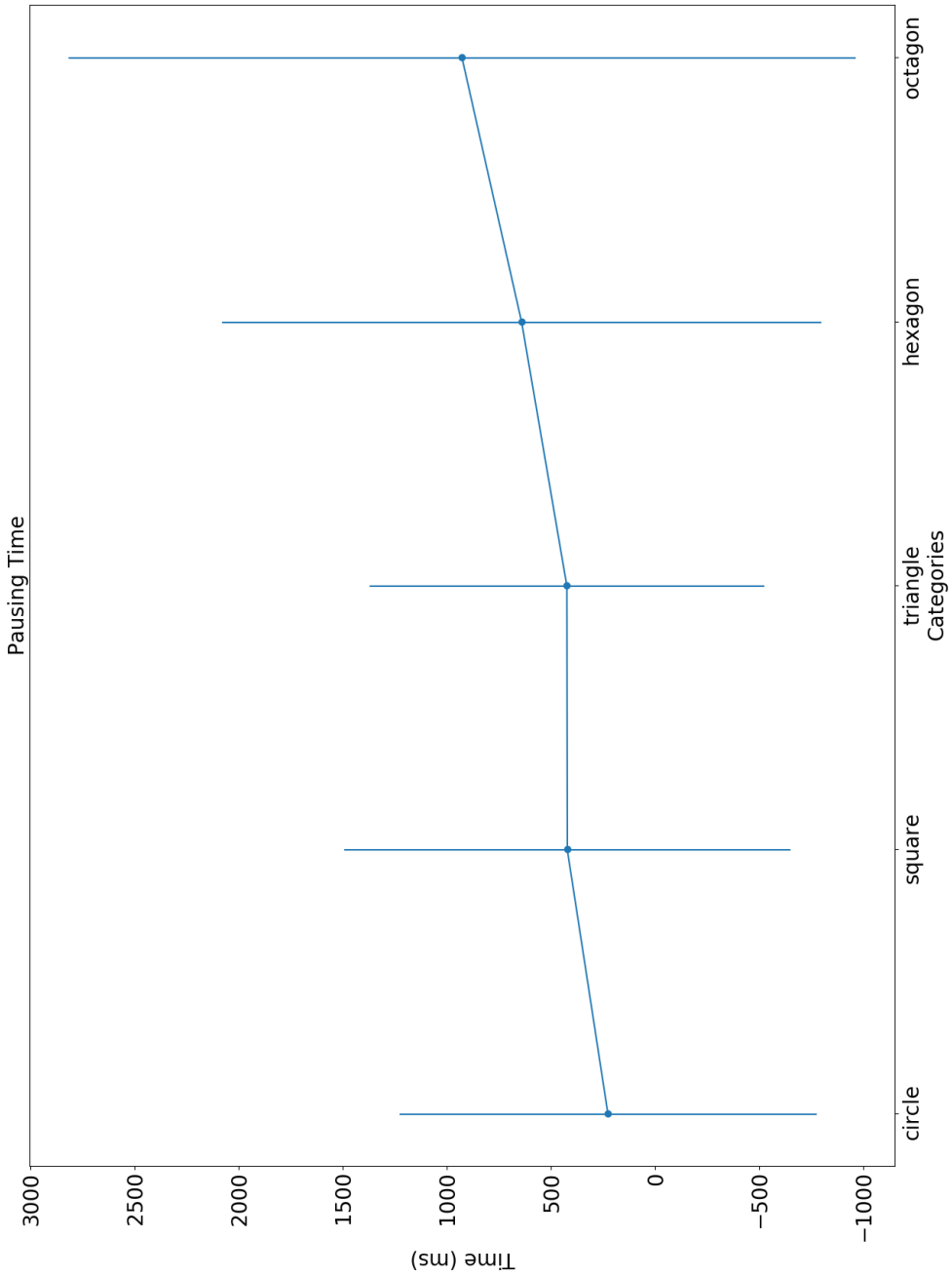


Figure 2.10: QuickDraw! pausing time statistics for each of the selected categories.

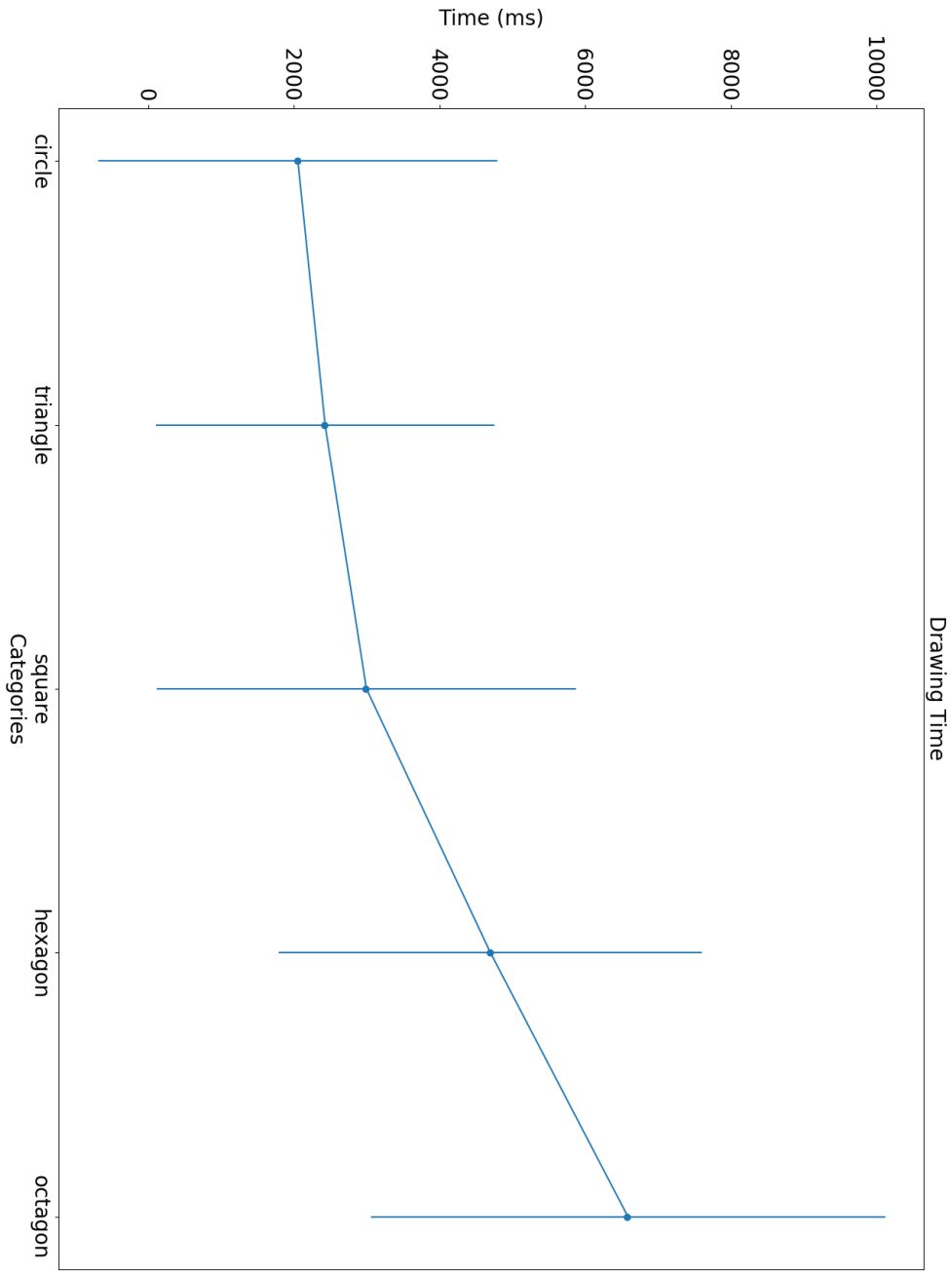


Figure 2.11: QuickDraw! drawing time statistics for each of the selected categories.

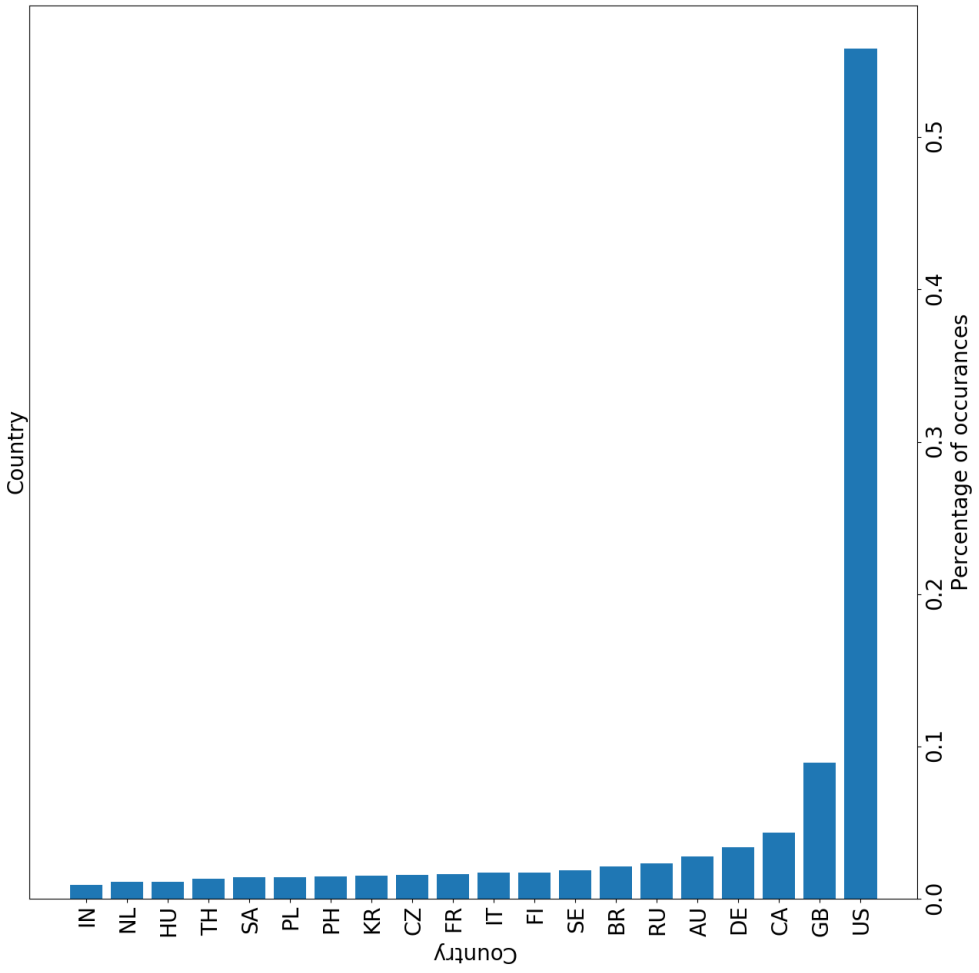


Figure 2.12: The most dominant countries for the players in QuickDraw! dataset. In the sample we analyzed, there is players from around 160 countries, but the majorty are from United States, followed by Great Britain.

2.3 Data representation

This section should be more general, discussing both pre-processing and represenation in the same time. Otherwise, we can discuss only the representation, and move the pre-processing steps to the appendix (which makes sense). Also, there is the pre-processing part Gerard performed on QuickDraw, which should be discussed.

The choices of data representation is a key factor in the success or failure of the machine learning based approaches. This choice, however, is also entangled with the task to be done (in this case, the study of styles).

A good representation tries to:

- Maximize the density of $data/patterns$ ratio: machine learning algorithms are statistical algorithms. It performs better when we have more examples for the patterns we want to learn (e.g., in case of cat/dog image recognition, having more example images for these two categories will always lead to a better performance). Another way is to reduce the number of irrelevant/unnecessary patterns to be learned from the data. This is the task of *feature selection*, which is a fundamental step in machine learning. The target is to increase the amount of signal-to-noise ratio in the data. All-in-all, the objective is to increase the ratio of $data/patterns$, either by adding more data (if it is possible, or by using synthesis data using methods like *Generative Adversarial Networks* (Goodfellow et al., 2014)), or removing irrelevant patterns.
- Simplify the learning procedure (differentiability/richness of distributions): while working with deep neural networks provide us with a lot of power in modeling a large variety of task, it also imposes some constraints. For example, the whole learning pipeline must be differentiable (this is a limitation imposed by the optimization methods, which will be discussed later). There are a lot of already available *off-the-shelf* functions that can be used, but care must be in the design of the experiment, to make it fit with these functions. Sometimes however, these functions are not enough to model the task properly. Thus, there is a need to adapt new tools to fit within the neural networks paradigm. This including finding a proper way to differentiate them, or, if not possible, to move around the non-differentiability problem (usually by finding a surrogate function to optimize), which is not always a straightforward task for deep learning practitioners.
- Keep enough patterns to perform the intended study: it is quite tempting to focus on the scores of the machine learning algorithm, while forgetting about the original task. In our case, machine learning is a tool to help us perform our analysis, but not the final objective. For example, a low-level quantization of the X and Y traces of the pen will remove a lot of patterns, and will make it easier for the algorithm to learn the data distribution. But it will also remove essential information about the styles in this case.

In the following subsections, we will discuss two different data representation choices (continuous versus discrete representation, and the features used), and the implications of each of them on the machine learning, and the study of the styles.

2.3.1 Continuous or Discrete representation?

The X, Y and pressure of handwriting tracings are always recorded as continuous distribution, while the pen state is discrete (categorical). Thus, it probably makes sense to model the data in their native form. In the neural network design, a typical design choice will be to use a linear activation function as output function, and the *Minimum Square Error* (MSE) as loss function. Unfortunately, it is not that straightforward.

Continuous Data Representation To understand the problem, we first need to consider how a simple feed-forward neural network works, figure 2.13. The input is x, the output of the network is a, and we want to predict y. The neural network tries to project x to a new space z_{21} through a series of continuous folding of space. Then, the rule of the last activation layer is to model $P(y|z_{21})$.

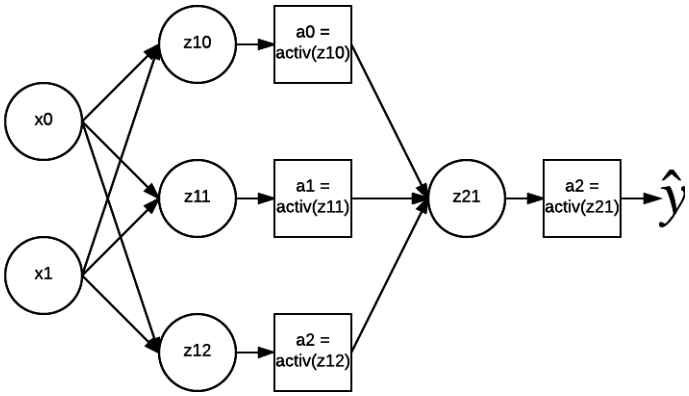


Figure 2.13: Example of a single-hidden layer neural network. The circle units is the sum of the weighted neurons connected to this unit, and the square units is the application of the activation function on that sum.

Although a linear activation is very simple, $y = z_{21}$, it is also shown to have limitations. In (Bishop, 1994), a simple example is shown (see figure 2.14). If each input x gets a unique output y (one-to-one mapping), then linear activation performs well (figure 2.14, left). But if the input can have multiple possible outputs (one-to-many), the model learns to average over these outputs (figure 2.14, right). The author concludes that a simple linear activation function is not powerful enough to represent a complex/rich distributions. He then proposed the use of a *Gaussian Mixture Model* (GMM) (Murphy, 2012) as the final activation of the neural network, which is powerful

enough to enable modeling complex continuous distributions, and avoids the problems of a simple linear activation function. This combination of neural network and GMM is called *Mixture Density Network* (MDN). The loss function in this case is changed from the MSE to the a posterior log-likelihood of the GMM. The neural network output in this case is not the required prediction directly, but the parameters of the GMM (means, variances, weights, correlations). The required prediction is then sampled from this parameterized GMM.

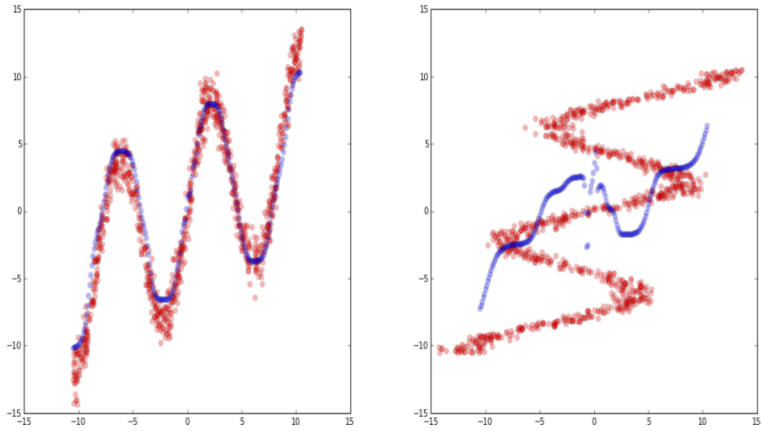


Figure 2.14: The performance of the simple linear activation function on two setups. Left: in case of one-to-one mapping between the input and the output, it performs well. Right: In case of one-to-many mapping, the function starts to average over the seen observations, leading to undesirable behavior. Source of this image is (Ha, 2015).

The work done in Graves (2013) demonstrated a system which generates impressive results on handwriting demonstration. He used an adaptation of the MDN for temporal data. Other applications for MDN can also be found in speech synthesis (Wang et al., 2016, 2017a; Zen and Senior, 2014). While there is no question about the power the MDN approach provides, it was reported in (Graves, 2013) that training model kept collapsing (the explosion of gradients, the loss going to inf or NaN), thus requiring tweaks and tricks in order to train properly. This is in-line with the experiments I performed with MDN, leading to similar conclusions.

Discrete Data Representation Discrete representation of originally continuous data requires transforming the data via feature engineering and/or quantization step on the raw data. While it is guaranteed there will be some information loss in the discrete data, discretization provide a lot of robustness to noise (this is why it is used in digital communication for example), and flexibility (many tools based on information theory do exist to handle digital data). Plus, a categorical distribution does not make assumption of the shape of the data distribution, unlike continuous distributions. The use of discrete distribution and how to infer from a discrete distribution will be covered later

in section 3.1.

The challenge in this case is to choose a good quantization technique, that preserve the relevant information in the original signal one one side, and while keeping the dimensionality of the problem to a tractable level. For example, in case of speech synthesis, the authors in (Oord et al., 2016a,b) showed that applying the μ-law to the raw speech signal, and then quantize it (thus, the quantization is not linear), revealed a superb sound quality. This is better than performing naive linear quantization on the data, and saves a lot of memory as well: with non-linear quantization, they only need 256 levels. To get a similar behavior with linear quantization, around 65K quantization levels or required.

2.3.2 Feature engineering: Direction and Speed

As argued in the previous section, we choose discrete representation for our data. A requirement for a good quantization scheme should keep the important information in the original signal[4]

In case of *IRONOFF*, the letters tracings has been cleaned by removing points related to false starts or corrections as well extra strokes. Tracings were re-sampled with $10ms$ time interval, and the ones with length exceeding 1 second has been removed, as well as tracings more than 100 time steps. This is because they are quite rare, thus, their existence would significantly degrade the performance of our model.

For *QuickDraw!*, the re-sampling interval is $20ms$, and consider tracings that are less than 200 time steps. This is because the tracings tend to be longer than in *IRONOFF*. This start to became clearer when the task gets more complex (the octagon being the most complex one).

We represent each letter tracing by two features: directions and speed. Each feature is quantized into 16 levels and represented as a one-hot encoded vector.

Freeman codes (Freeman, 1961) is used in order to encode the direction feature. It belongs to a family of compression algorithms called *Chain Codes*. This set of algorithms proved to be useful to encode an image with connected components. They can transform a sparse matrix to just a small fraction of the size of the image, in the form of a sequence of codes. Thus, they are being used as compression algorithms as well.

Freeman codes can N-directional codes (where N are the directions), depend-

[4]In some applications, like in digital communications for example, the quantization (or digitization process) has other rules, like compressing the signal, increasing the signal-to-noise ratio, and increase the robustness of the signal. This is outside the scope of our work however.

ing on the needed resolution. It is quite simple as it encodes each direction with a unique number from 0 to N-1. A direction is defined as the directed vector connecting two neighboring pixels on the contour of a connected component in the image.

We compute the change of directions between three consecutive points. Then, we map this change to its corresponding freeman code number, as shown in figure 2.15. Last, we transform the direction number into one-hot encoding scheme, and use this as input to our network. We also quantize the speed of each displacement.

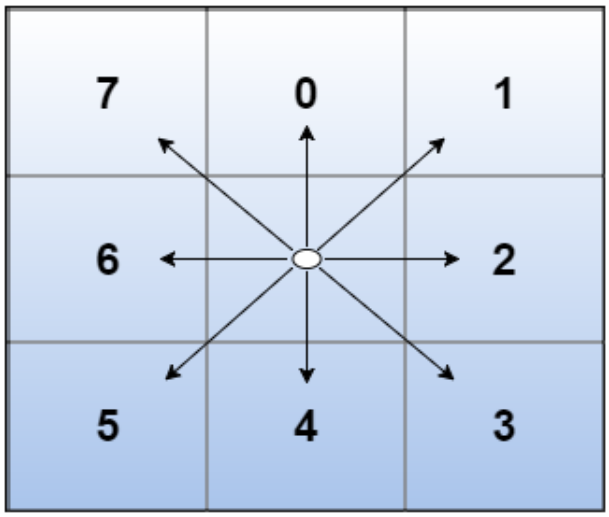


Figure 2.15: Example for freeman code representation for 8 directions. Each direction is given a unique number.

2.4 Summary

In this chapter, we explored two datasets: Cursive Handwriting dataset, *IRONOFF* and sketch drawing dataset *QuickDraw!*. Each of these datasets contains several tasks/categories (the letters/digits in case of *IRONOFF*, and the shapes in case of *QuickDraw!*), and we can explore the styles around each of those tasks. We explored basic information about each task (number of strokes, drawing time, and the pausing time).

We motivate the use of these datasets because we can see that there is a variance in these information of each task, suggesting different styles for each task. This variance differs a lot depending on the complexity of tasks: in *IRONOFF* for example, letter *c* seems to be the simplest task, thus, the variance in it is relatively not that large. This variance increases as the task get more complex (letter *E* for example). In *QuickDraw!* dataset, a similar trend can be observed (the *circle* being the simplest

task, and the *octagon* being the most complex one).

Last, we argue that although that *QuickDraw!* has simpler tasks than *IRONOFF*, it is actually more complicated. We hypothesize that this is due to the fact that there is a huge variety in the players (many different countries, compared to mostly *French* people in *IRONOFF*). The players use mostly the mouse in order to draw[5], which leads to behaviors usually unobserved with hand drawing/writing. Also, in such environment, it is expected that the human curiosity will prevail, and people will try to draw complex shapes, outside the limit of the required task, in order to see if the neural network classifier will recognize it correctly or not.

We will consider that there are two hierarchies of tasks in *IRONOFF*: the first is uppercase, lowercase letters, and digits. The second is the individual letters and digits. When we perform transfer learning, we will do it on the tasks of the first hierarchy only (for computational reasons).

We will use *QuickDraw!* side-by-side to *IRONOFF* in our last part of our work in order to validate our approach and conclusions, but the first two parts will be done on *IRONOFF* only.

[5]The players did not receive any special training on drawing with the mouse beforehand.

Part II

Experiments

Chapter 3

Generation, benchmarks and evaluation

Contents

Since styles are ill-defined, they can not be evaluated explicitly (we can not quantify them in advance). Thus, we need a proxy method in order to implicitly evaluate them. We can do this by using them in order to generate behaviors (i.e., synthesis handwriting traces), and evaluate the quality of those behaviors relative to the target behaviors (i.e., the original letters traces). In other words, to study styles, we would like to reconstruct the target behaviors using generative models, and the task and the styles to perform the behavior. In this chapter, we discuss the recent advances in neural

generative models: how are they trained? how do we infer from them? Then we look at paradigm for reconstruct target behaviors, with focus on the usage of neural networks.

After we generate the behaviors, we need to evaluate them in a manner that evaluates the styles. This is still an open research question, and a complicated one as well. It also goes hand-in-hand with the choices made in generative models. We discuss the used evaluation metrics across different domains (e.g., text, speech and handwriting evaluation). We also discuss the difficulties associated with finding the suitable evaluation metric.

I then present the experiments done in order generate drawings and ground our proposed evaluation metrics. I will explain our choices for the model design, the inference methods, and our approach to ground the metrics.

Questions addressed in this chapter

- How to generate handwritten letters using deep learning framework?
- How to evaluate the generated traces?
- What benchmarks are suitable to compare to?

3.1 Background

We start by introducing the different basis in the literature for our work. Most of this work uses deep learning and generative models. How and why deep learning emerged in the last few years is quite important to keep in mind, since it is the starting point for any successful usage of deep learning. We then explore a particular aspect of deep learning: generative models. In particular, we focus on the domain of sequential data, where the problem gets more challenging. A direct consequence for using generative models is the challenge of evaluating what is generated. Generation, after all, has an artistic aspect, unlike well-defined machines learning tasks like regression and classification. We try to make sense of what exists in the literature, and try to deduce what would be a good criteria for new evaluation metrics for a generative task.

Deep learning is a subset of machine learning (Goodfellow et al., 2016; LeCun et al., 2015), mainly applicable on neural networks. These set of techniques perform remarkably well nowadays on a wide range of tasks and benchmarks (for example, in image recognition (He et al., 2016; Krizhevsky et al., 2012; Simonyan and Zisserman, 2014), speech synthesis (Oord et al., 2016a), image segmentation, handwriting recognition, image captioning (Karpathy and Fei-Fei, 2015; Vinyals et al., 2014), language translation (Sutskever et al., 2014a)), even outperforming humans in some of them (GO game (Silver et al., 2016)).

Before deep learning, in order to use classical machine learning algorithms, an important step was *feature engineering*: extracting the relevant features from the data, in order to get the relevant information needed to perform the task. In images, techniques like *Scale-invariant feature transform* (SIFT) (Lowe et al., 1999), and in speech, feature like *Mel-frequency cepstrum* (MFCC) and *Probabilistic Linear Discriminate Analysis* (PLDA) (Narang and Gupta, 2015), were quite dominant at that time. There is no one solution that fits all here; it depends on the task in hand, and that required experience. Thus, feature engineering was quite challenging.

The advantage of deep learning techniques is that it overcome (to a big extent) the need for the daunting task of feature engineering. Instead, it tries to learn, from the raw data, hierarchy of features, that are optimal in order to solve the task in hand (i.e., it performs *automated feature engineering*). In our work for example, we did not need to do any kind of temporal feature extraction[1].

A detailed discussion into the basics of deep learning is out of the scope of this document (and has been done properly in many books and tutorial available online). We refer you to the book (Goodfellow et al., 2016) for more theoretical treatment of deep learning, and to (Chollet, 2017; Géron, 2017) for a more practical aspect.

[1]When it comes to spatial features, we performed feature-extraction using Freeman codes and Speed modalities, as discussed in the previous chapter.

3.1.1 Sequential data

Sequential data appears in a lot of our daily life, for example: text, speech, the weather status, etc. In a more formal manner, a sequential data example is formed of tokens, and can be represented as $x_1, x_2, ..., x_N$, where x_n is one token, and N is the total length of this sequential data point.

Let's take a more concrete example: text. We have multiple sentences in a given text. We can consider each sentence as a data point/example. If we consider an example sentence: *the cat is eating the food*, and we define our tokens to be the words[2]. In this case, $x_1 = the$, $x_2 = cat \ldots etc$.

What is common between all the sequential data (and what also distinguish them from non-sequential data points) is that each token is dependent on the previous token/s. In general, when we want to model sequential data, we in essence want to learn the following probability distribution:

$$p(x_1, x_2, ..., x_N) = p(x_1) \times p(x_2|x_1) \times p(x_3|x_2, x_1) \times .... \times p(x_N|x_1 ... x_{N-1}) \tag{3.1}$$

In order to model such a distribution using neural networks, we can either:

1. Adding a *state variable* in order to factorize the problem. In this case, we introduce an intermediate state variable, h, which model the dependency on the previous tokens. Our objective in this case will be

   $$p(h_n) = p(h_{n-1}, x_n) \tag{3.2}$$

   This can be done using *Recurrent Neural Networks* (RNN), which will be explained later in detail (section Section 3.1.2). In this case, the network is looping over the tokens one-by-one, updating the state variable each time.
2. Use the whole sequential data (all its tokens) as input to the network in the same time. In this case, the network is trying to model equation (3.1) heads-on. This approach has gained popularity recently with the work done in (Gehring et al., 2017; Oord et al., 2016a), using convolution networks in order to model sequential data (speech for the first, language text for the second). The advantages of this approach is removing the sequential aspect of the network, and leveraging parallelism when treating the whole sequence. This results in faster training, while still achieving state-of-the art results.

[2]There is no rule here for how we define the tokens. We can, for example, define the letters to be our tokens. In the case of text, it is a trade-off: considering words as tokens allow the model to be more fluent, but it also means that the learning space is very large (sometimes the number of unique words to predict at each time step is in the order of tens of thousands). If we consider letters as tokens however, it makes the model job more tractable (all the letters, symbols, digits, can be in the order of tens), but it leads to less quality for the model.

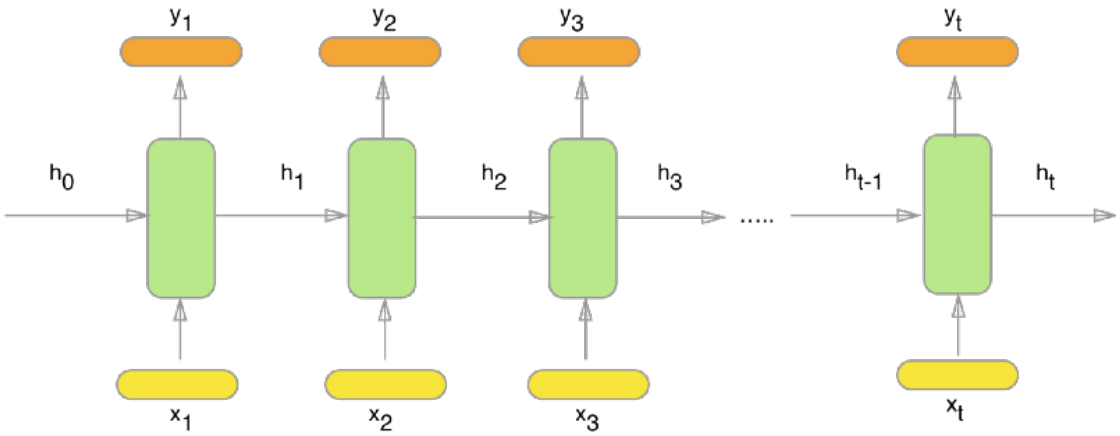


Figure 3.1: A demonstration of how RNN works: the network is applied on each token in the input $(x_1, x_2, ..., x_t)$, while update the hidden state variable every time $(h_0, h_1, ..., h_t)$. The output at each step is a function of the hidden state variable (not demonstrated here). Source of the image is from (Kostadinov, 2017).

There is no one solution that fits all here, it simply depends on the problem and the constraints on the solution. In case of variable-length sequential data, the first approach is the one to go for. In the case of fixed-length sequential data, the second approach should be considered (easier to train the model, faster to optimize, faster to infer from).

3.1.2 Recurrent Neural Networks and Sequence Modeling

As mentioned briefly in the previous section (Section 3.1.1), *Recurrent Neural Networks* (RNN) is a type of neural networks, that can handle sequential aspect in data. In its simplest format, it is a simple feed-forward network, applied on each token in the sequential data point, while carrying the information about the previous tokens using a latent variable h, commonly referred to as *the hidden state variable*. This is demonstrated in figure Figure 3.1 on page 59.

To describe it in a more formal manner, let's first assume the following:

- The input x is first processed by a set of weights, W_{ih}, and bias b_{ih}.
- From each step to the other, the hidden state h is processed via a set weights, W_{hh}, and bias b_{hh}.
- Last, the output is given by processing the hidden state h via another set of weights, W_{ho}, and bias b_{ho}.

If we assume that output activation is a simple linear layer (thus, it is a

regression task), then the equations of this simple RNN are the following:

$$\begin{aligned} z_n &= W_{ih} \cdot x_n + b_{ih} \\ h_n &= W_{hh} \cdot [h_{n-1}, z_n] + b_{hh} \\ y_n &= W_{ho} \cdot h_n + b_{ho} \end{aligned} \tag{3.3}$$

Where the brackets [ ] indicate a concatenation process. In case of a simple regression task, a typical loss function is the *Minimum Square Error*, which can be formulated as:

$$Loss = \frac{1}{T \times N} \sum_{n}^{N} \sum_{t}^{T} (y_{n,t} - \hat{y_{n,t}})^2 \tag{3.4}$$

where T is the length of the sequence, and the N is the number of sequential data points available, and $\hat{y_{n,t}}$ is predicted output by the model, while $\hat{y_{n,t}}$ is the ground truth output.

Given this formalization, the objective is to find the set of weights and biases of the network, that will minimize the chosen loss function. We will explore the optimization algorithms in section 3.1.3.

3.1.3 Optimization Algorithms

One of the important factors in the recent success of deep learning is the advances in optimization algorithms. These algorithmic advances make the optimization easier, converge to a better solution, and require less tweaking in order to achieve a good performance. All these methods are derivatives of gradient descent optimization, defined as (Wikipedia, 2019b):

> Gradient descent is a first-order iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.

If the function to be optimized is convex (Wikipedia, 2019a) – differentiable, and have a global optima –, then gradient descent can find this global optima. An example for the different optimization iteration can be seen in figure Figure 3.2 on page 61.

Gradient descent can be formalize by the following equation:

$$\theta_{t+1} = \theta_t - \alpha \times \frac{dLoss(\theta_t)}{d\theta_t} \tag{3.5}$$

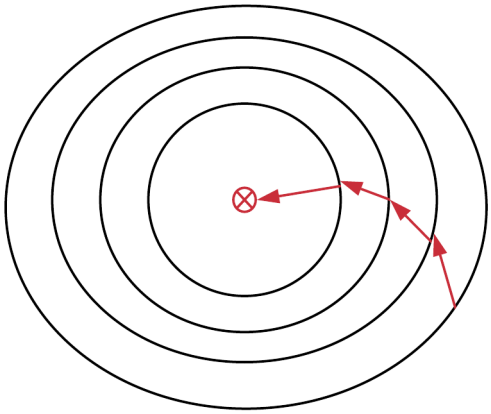


Figure 3.2: A demonstration for how a gradient descent algorithm work. The optimization process progress each step towards the global optima (the indicated point in the center).

where θ refers to the parameters we want to optimize, $t, t+1$ refers to the current and the next iterations respectively, α is the learning rate (how large the step to take at each iteration), and $Loss$ is the loss/objective function we want to optimize our parameter for. Optimization by gradient descent is a *batch optimization*: the loss function and its gradient is calculated all given data examples. This iterative process keeps going till we wither do not observe a tangible change in the performance of the parameters, or we exhaust our computational budget. An illustration for this process can be seen in figure Figure 3.3 on page 62.

Before we dive into the different optimization methods currently used, it is important first to stress on an important characteristic of all modern deep neural networks, which are (Goodfellow et al., 2016):

- **Differentiability**: All the components of the neural networks (activation function, loss objective, regularization, ...etc) are differentiable components. This issue is not about mandatory by theory, but by convenience: good optimizers exist that can use this feature in order to optimize the network faster, while being able to scale with appropriately with the given number of data points and the number of parameters to be optimized.
- **Non-convexity**: While the neural network is built of convex parts, the composition of those parts together is not convex. The reasoning for this particular point is out of the scope of this document. We refer you to the book (Goodfellow et al., 2016) for more details –. While this seems to be a disadvantage, it is actually a powerful advantage of the neural network. Neural networks are *universal approximator*, meaning that they can approximate/learn any function. A convex function can not

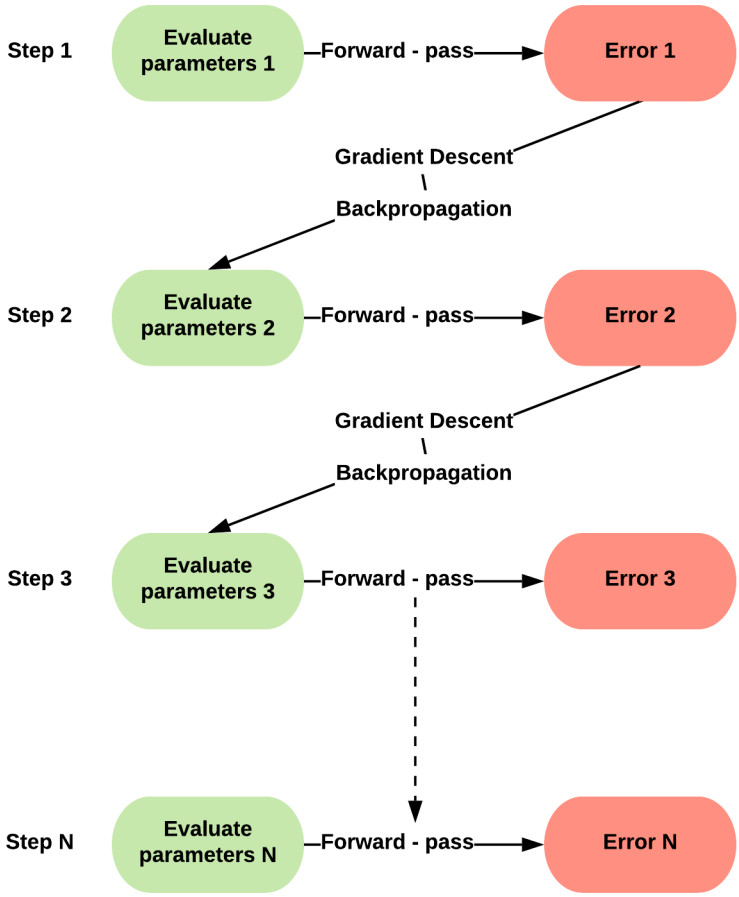


Figure 3.3: The process of gradient descent is iterative, and include 3 steps: evaluate the quality of the current parameters (forward-pass step), get the error value, use the error-value in order to update the parameters/getting new set of parameters (back-propagation step). This process is repeated N times, which is decided by either not observing any update in the error, or we ran out of computational resources.

approximate non-convex function, but the other way around works (De Sa, 2017). Using gradient descent optimization methods in this case leads to convergence to local optima points. The initial conditions of the network (the initial random parameters, and the strategy of their selection) and the optimization strategy and parameters (learning rate, decay factor$\cdots$, etc)$will\,play\,an\,important\,rule\,to\,determine\,the\,convergence\,characteristics$($speed\,of\,convergence$,$a$

Another important aspect of the success of deep learning is the availability of large amount of data. To optimize a deep neural network (can be the order of millions of parameters) using large amount of data (can be in the order of millions of data examples) using gradient descent, is not physically feasible. We do not have the hardware capability to process all of these data in the same time.

To get around this issue, *mini-batch* optimization is used: instead of performing gradient descent based on the information from the whole data, we divide the data into chunks (mini-batches). For each mini-batch, we calculate the loss and the gradient, and update the parameters, and then move on the next batch. The gradient descent in this case is called *Stochastic Gradient Descent*, SGD, (Robbins and Monro, 1951), following the same equation (3.5), but applied to only a mini-batch at a time, instead of the whole data (batch).

Many advances built on top of SGD helped in advancing deep learning, like combining SGD with momentum (Rumelhart et al., 1988), RMSProp algorithm (Hinton et al., 2012) and Adam algorithm (Kingma and Ba, 2014).

3.1.4 Inference: How to generate sequences from the network?

There exists several approaches in order to generate information from the networks. In the case of recurrent neural networks, this is a sequential process (one step at a time, till we generate the whole sequence), as illustrated in figure Figure 3.4 on page 64.

During the inference mode, the objective is to generate the most likely sequence. We want to solve the following problem

$$\begin{gathered}\text{Find } x_1, x_2, ..., x_N \text{ that} \\ argmax_{x_1,x_2,...,x_N} p(x_1, x_2, ..., x_N)\end{gathered} \tag{3.6}$$

This problem, however, is not tractable, as it scales badly with the number of options (i.e, dimensions) per time step, and the number of time steps required. One simple way is to perform *greedy sampling*, where, at each time step, we select the most likely token. This, however, leads to repetitive and predictable patterns (Chollet, 2017).

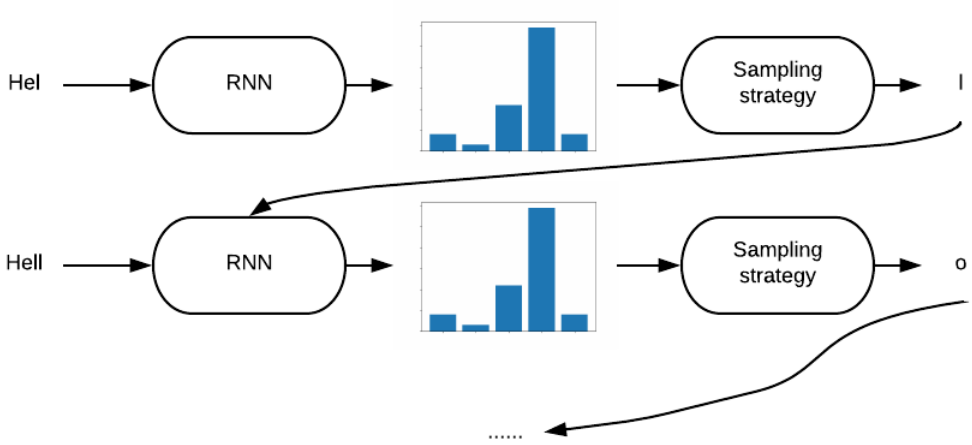


Figure 3.4: An example of using RNN in order to infer a sentence. The token to be generated here is a character. At each time step, the model is given the part of the sentence that has been generated so far, and asked to give the probability distribution over the next character. This distribution is then given to the selected sampling distribution, which sample the next character, and so on.

A better way will be to use *stochastic sampling*, by leveraging the fact that at each step, we have a probability distribution over all possible tokens. This allows more diversity in the generated tokens, and also allow unlikely tokens to be sampled some of the time.

But what if we want to control the level of randomness in the sampling process? Having such control will allow us to explore different ways to infer from the model, in order to determine the most satisfying way. This control can be done using *temperature sampling*. The idea is to reshape the probability distribution over the different token. On one extreme, very high temperature (going to infinity) will flatten the distribution, making the distribution equivalent to *uniform distribution*. On the other extreme, a temperature of zero will be mount to greedy sampling. This is illustrated in figure Figure 3.5 on page 65.

A clear advantage for temperature sampling is its simplicity, ease of implementation, and ability to generate diverse outputs. However, this also create a challenge when we want to focus on issues like repeatability and reproducibility.

Another important thing to notice is that the training/optimization part is not the same as the generation part. Usually in machine learning, there is a symmetry between the training and testing procedures. However, in generation, we are suddenly

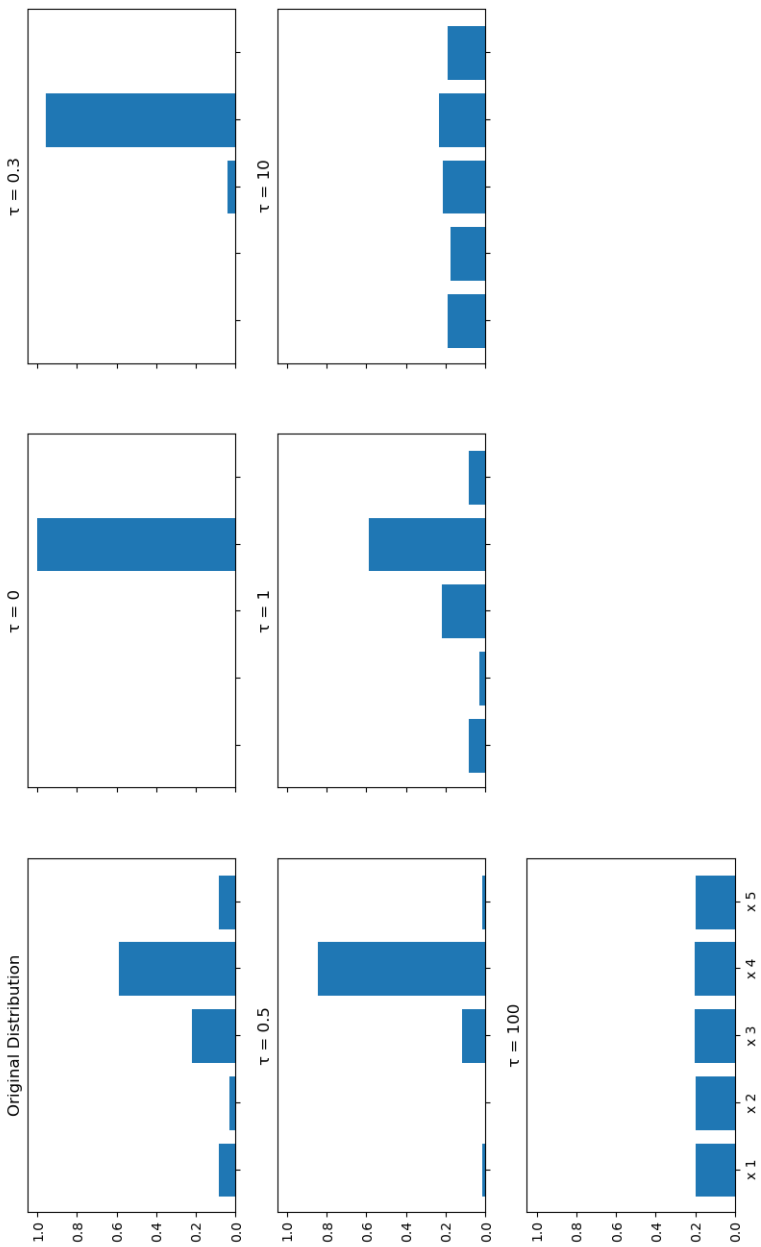


Figure 3.5: Illustration of temperature sampling. When the temperature τ is very low, it becomes greedy sampling. With the increase of temperature, we can see more higher possibility of sampling the lower-probability tokens. When the temperature is too high, it becomes uniform sampling.

letting the model generate long sequences, and use it output from the current step as the input to the next step. The model is not trained to see its own output in the input. It is trained to see always the ground truth in the input. Overtime, there is an accumulation of errors that build up, leading to the degradation in the quality of generation (Ranzato et al., 2015). This discrepancy between the training and generation will have consequences in the evaluation process, as we will see shortly.

3.1.5 How to introduce prior to the model? (conditioning the model)

When training a RNN network, we initialize the first hidden state with zeros. This way, we are informing the model that we are not making any prior assumptions about that particular sequence, and that all sequences in the data have the same 'no prior assumption' condition.

However, what if you want to add some prior knowledge about the sequence to the model? There are two reasons why we may want to do that:

- Increase accuracy: In case we are doing some classic pattern recognition task (classification, regression), having extra useful information will definitely help increasing the model final performance.
- Act as a command: In case of generative model – during the generative mode – we need a way to *trigger* the model in order to start generating a sequence in a particular context. For example, we want to tell the model to generate letter 'A'. Initializing the model with this value allow it to start generating letter A.

There are multiple ways to bias the model, all targeting the same thing: initializing the hidden state of the model. Some work in the literature combines multiple of these approaches in the same setup. To the best of my knowledge, there is no one place which all these methods are discussed together.

Initialize the first hidden state directly This is a common approach, used image captioning (Karpathy and Fei-Fei, 2015), machine translation, and sequence-to-sequence autoencoders. This is illustrated in figure 3.61.

Using the first time-step This method was used in the work done in (Vinyals et al., 2015), in the area of image captioning. The prior in this case is the image, and the objective is to generate the text caption for it. In order to condition the model, the authors projected the image information into the same size as the word embedding used, thus creating a *fake* word, and concatenated this new word with the rest of the words. This is similar to have a word at X_{-1}. The hidden state after this fake word is now condition on the information from the image. This is illustrated in figure 3.62.

Using context sequence – multiple time-steps – In this approach, the model is provided with some time-steps from the ground truth, in give it a context. At the end of the these given time-steps, the hidden state is initialized with information about what to be done. A use case for this scenario – for example – is when training a language model on multiple authors. When asking the model to generate, you can provide some sentences from the author you want, so the model can follow on this.

Concatenating input time-steps with the condition In the work done by (Ha and Eck, 2017), although not mentioned explicitly, the dataset used – the *QuickDraw* dataset, discussed earlier – is quite complicated – they choose different tasks than us –. It is hard to make the model remember the information about such a complex task over a long time span. Thus, the authors use a mix of *initializing the first hidden state* and *concatenating with the first time-step* in order to make it easier for the model to remember the task. This is illustrated in figure 3.63.

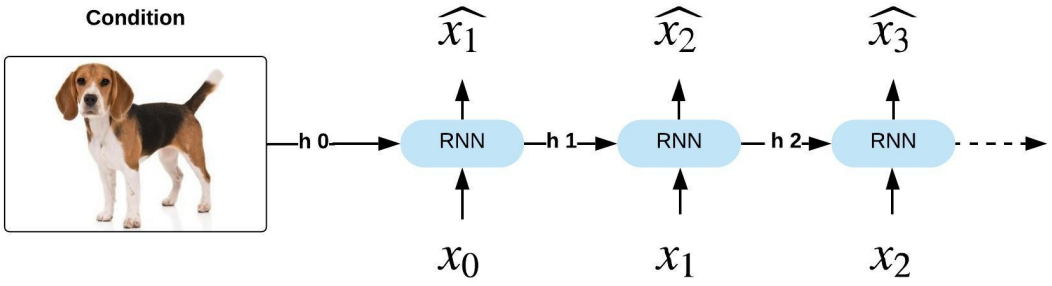
Concatenating the hidden state with the condition This approach is more popular now, since it allows the use of *attention mechanisms* (Denil et al., 2012; Larochelle and Hinton, 2010). Examples for this in image captioning (Xu et al., 2015a), speech synthesis (Wang et al., 2017b).

3.1.6 How to evaluate the quality of generation?

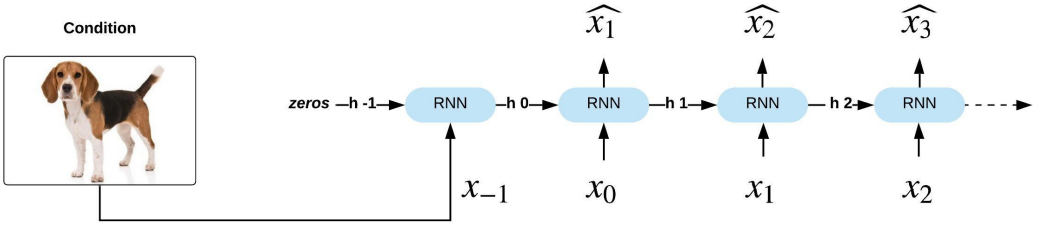
The objective evaluation of a generative model is a challenging task, since there is no consensus for objective evaluation metrics. We can not simply make strict comparisons between the generated data and the ground truth, since, as we saw earlier in the inference section, there is asymmetry between the training and the generation task. While a subjective evaluation can be a workaround to this problem, it is quite slow, expensive and hinders the development of new models and approaches. The need for objective evaluation that are consistent with the relevant subjective evaluation is thus a necessity.

Let's see how this problem is approached in different domains:

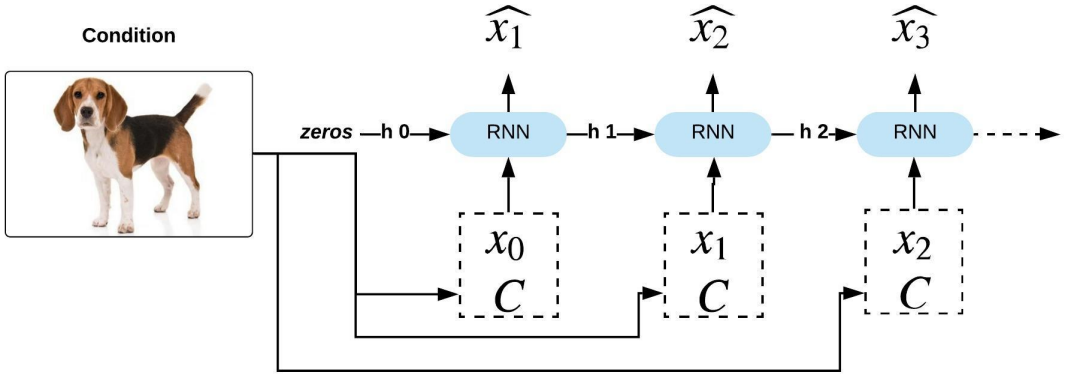
Text Evaluating text is an essential task to assess the quality of generation in many applications, like image captioning, language translation and language modeling. An example for an objective text evaluation metric is *BLEU score* (Papineni et al., 2002), which stands for *bilingual evaluation understudy*, it is a well known metric to evaluate text generation applications, like image captioning (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015) and machine translation (Sutskever et al., 2014b). It tries to capture the human evaluation criteria for the quality of the generated



1 Initializing the hidden state, similar to the one used in (Karpathy and Fei-Fei, 2015).



2 Initializing using the first time-step, similar to the one used in (Vinyals et al., 2015).



3 Concatenating the condition with time-step, similar to the one used in (Ha and Eck, 2017).

Figure 3.6: Different conditioning method for RNN

text. It compares individual generated segments – the size of the segments is a parameter to set – to see if they exist in the ground truth. It does not focus on the location of that segment, just the fact if it exists or not. The small segments – letters or words – are used in order to measure the *adequacy* of the generation, while the longer segments are used to measure the *fluency* of the generation. Usually, the metrics is reported on segments from one to four words, to give and overall idea about the quality of the system.
Other objective metrics were developed and commonly used, such as *METEOR* (Denkowski and Lavie, 2014) and *Word Error Rate* (Klakow and Peters, 2002).

Speech synthesis We are not aware of commonly used objective metrics in this area. A commonly used subjective criteria is Mean opinion score (MOS), which is the average of individuals' opinions on the quality of the system. As an example, the *Blizzard Challenge* (of ISCA , the International Speech Communication Association) is an annual competition in speech synthesis, to encourage the development of better speech synthesizers. The evaluation is much more detailed that MOS, by trying to assess multiple levels for speech, like intelligibility, naturalness, utterance and the efficiency of the speech.

Handwriting of offline Chinese letters Chang et al. (2018) proposed two metrics to evaluate the quality of their generated letters: content accuracy and style discrepancy. For the first metrics, they train an *evaluator* model on the ground truth data, and use it to recognize the letters produced by their generator. For the second metric, they follow an approach developed in Gatys et al. (2015), where the authors studied the problem of image style transfer, by measuring the correlation between different filter activations (in convolution neural network) at one layer, which represents the style representation.

I would like to phrase the objective from these metrics here as *the desire to capture the distance between the generated and the ground truth distribution*. It is not important if some individual mistakes happens in the generated sequence, what matters is to capture the essence of the ground truth distribution. That being said, for complex distribution, capturing the distribution, or even measuring the distance between two distributions, is not always a tractable task. As noted in the speech synthesis point, sometimes it is better and more practical to consider multiple criteria in the same time in order to better understand and evaluate the behavior of the system. This point will later reflect our decisions concerning the evaluation criteria for our work.

Another way to perform the evaluation seen in the literature – like in (Wang et al., 2018) – is to use a machine learning model (sometimes called the *Oracle*), trained on the task needed (recognizing the speaker in the synthesized sound for example). In that case, the model is trained on the ground truth, and used in order to evaluate the samples generated by the model. To put simply, I strictly disagree with such approach, for two reasons:

Improper data distribution It violates the rules of machine learning: if the model is

trained on some distribution, then we expect it to perform well on that distribution. What another model generates is simply a different distribution from the ground truth (even though we want this distance to be as close as possible). Thus, the model behavior in this case is not covered in the statistical learning theory. If the oracle – on the ground truth – has an accuracy of 90%, and then we change the data distribution, we can not really trust the prediction of the system, it is no longer the 90% percent, and we do not know what it is.
The problem continues when we have two generator, and we want to determine which is best, using an oracle, then we can not really compare them. If one generator achieves 80% and the other is 85% accuracy, we can not make a conclusion about which is better. The process itself is flawed. Besides, we do not have access to confidence level on the quality of the oracle, thus it is not possible to perform such comparison using the oracle.

Adversarial problem The problem gets worse when we use the oracle in order to evaluate the generator, then use this value as feedback to the generator in order to "improve" it. We experimented on this part to better understand it. Check appendix C for more details.

The takeaway message here is: it is not about numbers and face value. It is about what these numbers actually means and implies.

3.2 Putting it all together

In the previous section, we explored multiple building blocks for the work to come, like recurrent neural networks (architectures, training, inference and conditioning). We also discussed the issue of evaluation the output of generative models, discussing three aspects: evaluation in case of text (image captioning, translation, and text generation in general), speech synthesis, and the dangers of using another model (oracle model) in order to evaluate the generation quality.

In this section, I discuss how to put all these elements together, the decisions made, and the experimental setup used, in order to address the following three questions:

- How to generate handwritten letters using deep learning framework?
- How to evaluate the generated traces?
- What benchmarks are suitable to compare to?

3.2.1 Our proposed evaluation metrics

Evaluation is a challenging problem when using generative models. We want metrics to capture the distance between the generated and the ground truth distributions. One of our contributions is to propose using the following metrics:

BLEU score As mentioned earlier, BLEU score is used in the evaluation of text generation. Since we discretized the letter drawings, this fits nicely within our work. The general intuition is the following: if we take a segment from the generated letter, did this segment happen in the ground truth letter? We keep doing this for segments of increasing length (the length of the segment here is the number of grams used in the BLEU score). For our work, we report the results on segments from 1 to 3 time steps.

Each part of the letter has two parallel segments: freeman codes and speed, thus, we report the BLEU score for both of them. The equation to compute the BLEU score is the following:

$$BLEU_N = \frac{\sum_{C \in G} \sum_{N \in C} Count_{Clipped}(N)}{\sum_{C \in G} \sum_{N \in C} Count(N)} \tag{3.7}$$

$$Score_N = \min(0, 1 - \frac{L_R}{L_G}) \prod_{n=1}^{N} BLEU_n \tag{3.8}$$

where: G is all the generated sequences, N is the total number of N-grams we want to consider. $Count_{Clipped}$ is clipped N-grams count (if the number of N-grams in the generate sequence is larger than the reference sequence, the count is limited to the number in the reference sequence only), L_R is the length of the reference sequence, L_G is the length of the generated sequence. The term $\min(0, 1 - \frac{L_R}{L_G})$ is added in order to penalize short generated sequences (shorter than the reference sequence), which will deceptively achieve high scores.

In order to get into the intuition of using such a metric in evaluating the quality of handwriting generation, we can imagine that we are comparing segments of a generated trace to segments in the ground truth letter, by asking the following question: does the segment in the generated trace exist (anywhere) in the original trace? The shorter segments represent *adequacy* (i.e., does the generated trace use the same elementary moves like the ground truth trace). The longer segments represents *fluency* of drawing (i.e., how good the drawing is)[3]. See figure Figure 3.7 on page 72.

[3]This last interpretation of short and long segments is adapted from the (Papineni et al., 2002), which uses them for text translation evaluation.

Figure 3.7: Using BLEU score of different sizes, we compare segments of variable length in the generated trace to the target trace.

End of Sequence The length of the letter is another aspect of the style. The distribution of length in the generated examples should follow the ground truth examples. In order to perform this analysis, we compute *Pearson correlation coefficient* between the generated examples and the ground truth data.

3.2.2 How to ground the metrics?

We assess multiple methods to condition our handwritten letter generator, and evaluate their ability to capture of writing styles. We know their cardinal order of the power of these methods (depends on the kind of information available to each method). Knowing this information beforehand, we can use it to ground our performance metrics. The methods are:

Letter identity : the letter id only is used as bias. No style information is thus included. The model will try to average over the different example for the same letter. We consider this as a lower baseline.

Letter + Writer identities : the letter id and writer id are used as a bias. Thus, the model has an explicit access information about the writer. This method is expected to perform the best. This model will also serve as a upper baseline.

Image classifier embedding : we train a convolution neural network (CNN) to classify the letters images[4], as shown in figure Figure 3.8 on page 73. We use an intermediate layer as to extract embeddings, that will encode information about the letter images. This model should perform the same or a more performance than using the letter identity only, since it learns to clusters the letters, and there are classification errors. But we expect it to perform less than the letter + writer identities.

Image auto-encoder latent space : we train a letter image autoencoder, using reconstruction error, and use the latent space as a representation of the letter + style. The architecture we use can be seen in figure Figure 3.8 on page 73. The la-

[4]This letter classification task achieves 95.1% classification accuracy, which we consider very good.

tent space encodes the similarity between the letters. This model should perform worse than using the letter identity only, since, while it capture the similarity between the letter images, it does not capture discriminative features about each letter itself.

From this discussion, we can say that cardinal power of the different conditions is:

$$autoencoder < letter \leq classifier < letter + writer \tag{3.9}$$

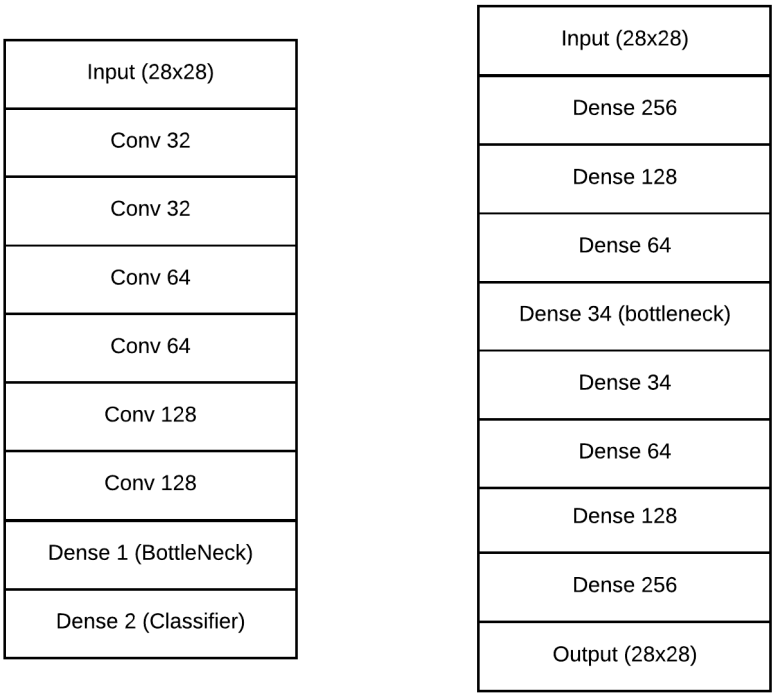
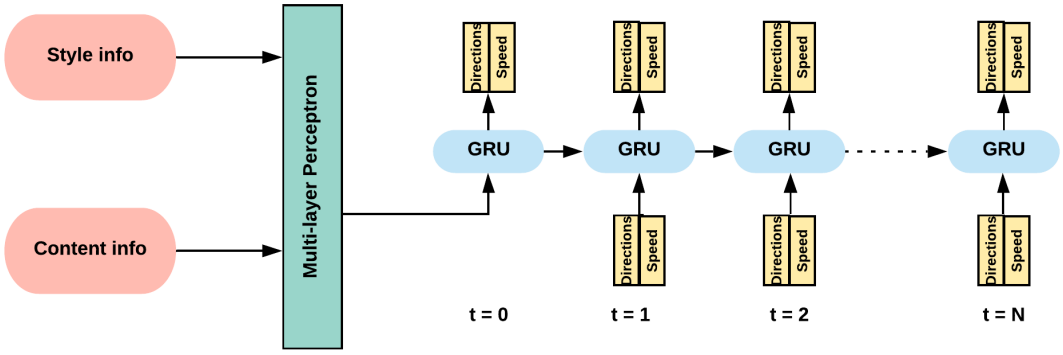


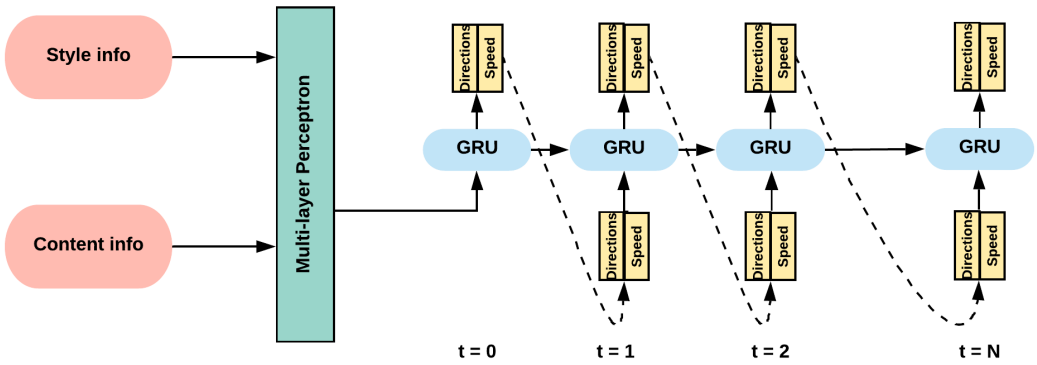
Figure 3.8: Left: architecture of the CNN letter classifier. Batch normalization is used after each convolution layer. The *Dense 1* layer is the embedding that is used to condition our generator. Right: the autoencoder architecture we used. The first *Dense 34* layer provides the latent space used to condition the generator.

3.2.3 Proposed model

We use a type of RNN called Gated-Recurrent Network (GRU) (Chung et al., 2014), which is known for having a better memory ability than basic RNN, thus, making it suitable choice for long sequence. Our model is a conditioned-GRU model, demonstrated in figure Figure 3.9 on page 74. Using this model, we compare different style approached discussed in section 3.2.2, and use the generation results from that model in order to ground the proposed evaluation metrics, discussed in section 3.2.1.



1 Training mode, using *teacher forcing* methodology (Goodfellow et al., 2016; Williams and Zipser, 1989).



2 Generation/Inference mode

Figure 3.9: The conditioned-GRU model used in this work. During the training mode 1, the input of the model is always the ground truth, and the predicted value is compared to the ground truth. During the generation mode 2, the input to the model at each step is the model prediction from the previous step. The 'style info' and the 'task info' inputs are separated here for demonstration (they could be mixed).

3.2.4 Results

We train our different models and generate the traces from them as explained earlier. In this section, we compare the different models using the evaluation metrics discussed before. We observe the consistency of the reported metrics with the prior information about the cardinal power of the different methods, equation (3.9). This is how we ground our metrics.

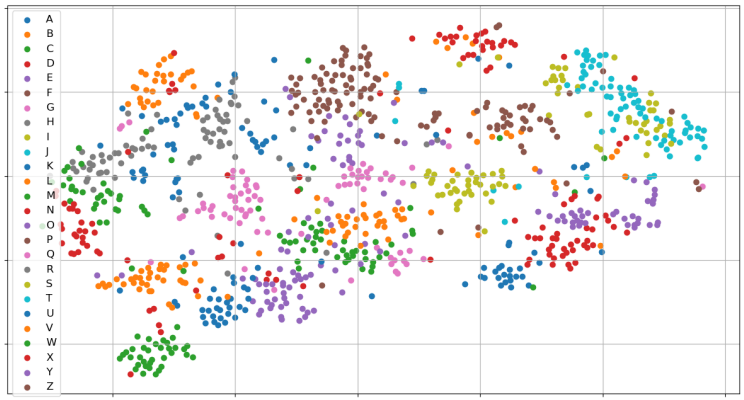
BLEU score

The final results using the BLEU score can be seen in table 1. The results vary when measuring BLEU-1. But, as we increase the number of grams, BLEU-2 and BLEU-3, to measure the similarity between larger segments of the traces, we can observe:

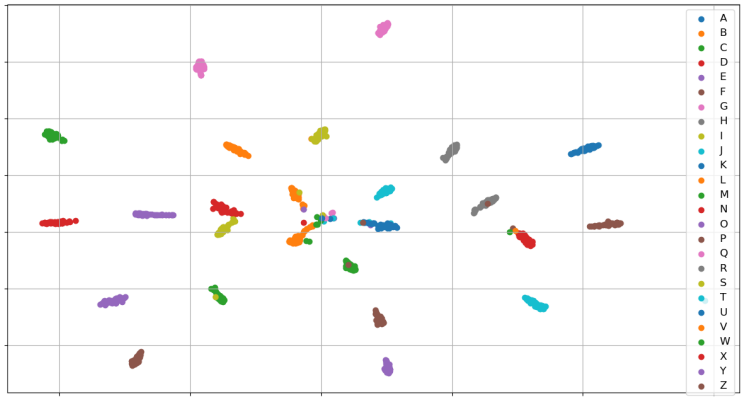
- The letter + writer condition performed better than all other conditions, thus showing that having access to information about the writer, like the writer id, improve the quality of the handwriting synthesis.
- The image classifier condition performs better than the letter identity only, but less than the letter + writer bias. Since the classifier is trained on a single objective only (to classify the letters), and the classifier performs well, we expect the embedding to cluster the letters well, as seen in figure 3.102. We can expect the model to capture some of the writer style, possibly in the inter-cluster variance. This is an interesting result, suggesting that some fine tuning for the image classifier while in the generation task could be beneficial to capture more details about the styles.
- The image autoencoder bias performed the worst. To understand why, we plot a 2-D projection of its latent space using t-SNE (van der Maaten and Hinton, 2008), figure 3.101. Since the autoencoder is trained to minimize the reconstruction error, the distance in the latent space encode the proximity between the images. It can be observed also that this latent space does not encode discriminative features for the letters. Using this latent space for our generator, we find the model gets confused between nearby letters, resulting sometimes in generating different letters than requested.

Aspect/Feature	Speed			Freeman		
Model / B-score	B-1	B-2	B-3	B-1	B-2	B-3
Letter identity	49.7	37.3	24.2	47.4	36.6	26.8
Image classifier	50.9	38.2	24.6	48.5	37.9	28.1
Image autoencoder	**51.9**	37.9	23.1	46.4	35.0	24.5
Letter + Writer identities	51.5	**41.4**	**25.1**	**56.7**	**39.4**	**28.3**

Table 1: Comparing different approaches for style extraction using clipped n-grams



1 Bottleneck of the auto-encoder



2 Bottleneck of the classifier

Figure 3.10: Figure 1 shows the autoencoder latent space, there is no clear separation between letters; the encoding is based on the similarity of the images only, while figure 2 shows the classifier embedding, there is a clear separation between the letters - with few exceptions -.

Sequence length

As mentioned earlier, we performed a statistical test between the paired distributions of lengths of the generated and the reference tracings. The results are shown in table 2. We can see the following:

- The results from the statistical test shows that the letter + writer bias outperform the rest of the biases, achieving p-value < 0.05. This is quite reassuring, since it is also in line with the results from the BLEU score.
- The results from the Pearson correlation coefficients are also consistent with the rest of the results. High coefficients are given to the letter + writer bias, compared to the other methods. The image classifier and autoencoder gives the lowest results. This could be due to insufficient information about the letter length that can be inferred from the image. For the image classifier, as noted earlier, a fine-tuning during the generation task is worth exploring.

Models	Pearson coefficient	p-value
Letter bias	0.38	0.84
Image classifier	0.32	0.62
Image autoencoder	0.25	0.29
Letter + Writer bias	**0.55**	**0.04**

Table 2: Pearson correlation coefficients and associated p-values for the EOS distributions of the different style biases.

3.2.5 Examples of the generated letters

The design choices of our experiments (discretization, and ignoring the pen state) affects the final shape of the letters, yet, the letters and their style are quite recognizable. See examples for the original letters in figure Figure 3.11 on page 78. Examples for the generation with our methods are in figure Figure 3.12 on page 79. This is a subjective indication that our model is working properly, producing real-like comprehensible letters.

3.3 Summary

In this chapter, we sit the foundations for the work done in my thesis. I first discussed the relevant areas in the state-of-the-art concerning recurrent neural networks, optimization, inference/generation and evaluation of generation. I then detailed how we

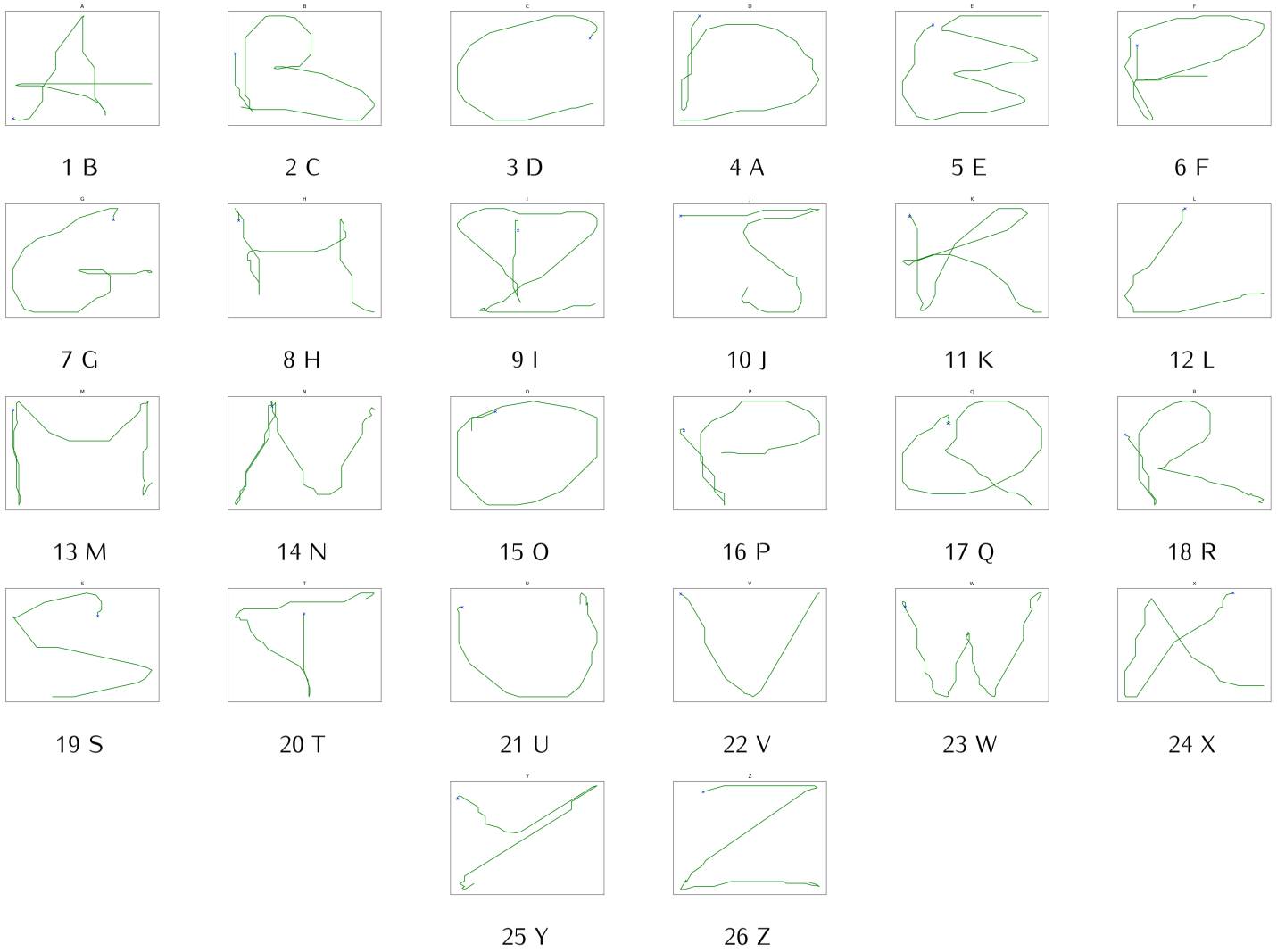


Figure 3.11: Examples of original letters. The blue *x* mark is the starting point. These ones are generated using the letter + Writer bias. E and F are visually harder to recognize, since we do not model the pen pressure, otherwise, the rest of the letters are well recognizable.

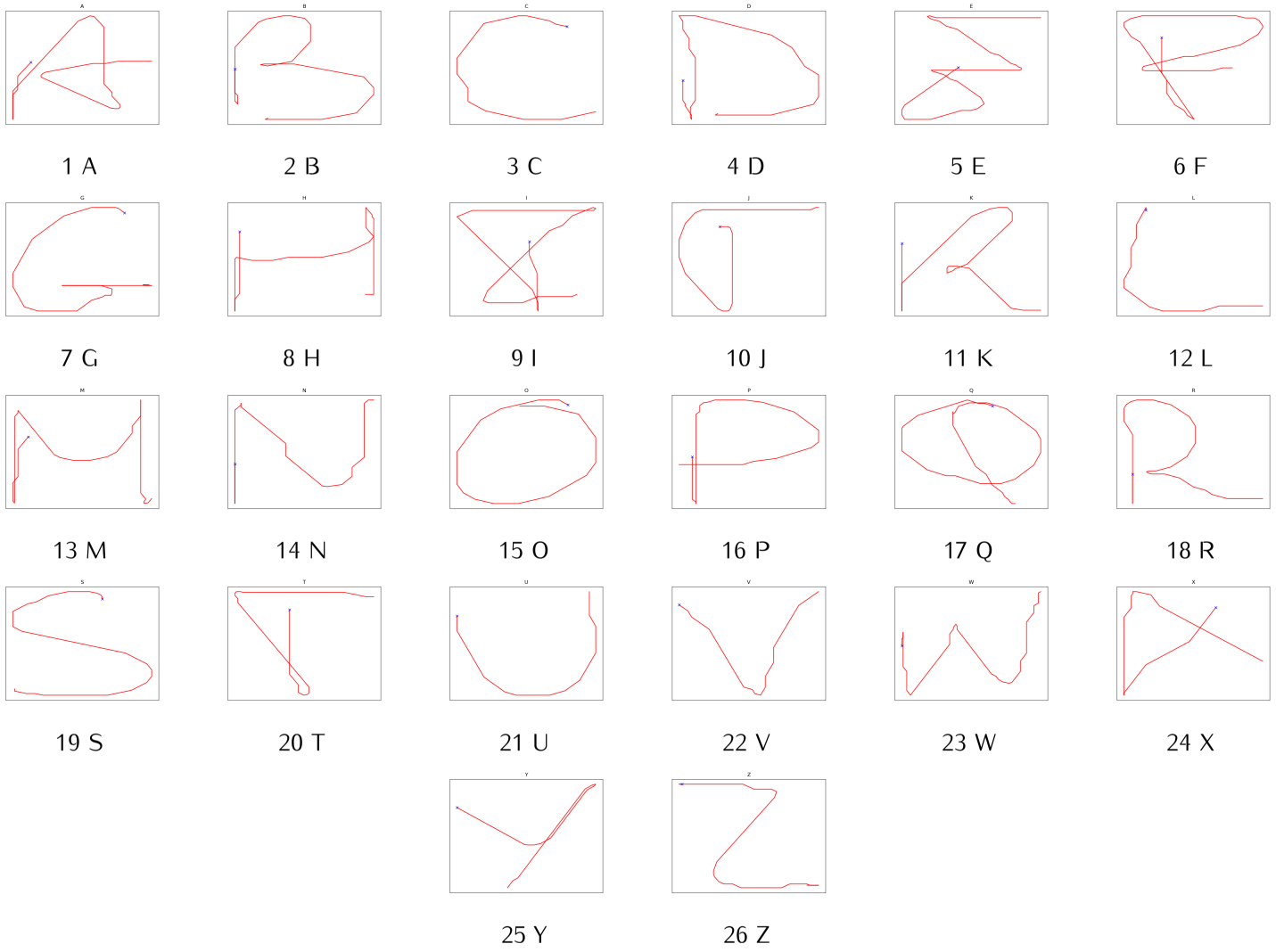


Figure 3.12: Examples of generated letters. The blue *x* mark is the starting point. These ones are generated using the letter + Writer bias. The general quality of this quite acceptable.

combine these elements in order to address the questions in this thesis. We addressed three points:

- How to generate traces using deep generative models?: we proposed to use a conditioned-GRU model. I explained the process behind training, optimizing and choosing hyper-parameters for this model, and showed some of the generated letters.
- How to evaluate the quality of the generated letters?: we proposed two evaluation criteria, BLEU score metric – inspired by its usage in evaluating text quality – and End-of-Sequence quality, as a way to capture some aspects of the distribution, and a feasible way to measure the distance between two distributions.
- We proposed multiple benchmarks for future evaluation of style extraction models, and to ground the proposed evaluation metrics as well.

Now that we have our benchmarks and evaluation metrics, it is time to go for the next chapter...

Chapter 4

Framework

Contents

In the previous chapter, we set the foundations of studying styles, which is the work done in the PhD. We explored our choices for generative models and the evaluation metrics, and how to ground them. We also had a discussion about lower- and upper-bound benchmarks. These foundations are necessary in order to have baselines to compare to, and performance aspects (i.e., metrics) use for this comparison.

In this chapter, we build on these foundations, by proposing the use of conditional temporal auto-encoder framework in order to study and extract styles, in the context of sequences (i.e., a time aspect exists in the data). We take advantage from the fact that the content is well known in our dataset (i.e., the identity of the letters in *IRONOFF*, and the identity of the shapes in *QuickDraw!*).

Questions addressed in this chapter

- What possible framework to study styles? and why?

- How does this framework performance compare to the benchmarks?
- What kind of styles we can extract from this framework? and how do we extract them?

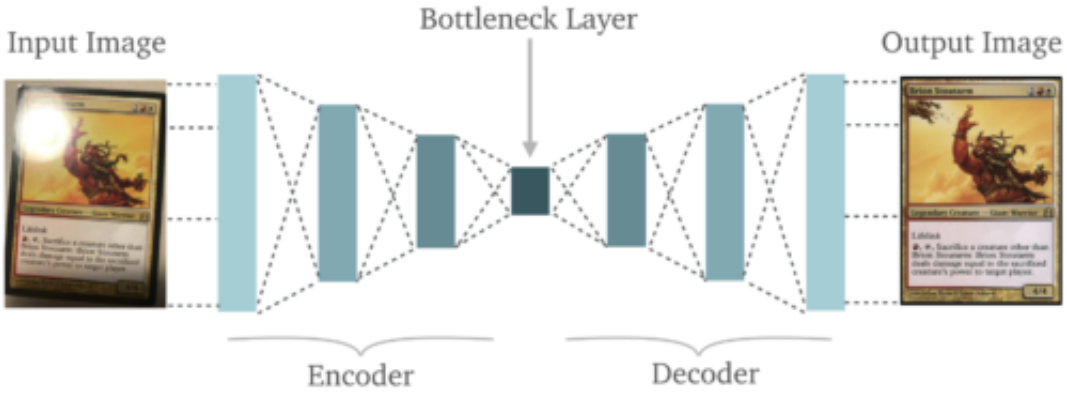


Figure 4.1: An example for the main components of an auto-encoder, used on image compression: an encoder takes the image, transfer it via a set of transformations into a bottleneck code, which is a compressed representation for that image. The decoder then takes this bottleneck code, and apply a series of transformations on it, in order to reconstruct the original input image.

4.1 Background

4.1.1 What is an auto-encoder?

Auto-encoders (Hinton and Salakhutdinov, 2006) is identity-capturing framework, that allow the emergence of interesting behaviors. To expand on this, the objective of the auto-encoder is to capture/learning the distribution of the input data (identity-capturing). An auto-encoder consists mainly of three parts (see figure 4.1):

1. Encoder: it takes the input data, and project it into a manifold (the bottleneck).
2. Bottleneck code/learned representation: this is the representation learned by the encoder. In the basic form, this just the output of some linear/non-linear activations. However, a lot of the literature exists on how to organize and shape the bottleneck, in order to allow the emergence of interesting information about the data.
3. Decoder: it takes the learned representation, and reconstruct the original input from it.

There are many reasons for using auto-encoders, to name a few:

Dimensionality Reduction It is one of the main motivations to study auto-encoders. By mapping a high-dimensional data into a smaller low-dimension space, we can better explore the data, or use it as a step in a pipeline of machine learning/data analysis operations, where it provides a more manageable format of the original

data, as in (Ha and Schmidhuber, 2018). It can also be used in classification system (Goodfellow et al., 2016).

Information Retrieval The kind of search used in information retrieval is quite efficient in low-dimensional space. In (Salakhutdinov and Hinton, 2009), auto-encoder is trained to produce low-dimensional binary code, which can be then used in queries (by returning the entities that have the same binary code).

Anomaly Detection Multiple works have explored the use of auto-encoders in order to perform anomaly detection (An and Cho, 2015; Ribeiro et al., 2018; Sakurada and Yairi, 2014). The main hypothesis is that the auto-encoder will learn the most salient features in the data. Thus, when faced with an anomaly, a significant degradation in the quality of reconstruction will be noticed. The reconstruction error can be used in this case as an indicator if the input data point is an anomaly or not.

Image De-noising Multiple works have explored the use of auto-encoder in order to de-noise images (Cho, 2013a,b; Gondara, 2016). The applications of image de-noising are diverse, from post-processing of digital images, to more sensitive areas like medical images.

The idea of compressing data is not new. Techniques like *Principal Components Analysis* (PCA) (Jolliffe, 2011) do exist in order to project the data into smaller dimensions. But they are usually restricted by several assumptions. In case of PCA, it assumes linearity and orthogonality in the dimensions of variation in the data. Neural networks enables us to get around this issue, by leveraging nonlinearity and multiple layers, this giving us a more flexible approach to find dimensions of variation in the data.

4.1.2 Sequence auto-encoder

It a special case of auto-encoder, where RNN is used in order to compress a sequence into a fixed size bottleneck code. The sequence itself could have a varied size. The first architecture proposed for sequence auto-encoder was proposed in (Cho et al., 2014; Sutskever et al., 2014a), with statistical machine translation as the main application. They use RNN encoder and decoder parts, and consider that the last hidden state of the RNN encoder to be the summary/compression of the sequence. The RNN decoder uses this bottleneck in order to reconstruct the whole original sequence (see figure 4.2).

There many applications where a sequence-to-sequence model is used, for example:

Machine translation (Sutskever et al., 2014a) used sequence-to-sequence architecture in order to develop a machine translation system that surpassed the published results to that moment. The first sequence is the language of origin, and the

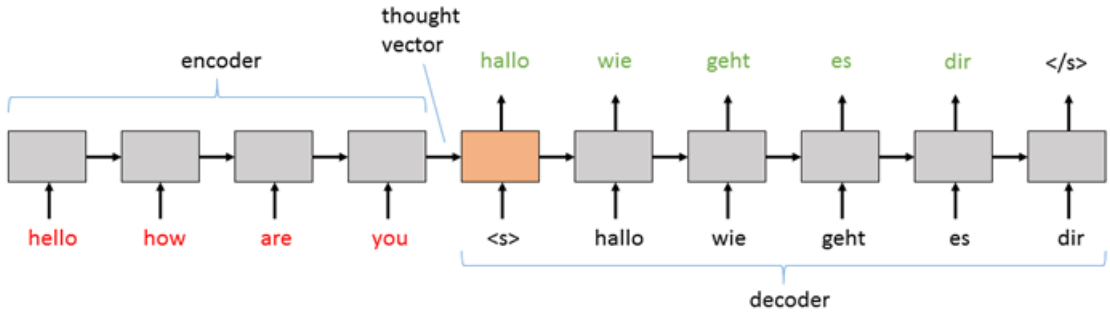


Figure 4.2: An illustration for a sequence-to-sequence architecture, used for language translation between English and German. The encoder summarize the English sentence, and the decoder use it as to bias its own output, to generate the equivalent German sentence.

second sequence is the target language.

Speech synthesis Speech synthesis also benefited from a sequence-to-sequence architecture, like the work done in in (Oord et al., 2016a; Wang et al., 2017b), achieving currently the state-of-the-art results. Another common approach for sequence-to-sequence learning is to use convolution network instead of RNN for the encoder and/or the decoder, like the work done in (Ping et al., 2017). A convolution network is generally faster than RNN and parallelize better, thus making it a lucrative approach. The underlying assumptions are the same however.

Video captioning Another application that benefited from sequence-to-sequence architecture is generating text (sequence of words/letters) that describe a video (sequence of frames), as nicely summarized in (Aafaq et al., 2018). In this case, the encoder – dealing with the video part – is usual a convolution neural network (because of its excellent ability to capture spatial features), while the encoder – dealing with the text part – is usually a RNN.

4.1.3 Conditioned auto-encoder

In the examples mentioned before, we focused on unconditional auto-encoder, where in the decoder only have the information given to it from the encoder part. A conditioned auto-encoder is when we concatenate extra information to the output of the encoder, and feed it to the decoder[1]. Why conditioning an auto-encoder? it frees the encoder from learning the condition information – since this information is given for free – ,

[1]The word 'conditioned' is used when a neural network has information from about the task. So a decoder is a conditioned network on the encoder information. We distinguish here in the terminology between 'unconditioned auto-encoder', where the decoder is not conditioned on anything else except the encoder, and 'conditioned auto-encoder' where the decoder has access to extra information other than the encoder.

allowing it to focus on other parts.

An example of this is the work done in (van den Oord et al., 2017), where they used an auto-encoder to compress audio, and condition the decoder on the speaker-id. This led the encoder to factor out speaker-specific information in the learned bottleneck, thus, learning speaker-independent information[2].

4.2 Putting it all together

In order to address the research questions stated at the beginning of the chapter, we chose to adopt the concept of conditioned auto-encoder as our framework to explore styles in handwriting, for the potential following benefits – discussed in the previous section –:

- Conditioning the decoder on the content identity of the task (i.e., the letter identity) will free the encoder from learning this information, thus allowing it to focus on learning the letter-independent style relevant information.
- The encoder will learn a bottleneck, that is a compressed information about the sequence style. This can allow us to explore this bottleneck via traditional techniques (PCA, tSNE, clustering, classification of the bottleneck...etc), thus, getting more insight about what the model actually learned

In this following subsections, I will present the exact model architecture we used. I will then discuss the quality of the generated lettered, using the evaluation metrics we discussed in the previous chapter. Then, we will briefly take a look at our first attempt to tackle transfer learning[3]. We then end with the style extraction part, where we explore what knowledge/information about the styles our model has extracted.

4.2.1 Model architecture

The model architecture is illustrated in figure 4.4. The input/output frames of the model are detailed in figure 4.3. The trace of the letter is first fed to encoder module. The final hidden state of that module summarizes the letter. In order to allow this module to focus on learning the style embedding, we complement this last hidden state with the one-hot encoding of the letter identity, and use a projection of them as the bias input to the generator. The encoder thus is free from the need to learn the letter

[2]Simple explanation and demonstration for that paper can be found in `https://avdnoord.github.io/homepage/vqvae/`

[3]More details on transfer learning in the next chapter

identity, and can focus learning the style information that enables the generator to better approximate the ground truth tracings.

In the decoder, we follow the framework proposed by Vinyals et al. (2015) in order to bias the model – as in the previous chapter – : we create an extra time step at the beginning, which has the information we want to bias the model with. In this case, this time step is the projection of the encoder last hidden state and the letter encoder. This has a much lower dimension than encoder hidden state. This further encourage the model to learn only necessary style information, as suggested in Skerry-Ryan et al. (2018).

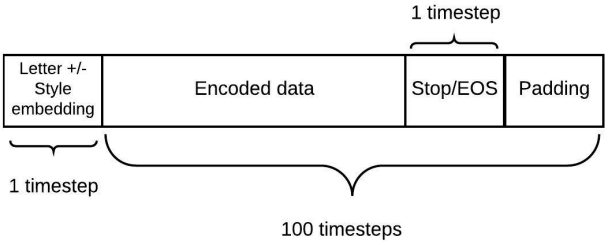


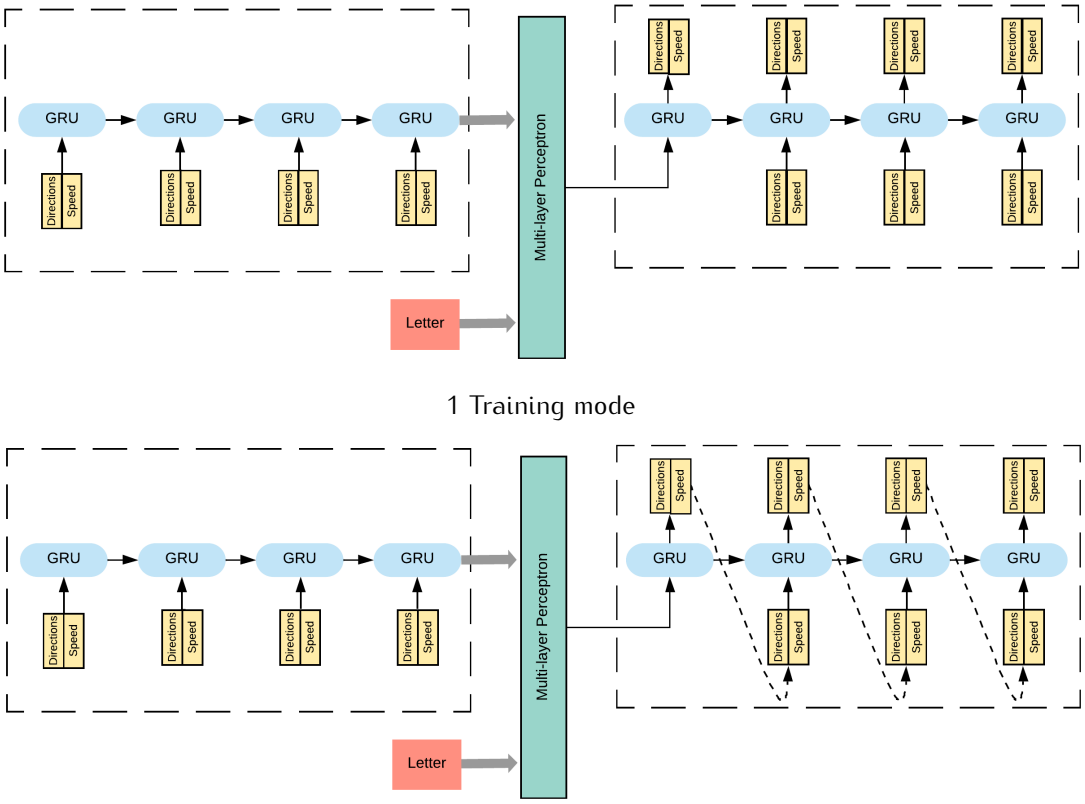
Figure 4.3: Input sequence to our model. The first time step contains the information necessary to condition/bias our model. In case of the encoder, this first time step (the bias) is not included.

4.2.2 Letter generation with style preservation

The objective here to compare the quality of the generated letters to the state-of-the-art benchmarks. As mentioned earlier, we compare using the BLEU score metric and the EoS analysis. The BLEU score results can be seen in table 1, and the results for EoS analysis results are in table 3. We can see that the BLEU-3 score results of our model achieves 32.3% accuracy in Speed feature and 38.7% accuracy in Freeman feature, compared to 25.1% and 28.3% accuracy using the benchmark model on both features respectively.

The same goes for the EoS analysis. In comparing the Person Coefficient, our model achieves 0.99 score compared to 0.55 for the benchmark model (the highest score is 1.0). This is a support that our model capture the style of handwriting better than the benchmark.

Examples for the generated letters can be found in figure 4.14.



1 Training mode

2 Inference mode

Figure 4.4: Schematic diagram of the model we used. During the training time 4.41, the input to the model is always the ground truth. During the inference time 4.42 however, the input to the decoder (generator) part at each time step is its own predication in the previous time step.

Aspect/Feature	Speed			Freeman		
Model / B-score	B-1	B-2	B-3	B-1	B-2	B-3
Letter + Writer bias	51.5	41.4	25.1	56.7	39.4	28.3
Style Extractor	71	51.7	32.3	65.6	51.5	38.7

Table 1: BLEU scores for different models for known writers.

Aspect/Feature	Speed			Freeman		
Model / B-score	B-1	B-2	B-3	B-1	B-2	B-3
Letter + Writer bias	55.4	39.6	25.3	50.2	38.6	27.7
Style Extractor	72.4	52.4	32.2	70.4	55.6	42.1

Table 2: BLEU scores for different models for style extraction for 30 new writers (style transfer).

Models	Pearson coefficient
Letter + Writer bias	0.55
Style Extractor	0.99

Table 3: Pearson correlation coefficients for the End-Of-Sequence (EoS) distributions for the different models on the normal gene ration scenario

Models	Pearson coefficient
Letter + Writer bias	0.5
Style Extractor	0.99

Table 4: Pearson correlation coefficients for the End-Of-Sequence (EoS) distributions for the different models on 30 new writers (style transfer).

4.2.3 Style transfer

One of the hypotheses we want to test is whether there is a limited number of styles needed, to generalize over new writers. To achieve this, the learned representation for styles should extract generic information about the styles.

In order to test this hypothesis, we expose our model to 30 writers that have not been seen before. We compare our model performance on these writers with a model is biased by the writer and letter identities (the benchmark model). The latter model was not constrained from seeing those writers (thus, the reported results of the comparison overestimates the actual performance of that model).

The BLEU scores can be seen in table 2. Our model achieves on BLEU-3 score 32.2% and 42.1% accuracy on the Speed and Freeman code features, compared to 25.3% and 27.7% on the benchmark model for the same features respectively.

The EoS analysis can be seen in table 4. Our model achieves a coefficient value of 0.99, compared to 0.5 for the benchmark. Thus, the new model clearly outperform the current benchmarks on the transfer task, on both BLEU score and EoS analysis.

4.2.4 Styles per letters

One of the nice consequences of using our model is that we can have a better look at the styles. We explore the latent space for multiple letters, and see that we can uncover interesting writing styles. A full scale analysis is beyond the scope of this paper. We project the latent space using *Principal Components Analysis* (PCA) (Jolliffe, 2011) and t-SNE (van der Maaten and Hinton, 2008).

As a start, we take a look at letter X. Beforehand, we identified a style feature in letter X: some writer draw X clockwise, and some draw it anti-clockwise. We manually annotated the whole dataset for this feature; the result can be seen in figure 4.5. Almost half of the writers draw the letter X clockwise, and the other half draw it anti-clockwise. If our assumption is correct, our model should be able to capture this feature. We project the latent of the model using PCA on all the letter X, which can be seen in figure 4.6. The model latent space clusters almost perfectly based on rotation. Examples for letters from both clusters are in figure 4.7.

Encouraged by the results on letter X, we explored more letters. For letter C, we can see the latent space project in figure 4.8. It can be seen that there are at least two main clusters. Examples from this cluster in the red ellipse are in figure 4.10. The indicated cluster represents the Edwardian handwriting style. The rest of the writers (in the big cluster) have a very similar style (this is expected, since the drawing of the letter C is quite simple).

For letter A, our model latent space create two main clusters, figure 4.9. We give examples from those two in figure 4.11, where we can see clear difference in the style. Some people start drawing the letter from down-left, other writers start from the top of letter A, move down, then continue drawing of the letter.

Another example is for letter S bottleneck, figure 4.12. There are three resulting clusters which we investigated. The indicated cluster (in red) is clearly different from the other two clusters (not indicated). Examples can be seen in figure 4.13. The indicated cluster is again for people with Edwardian handwriting style. We did not find a clear difference between the other two clusters though, but this is an expected outcome of using t-SNE (since it does not have the clear objective of clustering styles).

These examples show is that we can use our model to extract verbose style information.

4.3 Summary

In this chapter, we discussed the way we chose in order to study styles, by using the auto-encoder framework. We hypothesized that we can study styles implicitly by looking at how they contribute to the reconstruction/generation of the letters. We looked at the state-of-the-art concerning auto-encoders, the sequence-to-sequence case, and why we may choose to condition an auto-encoder.

I then presented our work, addressing the following points:

- What possible framework to study styles? and how to validate it?: we test the

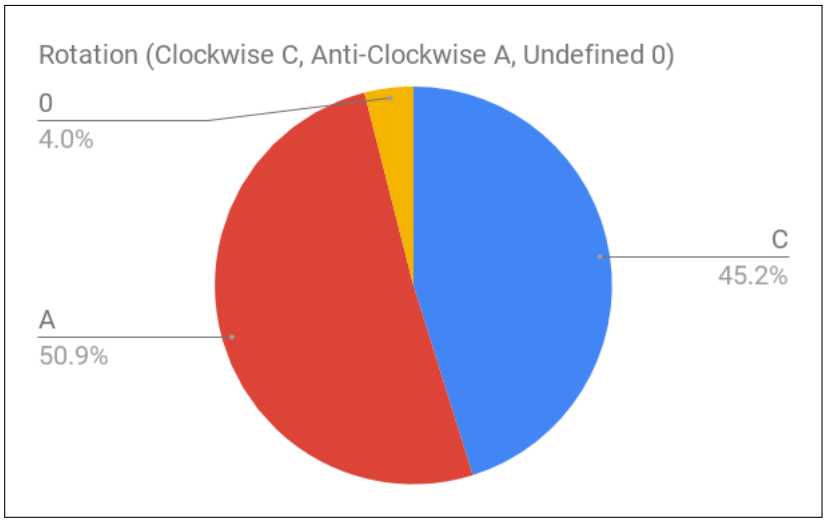


Figure 4.5: Results of the manual annotation for the rotation of letter X drawings over the whole dataset. Almost half the writers drew X clockwise, the other half anti-clockwise. The undefined styles were unclear to determine.

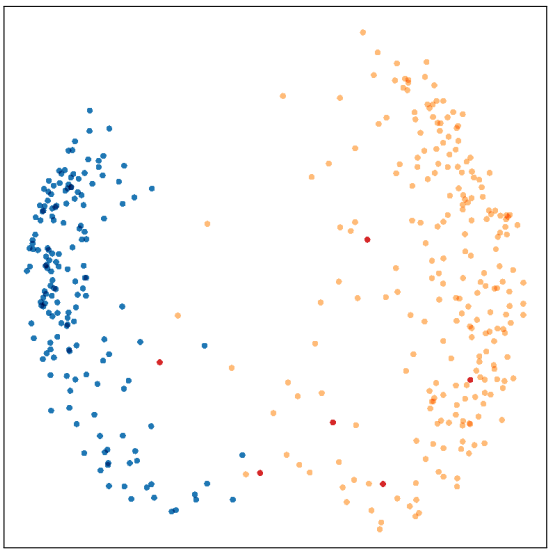


Figure 4.6: Projection for latent space for letter X using PCA. The colors show the ground truth of the X rotation: blue is counter clockwise, orange is clockwise, and the few red points are undefined.

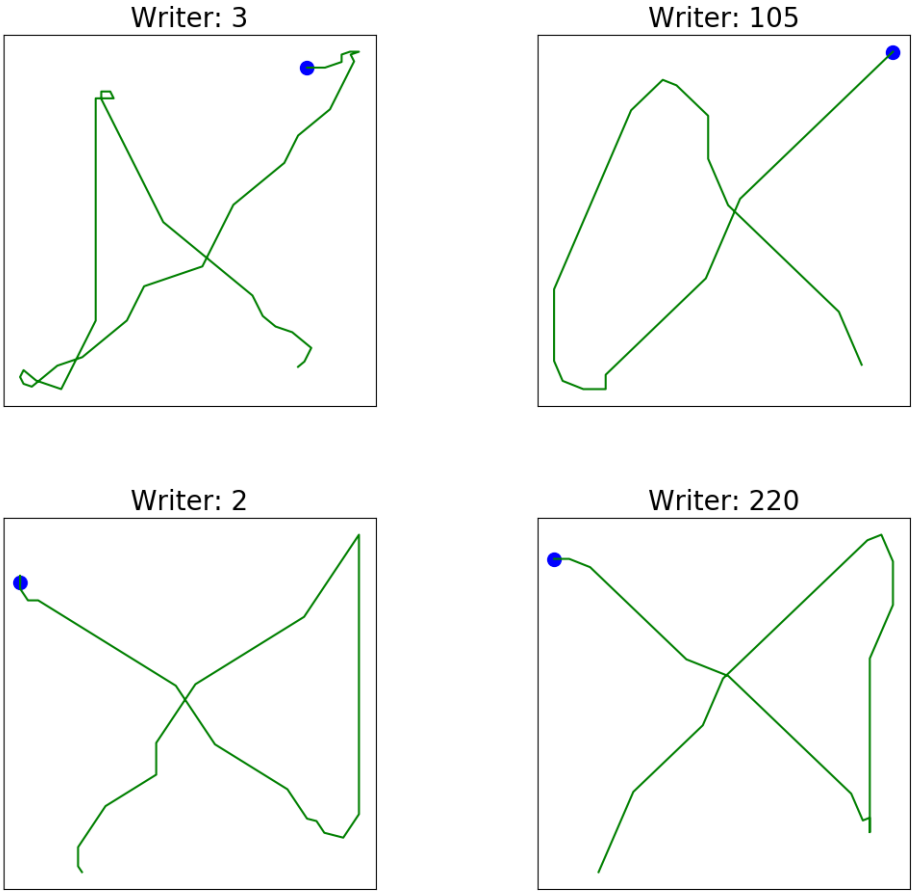


Figure 4.7: Examples for writing of letter X. Starting point is marked with the blue mark. Each raw is randomly sampled from each cluster in the bottleneck. The clusters shows that almost half the writers draw the letter clockwise (first row, first cluster), and the other half draw it anti-clockwise (second row, second cluster).

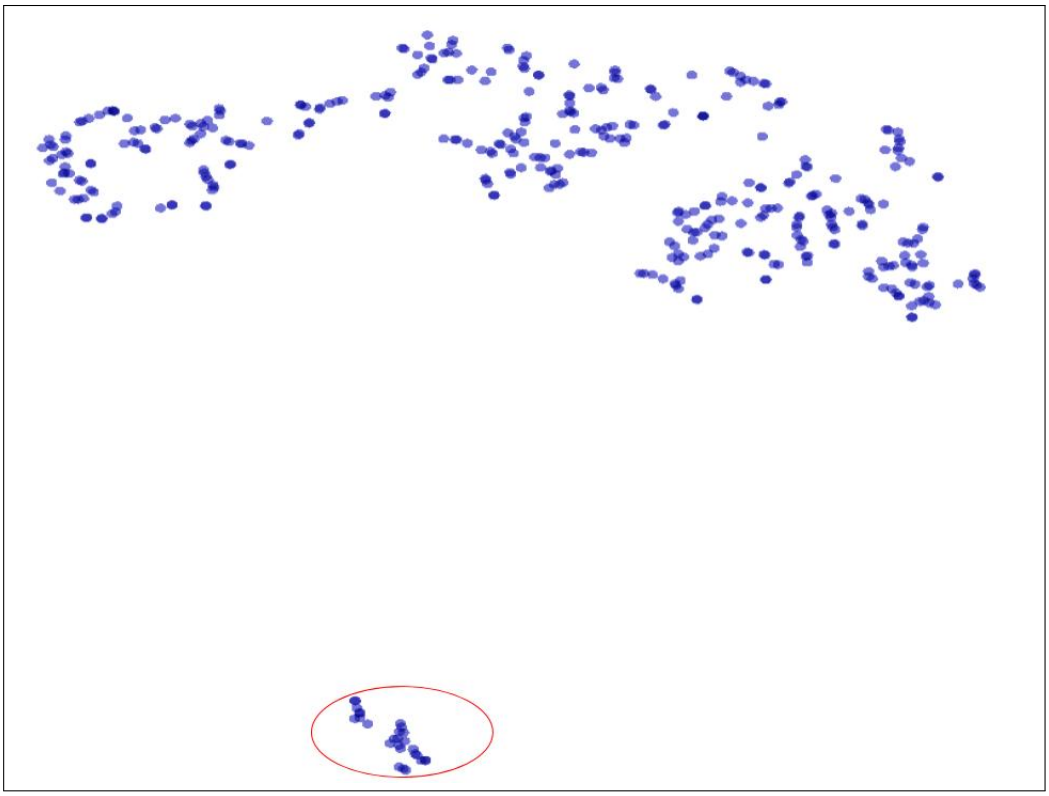


Figure 4.8: Projection for latent space for letter C using t-SNE. The cluster surrounded by the red circle has a clear interpretation, where writers have a cursive style.

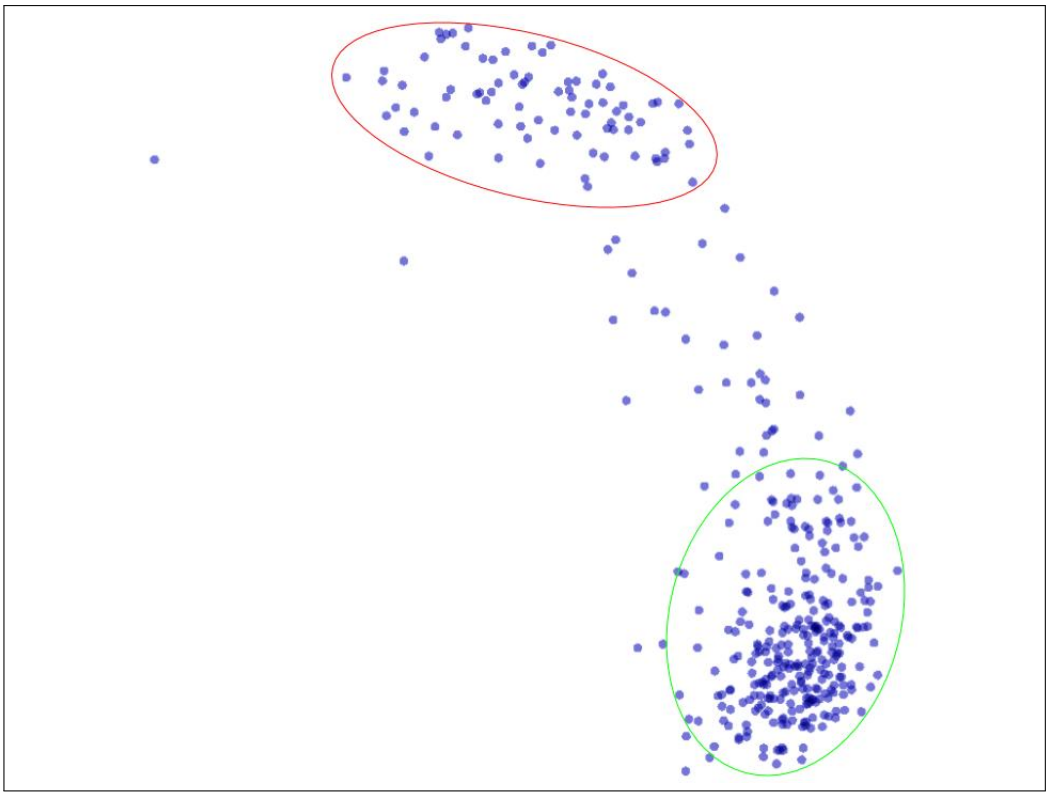


Figure 4.9: Projection for latent space for letter A using PCA.

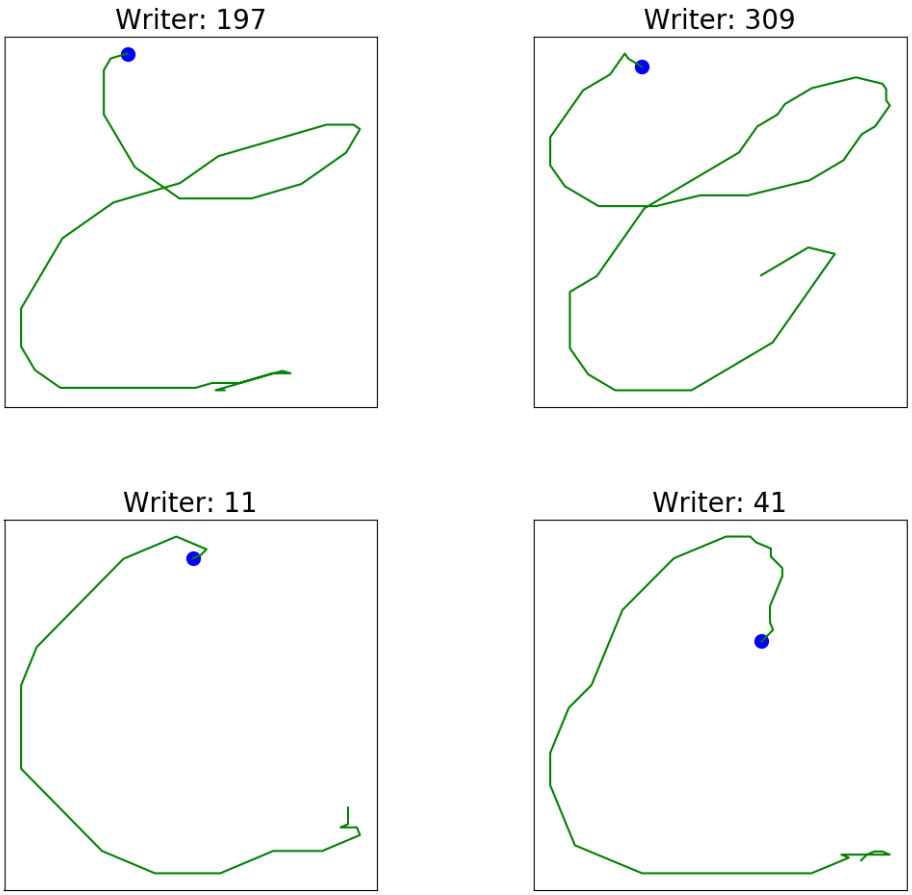


Figure 4.10: Examples for writing of letter C from the selected cluster (first row) versus the rest of the letter drawings (second row). Starting point is marked with the blue mark. The drawings from the selected cluster show people with Edwardian style of handwriting.

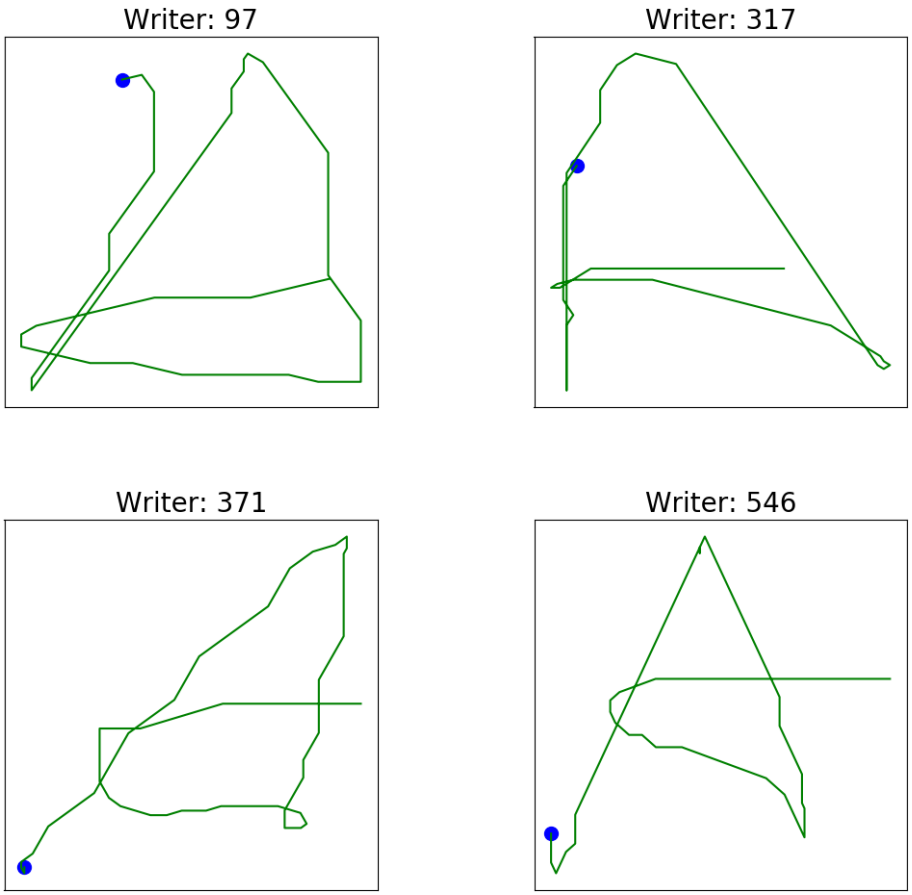


Figure 4.11: Examples for writing of letter A from the selected clusters. Starting point is marked with the blue mark. Each row is from one cluster. The first row show people who start drawing the letter from the top, going down, and then continue the drawing of the letter. The second row show people who start drawing from down directly.

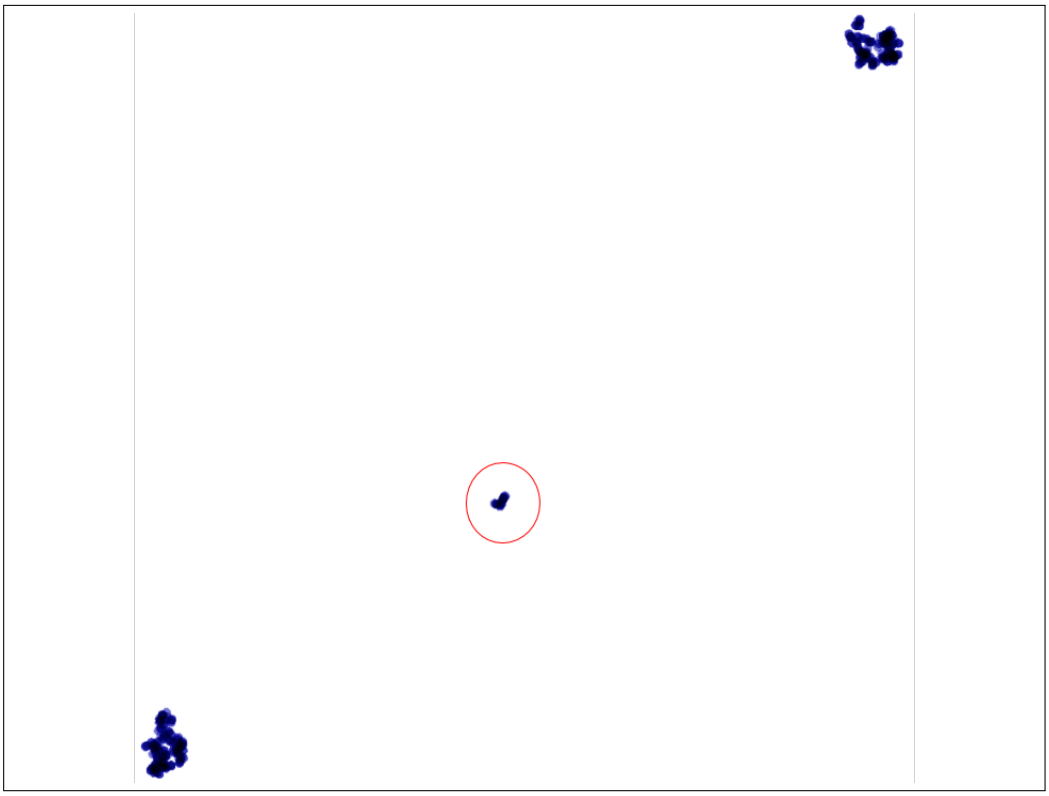


Figure 4.12: Projection for latent space for letter S using t-SNE. We manage to interpret the indicated cluster as the Edwardian style in drawing. The other two clusters (not indicated) did not show clear difference in the style, but this is an expected behavior from using the t-SNE algorithm, since it does not try to cluster styles as an objective.

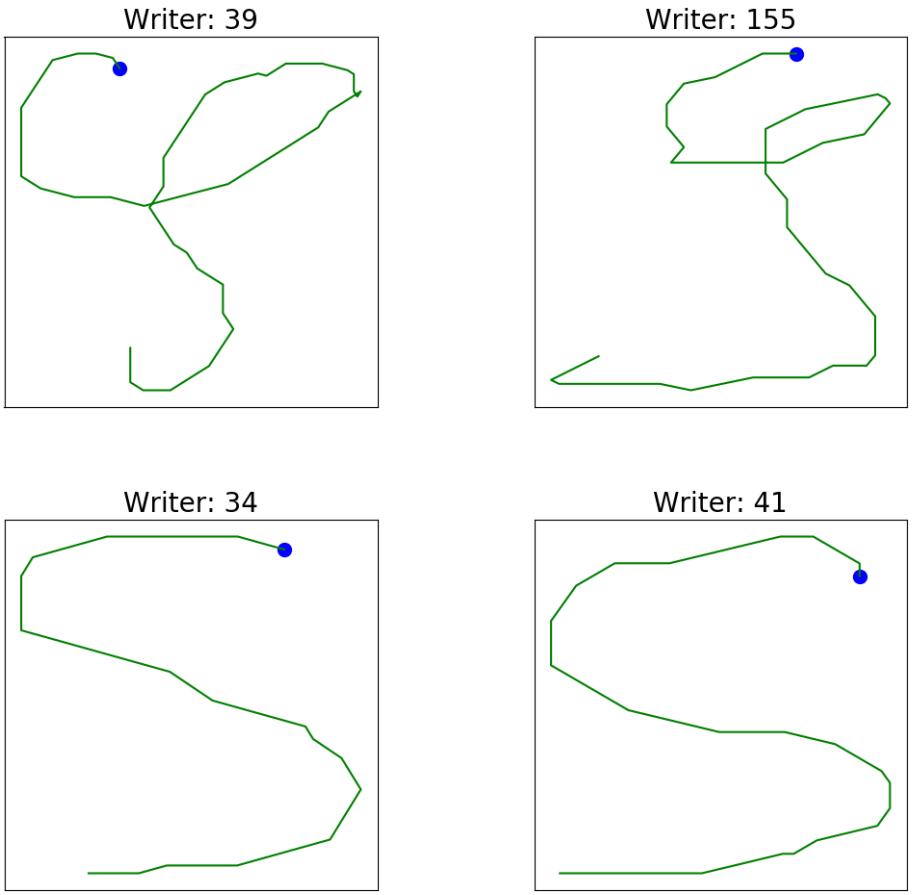


Figure 4.13: Examples for writing of letter S from the selected cluster (first row) versus the other two clusters (second row). Starting point is marked with the blue mark. The drawings from the selected cluster is always Edwardian style.

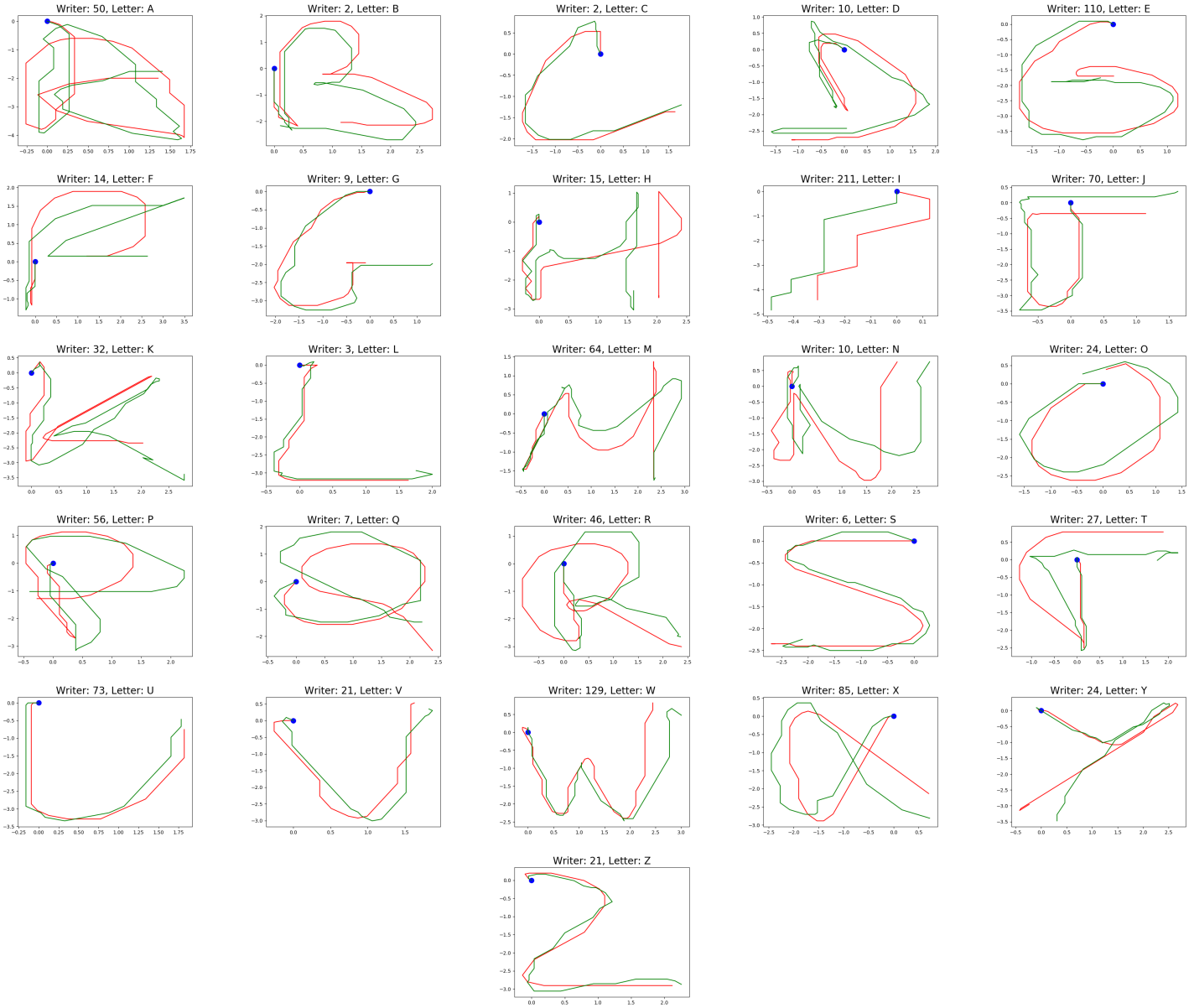


Figure 4.14: Examples of generated letters. The blue mark is the starting point. The traces in green is the ground truth, and the red is the generated ones by our model.

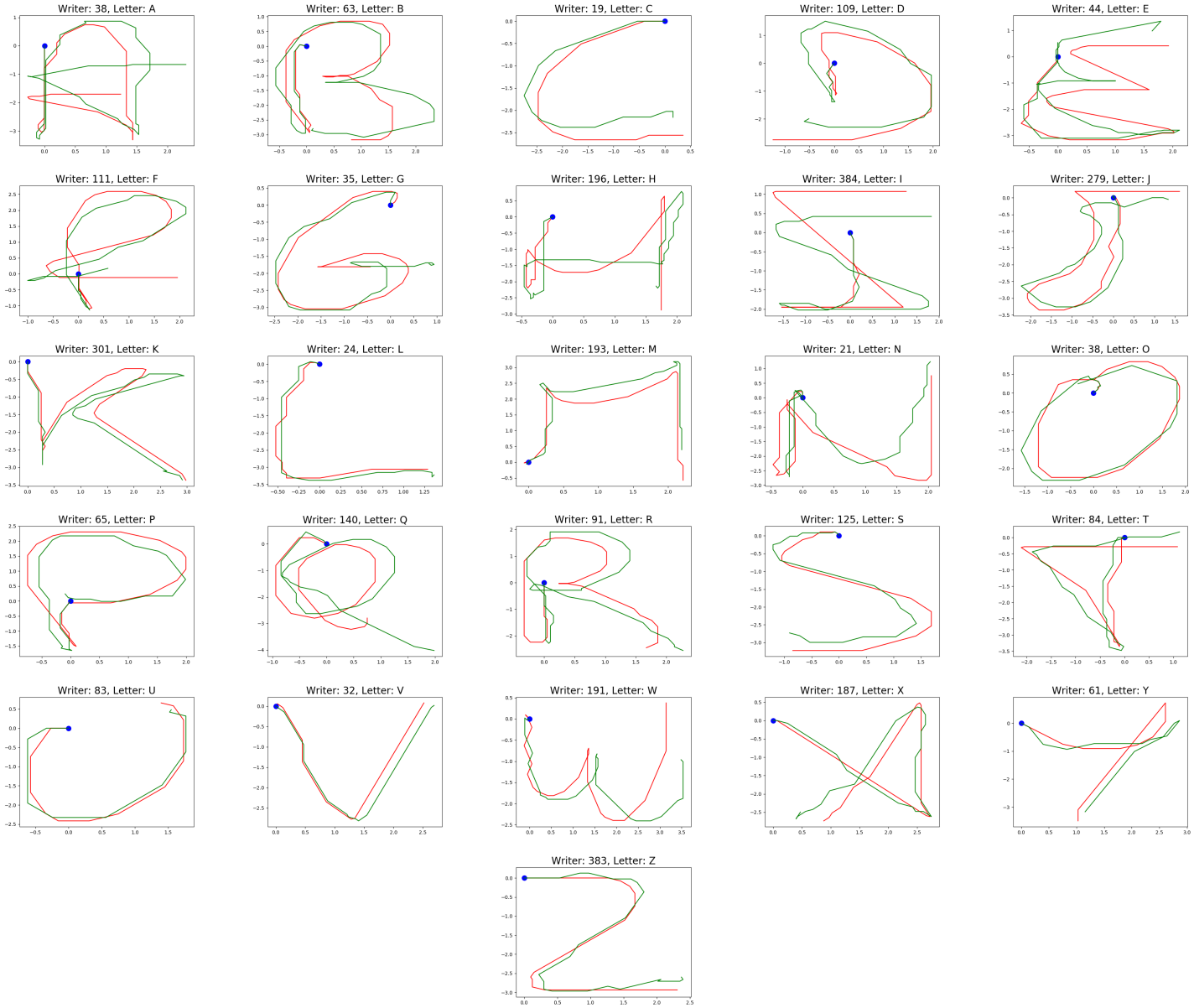


Figure 4.15: Examples of generated letters. The blue mark is the starting point. The traces in green is the ground truth, and the red is the generated ones by our model.

auto-encoder framework relative the benchmarks we proposed in the previous chapter, in order to determine if it is actually a valid framework to study styles. We find that an auto-encoder outperforms the benchmarks in all evaluation metrics, suggesting that this is a good approach.

- What kind of styles we can extract from this framework? and how do we extract them?: given the good performance results of the auto-encoder, we wanted to have further confirmation that our model is actual learning relevant style information from one side, and if we can learn something new about styles. We explored the model bottleneck for different letters, showing a strong evidence that the model is quite useful in the study of styles.
- Can we transfer the styles over different writers? We hypothesis that there is a limited number of styles in general – if we have enough writers in our training data, we will generalize well to new writers –. We test this hypothesis by hiding a number of writers from the training data, and compare a model trained on the other writers (transfer model) versus a model trained only on those writers (baseline model). We see clearly the transfer model outperforms the baseline, thus giving good evidence for our hypothesis.

We thus are ready for the next chapter, where we dive more into the world of transfer learning...

Chapter 5

Style Extraction and Transfer

Contents

Missing points

- Add a diagram explaining transfer Learning
- Add a diagram explaining which part of the model we are transferring, and which one we are retraining (maybe a series of diagrams explaining visually all the steps)
- Add examples of the generated letters/shapes

We finally arrive to the core objective of the PhD: how to leverage information of style from one (or more) task/s, in order to bootstrap the learning of a new task?

5.1 Background

Before we dive into the technical details, it is important to motivate why do we need to perform this in the first place. A common scenario is that we collect data over time for a particular task. This process can be quite expensive (e.g., collecting data from the robots is a slow and expensive process). Then, a new task of interest emerges. This task has common aspects with the old one. The question here is if we can leverage this common aspects, in order to bootstrap the learning the new task.

Examples of transfer learning:

- If you know how to hold a glass cup, with little adjustments, you can learn how to hold a plastic bottle.
- If you know probability and algebra, you can use this knowledge in order to bootstrap/accelerate your progress in mastering machine learning.

There are plenty of cases where transfer learning is/could be useful, like:

- In robotics (Konidaris et al., 2012a,b), collecting data can be quite expensive process, due to hardware limitations from one side, and human limitatio as well (in case of human-robot interaction scenarios). In addition, with techniques like reinforcement learning (Sutton and Barto, 2018), where the agent learns by trial and error, the process can be prohibitively slow, with safety concerns sometimes.
- In underwater acoustics (Malfante, 2018), an essential task is collecting and cleaning the data about the different fish sounds. This is a tediously manual job, and any change (type of fish, time of the day or place in the ocean) degrades the quality of prediction a lot. Transfer learning can be very useful in this case, to reduce the effort needed to collect, clean and annotate the data.

Points addressed in this chapter

- What is transfer learning? and what are the different approaches to perform it? I will explain the different paradigms and metrics used to characterise transfer learning.
- How do we approach the problem of style transfer, for both handwriting and sketch drawing?
- The experiments performed, the results, and our conclusions.

5.2 Transfer learning

An important research direction in machine learning nowadays is transfer learning. If humans and machines are able to learn how to perform a task, one of the thing that

separates humans from machines is the ability to leverage this knowledge in order to acquire new skills and perform new tasks, without the need for additional trials and errors from tabula rasa. This however, is not a straightforward thing for machine learning to do. The algorithms are fitted to data responding directly to the task required (i.e., has the same input feature space and same distribution). Thus, a change in the task can lead to degradation in the algorithm performance (Pan and Yang, 2009; Shimodaira, 2000; Tan et al., 2018; Weiss et al., 2016).

Let's first introduce some notations that will help in formulating the problem:

- We first introduce the concept of *Domain*. A domain defines a feature space (e.g., images of animals), and the probability distribution of this space (i.e., the distribution of pixels in the images of animals). We can consider the domain as the available *knowledge* to us. Thus, a domain $\mathcal{D}$ is defined as $\mathcal{D} = \{\mathcal{X}, P(X)\}$, where:
  $\mathcal{X}$ is the feature space, X is the data samples available to us from the feature samples, $X = \{x_1, x_2, \cdots, x_n\} \in \mathcal{X}$, where n is the size of the learning sample. $P(X)$ is the marginal distribution probability of this data sample.
- For a given domain (aka, the knowledge available to us), we define the concept of a *task*. A task is something we would like to achieve using the knowledge we have. For example, a task can be classifying animals given animal pictures, or perform robot navigation given a map of the building. Thus, a task $\mathcal{T}$ is defined as $\mathcal{T} = \{\mathcal{Y}, f(.)\}$, where:
  $\mathcal{Y}$ is the label space (the task objectives), $f(.)$ is the mapping function (mapping the domain knowledge to the task objectives). It can also be rewritten as a conditional probability over the domain knowledge, $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$.
- Based on this notation, we define two more concepts: *Source* and *Target*. A source defines a domain and a task/s that are available to us already (where we have plenty of domain knowledge, and examples on the task/s). A target defines a domain and a task/s as well, where we usually do not have enough domain knowledge and/or examples on the task/s.

Now that we clarified some basic terminology, we can move on to define transfer learning: given source domain data D_S, source task $\mathcal{T}_S$, target domain D_T and target task $\mathcal{T}_T$, we wish to improve the performance of the target task $f_T(.)$ by using D_S and $\mathcal{T}_S$.

Given this definition, we can categorize different types of problems that transfer learning covers:

- The source and taget domains are different, $D_S \neq D_T$, which means that the feature space is different, $\mathcal{X}_S \neq \mathcal{X}_T$, and/or the probability distribution of the feature space are not the same, $P(X_S) \neq P(X_T)$. If $\mathcal{X}_S \neq \mathcal{X}_T$, the transfer learning problem is *Heterogeneous*. Otherwise, it is *Homogeneous*.

- The source and target tasks are different, $\mathcal{T}_S \neq \mathcal{T}_T$, which means that the objectives are different, $\mathcal{Y}_S \neq \mathcal{Y}_T$, and/or the mapping function (from the feature space to the objectives) are different, $P(Y_S|X_S) \neq P(Y_T|X_T)$.

Many approaches in order to achieve transfer learning has been proposed in the literature. We will discuss the different approaches, and give examples from the literature on each one. We follow the categorization of transfer learning approaches done in (Tan et al., 2018)[1], by first identifying four main categories of transfer learning:

- *Instances-based*: in this case, we utilize examples from the source domain into the training of the new target domain, by defining weights on them.
- *Mapping-based*: the objective in this case is to project the instances from the two domains into a new manifold, that increases the similarity between the two domains.
- *Network-based*: the more common type of deep transfer learning. It is based on the idea that the layers of the deep neural network extracts basic and general information, that shared a lot with other domains. In this case, the network or some of its layers are re-used on the target task.
- *Adversarial-based*: similar objective to *Mapping-based*, by using generative adversarial networks (Goodfellow et al., 2014) in order to find a manifold that are fit for both source and the target domains.

In the context of *deep transfer leaning* – the main domain in our work –, *Network-based* and *Adversarial-based* transfer are the relevant categories in this case. *Adversarial-based* transfer is quite recent, and there is not much to talk about at the moment, so we will focus *Network-based* transfer, as it is the most common type of deep transfer learning.

When it comes to evaluating the success of the transfer, there is no one way to evaluate transfer learning in general. This depends a lot on the objectives of transfer learning, and the criteria of success. In the case of machine learning, the improve in end quality of the model is the primary performance aspect being measured and reported – like the classification accuracy (Chattopadhyay et al., 2012; Glorot et al., 2011; Long et al., 2013; Pan et al., 2010a), the reduction in the average error (Pan et al., 2010b), ...etc –. The transfer is considered successful if it achieves better performance than the baseline method.

We expect that, with the introduction of transfer learning via deep learning, that the *time to train* the model could be another aspect to consider – similar to what is being used in reinforcement learning (Taylor and Stone, 2007) –, although – to the best of our knowledge – we do not find studies mentioning this at the moment.

[1] Other categorization exists, like the one used in (Weiss et al., 2016). When it comes to deep transfer learning, we believe the one in (Tan et al., 2018) to be the most relevant.

5.2.1 Network-based transfer learning

The idea of using deep learning (LeCun et al., 2015) in order to achieve transfer learning has gain popularity during the last years, following the achievements in having better computational resources (Raina et al., 2009), and the availability of large benchmark datasets - most notably: ImageNet (Deng et al., 2009) for object detection, MS-COCO (Lin et al., 2014) for image captioning ....

The first notable success of deep learning happened in the area of computer vision, with the AlexNet architecture (Krizhevsky et al., 2012). It was found out that such a deep network manages to extract generic features about the images: it learns simple, hierarchical filters, that are generic enough to be applicable for different datasets (see figure 5.1). The filters can be also seen as a representation of knowledge learned on the given tasks, with the first layers learning primitive filters (like edge detection for example), and the subsequent layers learning more complex filters. This observation led to another surge in the usage of pretrained AlexNet – and later newer architectures, like VGG16 (Simonyan and Zisserman, 2014), Inception (Szegedy et al., 2015),...etc – as feature extractors for new, unseen datasets (i.e, transfer learning using these deep learning architectures).

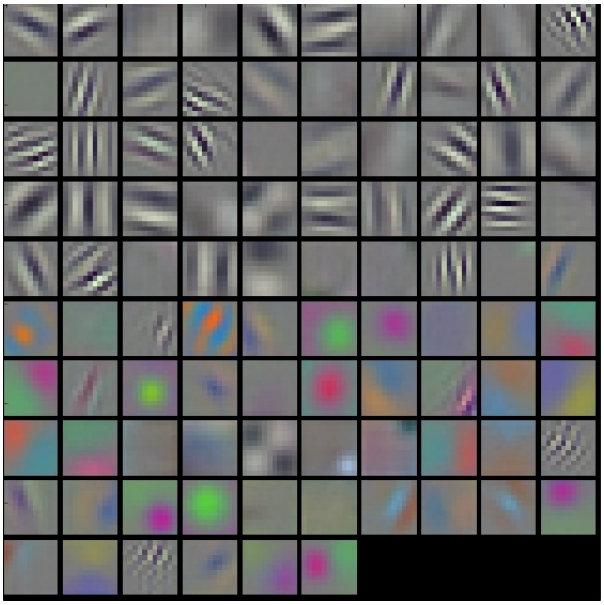


Figure 5.1: Visualization of the first convolution layer of a trained AlexNet. Note the basic shape of filters that resemble to Gabor filters widely used in image processing for decades (Fogel and Sagi, 1989; Jain and Farrokhnia, 1991). It is possible to see that those same filters will be useful in other computer vision or image-related tasks.

Examples on transfer learning in case of network-based transfer are:

- *Sentiment Classification*: (Glorot et al., 2011) discusses a deep learning approach for transfer learning for sentiment classification, by using stacked de-noising auto-encoders (Vincent et al., 2008) to correct the marginal distribution between the source and the target domain, by learning latent variables/features common between the two data sources in two steps: first, train an auto-encoder on the unlabeled data from the source and the target. This will produce latent variables that will make $P(X_S)$ closer to $P(X_T)$. Then, use those latent features to train a classifier on the labeled source data.
  Experiments are done on 12 different sources and target domain pairs. The data used reviews for different products (4 different products). The performance metric used in this case is the classification error rate when using a classifier trained on the source task only, minus the classification error rate of a classifier trained on the target task only. A SVM classifier trained on the source domain is used as a baseline, and a comparison with other transfer methods (Blitzer et al., 2006; Li and Zong, 2008; Pan et al., 2010a). All these methods performs better than the baseline, and (Glorot et al., 2011) peforms better than all of them.
- *Underwater Acoustics*: (Malfante et al., 2018b) compares the use of deep transfer learning to manual features (Malfante et al., 2016, 2018a), in order to recognize the sounds of fish underwater. Interestingly, the deep learning models are trained on ImageNet (Deng et al., 2009) dataset (which is completely unrelated to acoustics), then used in order to extract features from the spectrogram of the underwater recordings. The assumption here is the the filters learned with deep learning are basic and generic enough, to be used in other domains, thus, deep learning can provide a natural correction for $P(X_S)$, to make it close to $P(X_T)$. Without any fine-tuning, the deep learning achieves quite a high performance, although less than the state of the art, suggesting that a further investment in this direction is worth the time.
- *Speech to speech translation*: (Xu et al., 2014) studies the idea of transferring the speech knowledge learned from one language to another using deep neural networks. In particular, the study the transfer between Mandarin and English (in both ways). Their hypothesis is that when modeling the speech in a language, there is a part in the model that is language independent. Their work was to find this part, and use it to bootstrap the learning of a new language (in case of neural networks, this can be rephrased as trying to find the good part in the neural network that is language independent).
  They compared their proposed transfer approach to a baseline model (a deep learning model trained on the target task only), and compare the different the usage of different layers of a deep neural network trained on the source task (in order to determine the best part, or the language independent part). They report multiple performance metrics: segmental SNR (SSNR), log-spectral distortion (LSD), and perceptual evaluation of speech quality (PESQ). They find that, given insufficient data on the target task (which is one of the motivations to perform transfer learning), this transfer scheme works better than the baseline.

- *Image classification*: (Oquab et al., 2014) investigated the usage of a convolution neural network (CNN) trained on ImageNet (Deng et al., 2009) (where the object of interest is centered in the image), to extract low-level and mid-level features, that can transfer well to more complex dataset, PASCAL VOC (Everingham et al., 2010) (where there are several objects of interest in the image), and is smaller than ImageNet. They use the *average precision* as their performance metric, and they show that the transfer is better than training a model from scratch on this target dataset.
- *Styles of speech synthesis*: (Wang et al., 2018) discusses the problem of transferring the styles between different speakers, in the task of speech synthesis. Traditionally, the oucome of speech synthesis systems is the same style. One of the challenges is to capture the richness of style of different speakers on some training data, and transfer this style to new text. In their work, they propose an embedding approach called *Global Style Tokens*, in order to extract the styles from the different speakers during the training phase. They then show that they can use these tokens as extra information to the speech synthesis system, to bias/affect the style of the outcome. They compare this to a baseline model, called *TACOTRON* (Wang et al., 2017b), without these GST addition. They use *Mean Opinion Score* (MOS) performance metric – a subjective metric to assess the quality of experience experienced –. They conclude that transfer learning using GST outperforms the baseline model.

5.3 Putting it all together

In the previous section, we explored the concept of transfer learning, with focus on network-based transfer, with multiple examples from the literature on this type of transfer. In this section, I will explain the work done during the PhD on transfer learning. The main challenge we had is how to capture and transfer the styles between different tasks. This proposes multiple questions:

- Which framework/methodology of transfer we should use?
- How to assess the quality of transfer?

In this part, we study these questions on both *IRONOFF* and *QuickDraw!*. Our experimental setup is inspired by the work done in (Lathuiliére et al., 2019). We use a slightly different experimental protocol between the two datasets however. I will discuss the reasoning behind this, and will argue that both setups are good enough to address our questions concerning the transfer of styles.

What are we transfering? A typical approach to perform transfer learning in deep learning models is to:

1. Train the model on the source task,
2. freeze some parts of that model,
3. when going to the target task, use the frozen parts in a new model, and train the rest of that new model on the target task

The part that we freeze/transfer in our case is illustrated in figure 5.2.

How to evaluate the quality of transfer? Concerning the performance metrics, we decided to use multiple metrics to determine the effect of transfer learning, and compare it to the baseline (no transfer). While the usage of a single metric offers more convenience for decision making (i.e., which model to choose), our goal is have a better understanding for the transfer of style. To achieve this, we would like to shade light on the outcome from different angles, by using multiple metrics. We first evaluate the quality of the model in prediction by looking at the log-likelihood on the test data[2]. We then evaluate the quality of generation using our previously proposed metrics (the BLEU score and EOS analysis). We compare perform the generated strokes count, and include it in the comparison.

One of the things we changed from the previous work is that, when analysing *end of sequence*, instead of using Pearson Coefficient, we use Krippendorff-Alpha (Krippendorff, 2011). The reason for this is that it offers a smoother transition between what is good and bad. In Pearson coefficient, if there is a mismatch between the generation and ground truth, it will not matter how big is the mistake (the distance between the generated and the ground truth). With Krippendorff-alpha, we can take into account the distance between the generated and the target symbols, thus providing a softer and more realistic estimation for the correlation.

Another aspect we consider in our analysis is the number of *strokes* the model is generating. In the previous work, we did not model the strokes, in order to simplify the system and the analysis. Now that we are confident about our approach, we added the strokes to our model. Strokes are more complex to model, because the stroke signals are more sparse. A good stroke generation is an indication that the model can perform hard discrete decisions in order to generate the whole shape. We use Krippendorff coefficients as well to report and analyze the strokes. We also consider the confusion matrices for the generated strokes versus the ground truth ones, in order to shade more light on the behavior of the model.

One important choice to make is to decide the amount of data to be used in the target task. Normally, the whole point from transfer learning is to address the insufficient data available for the target task. However, since we do not have a particular scenario to solve (we are mainly interested in that 'what if this happens'

[2]Based on experience, the cross-entropy matters when it comes to generation. A difference of around 0.1 in cross-entropy between two models gives different generation quality.

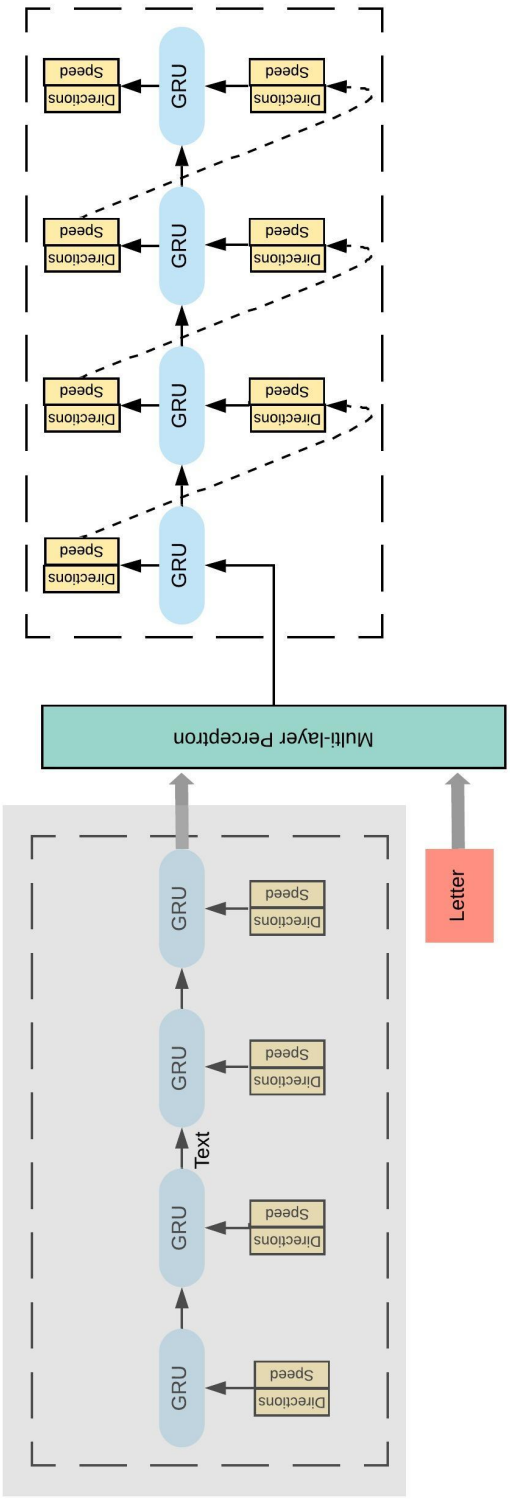


Figure 5.2: An illustration for model we use. The part identified by the grey area – the style extraction module – is what we transfer. Since we give the model the task content/identity, the grey part is expected to focus more on the style extraction. During the exposure to the source task, all the different components of the model are being trained. During the exposure to the target task, the grey area is the trained one on the source task, with frozen parameters. The other parts will be trained.

question), there is no clear criteria to decide the amount of data. If we select very few points, we may be biased towards giving transfer learning an advantage, while also sacrificing having generally bad results (even if transfer learning proved to be better than baseline on our performance metrics). It is important that the overall behavior of the system is acceptable (generating shapes with acceptable quality). To get around this design problem, we decided to use the entire data available to the target task, thus, testing the 'worst case scenario' for transfer learning (where the baseline has plenty of data, and it is used to train the whole model, compared to training only half the model in case of transfer learning). This will be the case for both *IRONOFF* and *QuickDraw!* datasets.

5.3.1 IRONOFF

In case of handwriting letters, we distinguish between three tasks: uppercase, lowercase and digits[3].

We explore the idea of transfer learning on all possible combinations of tasks:

- From uppercase and lowercase to digits,
- from uppercase and digits to lowercase,
- and from lowercase and digits to uppercase.

Experimental setup

Figure 5.3 details the protocol we used for this experiment. For each source/target task combination, we first perform hyperparameter search for the network for both the source and the target tasks. The best performing hyperparameters gives use the source and the target models. Using these models:

- We use the source model to extract the encoder module (aka, the style extraction module). We then add it to a new model, freeze it (it will not be part of training), and train the new model on the target task. We also search for the best hyperparameters for this new model. We call this model the 'transfer model'. We retrain this transfer model 5 times with different random weights each time, and report the stats on the cross-entropy of the test data for these repetitions.
- We retrain the target model 5 time as well.
- Then, we use the best performing transfer and targets models for generating the target tasks, and report the stats of the different generation metrics.

[3] I do not see a problem of working on the lower categories (A, a, B, b, ..., 7, 8, 9). However, it would have consumed a lot of computational time test on all possible combinations of source/target tasks.

We use 10% of the data for testing, another 10% for validation, and the rest is our training set.

Results

Loglikelihood of prediction In the prediction mode, the model is tested in a similar manner to the way it was trained: it is given the input from the ground truth, representing the current time step, and asked to predict the next time step. This tells us about the quality of the training procedure from one side, and shade light on the confidence of the model in predicting the next time step.

The result of 5 times repetitions, for all the different combinations of source/target tasks, are mentioned in figure 5.4. We can see that the transfer learning always gives a significant advantage over baseline models.

BLEU score Now we go to the generation part, our main concern in this thesis. As discussed earlier, we use BLEU score to assess the quality of matching segments between the generated and the ground truth letters, in a gradual manner (i.e., we increase the size of the segments to match gradually). Table 1 summarizes the numbers. We see that transfer learning always performs better than the baseline.

End-of-Sequence analysis One aspect to measure the quality of the generation, that we identified previously, is to analyse how the model predicts the end of the sequence generation. Table 2 shows the results of the Krippendorff coefficients for the different modes. In general, the different modes are performing quite well. It can be seen that transfer learning is actually performing better than the baseline models, adding another indication to the benefit of using transfer learning.

Strokes analysis As mentioned earlier in this section, we started to consider the the strokes in this part of our work. Table 3 shows the strokes results of the Krippendorff coefficients for the different modes. With the exception for the uppercase letters, transfer learning performs better than baseline models. For the uppercase, this could be due to the extra complexity of these letters (see figure 2.3, where it can be found that uppercase letters are usually the ones with higher number of strokes). A fine tuning for the style extraction module is the next logical step to perform here, to make it adapt to complexity of the this task. We also consider the confusion matrix for the generated strokes in comparison with the ground truth, figure 5.5. We can see that it is consistant with the Krippendorff results, where transfer learning outperforms the baselines. In the

Figure 5.3: Flow chart explaining the experiment protocol used in *IRONOFF* dataset.

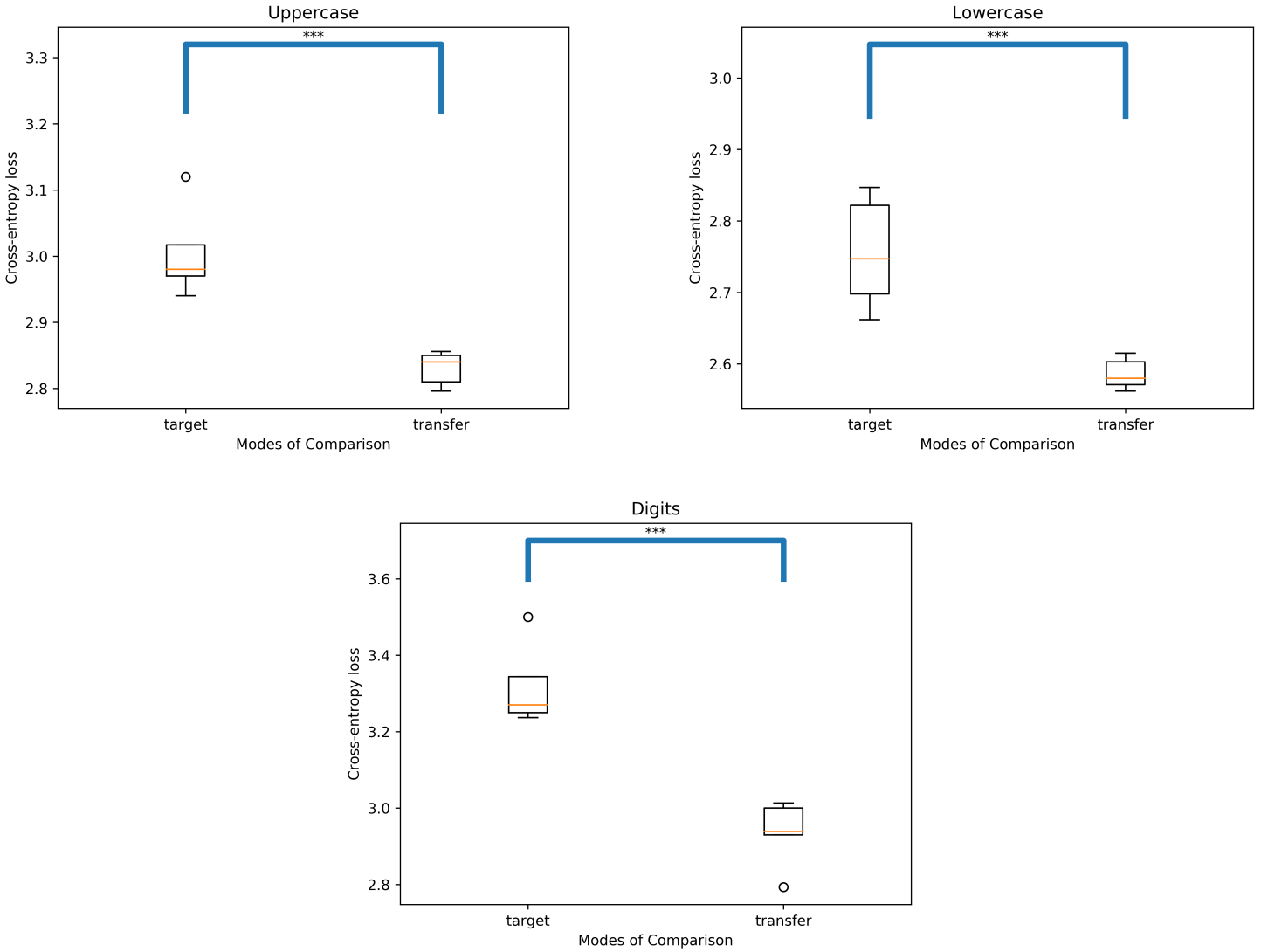


Figure 5.4: *IRONOFF*: log cross-entropy of prediction of test dataset for different combinations of source/target tasks, with 5 repetitions. We can see that, in all possible source/target task combinations, the transfer learning gives better advantage than just learning from scratch on the target task. The stars indicate the statistical significance level (1 for < 0.05, 2 for < 0.01 and 3 for < 0.001).

uppercase letters however, it is notable that the transfer learning performs better on the single stroke, while the baseline performs better on the 2 and 3 strokes.

Aspect/Feature	Speed			Freeman		
Model / B-score	B-1	B-2	B-3	B-1	B-2	B-3
Uppercase-baseline	66.1	46.3	27.2	62.8	49.4	37.1
Uppercase-transfer	**68.3**	**47.8**	**28.3**	**65.47**	**51.8**	**39.0**
Lowercase-baseline	73.1	69.7	55.9	54.8	37.2	40.9
Lowercase-transfer	**75.5**	**71.2**	**58.0**	**56.0**	**39.4**	**41.9**
Digits-baseline	68.7	65.2	49.1	49.6	29.3	34.6
Digits-transfer	**71.5**	**70.7**	**51.2**	**55.9**	**31.4**	**41.7**

Table 1: *IRONOFF*: BLEU score results on the generated letters, for the baseline models (trained on the target task only), and the transfer models (the encoder – style extractor – is trained on the source task, while the decoder is trained on the target task). The results show the advantage of using transfer learning.

Task/Model	Baseline	Transfer
Uppercase	0.95	**0.97**
Lowercase	0.96	**1.0**
Digits	0.92	**0.99**

Table 2: *IRONOFF*: Krippendorff correlation coefficients for the End-Of-Sequence (EoS) distributions between the transfer and baseline, for all tasks.

Task/Model	Baseline	Transfer
Uppercase	**0.38**	0.25
Lowercase	0.56	**0.65**
Digits	0.4	**0.71**

Table 3: *IRONOFF*: Krippendorff correlation coefficients for the strokes distributions between the transfer and baseline, for all tasks. Except for the uppercase case, transfer learning seems to perform well in the lowercase and the digits tasks.

5.3.2 QuickDraw!

In case of sktech drawings, we define the task by the class it belongs to. So we have 5 tasks: circle, triangle, square, hexagon and octagon. We explore the idea of transfer learning on all possible combinations of tasks: in each combination, one task is removed (the target task), and the other task are considered source tasks.

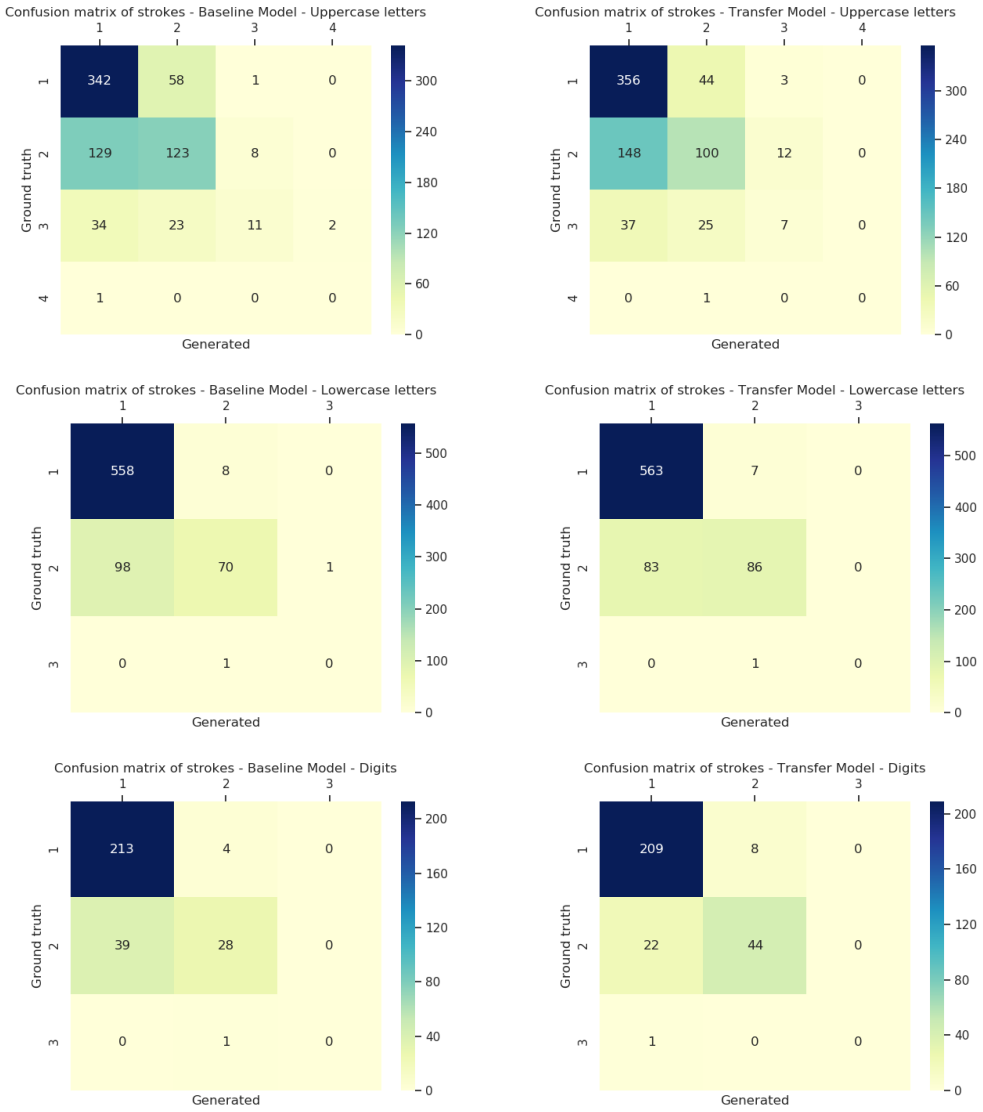


Figure 5.5: *IRONOFF!*: Confusion matrix for strokes for both baseline (left) and transfer (right) models, on the different tasks.

Experimental setup

Due to the increase the number of tasks – compared to *IRONOFF* –, and using the same experiment setup proved to be too much compared to the computational resources available to us. We identify that one of the expensive bottlenecks in our work is the number of times we need to perform hyperparameter search. In order to get around this problem, we are using a simpler (and less powerful protocol), see figure 5.6. We perform the hyperparameter search once in the beginning of the experiment, on the all the tasks combined, to find a good hyperparameter that is suitable to the domain of 'sketch drawing'. From there, we fix the hyperparameters for all the different steps in the experiment. The analysis steps are the same as in *IRONOFF*.

Another thing we noticed is that retraining the model 5 times only (like in *IRONOFF*) only leads to unstable conclusions, thus make it hard to quantify the difference between the baseline and the transfer modes. We increased the number of models to 30 in this case, in order to have a more consistent trend.

Results

Loglikelihood of prediction The results can be seen in figure 5.7. For all the combinations, there is a clear advantage for using transfer learning over the baseline model. The difference in all cases is statistically significant.

BLEU score The BLEU score results are summarized in table 4. Transfer learning outperforms the baseline models in this case.

End of Sequence analysis Table 5 summarize the results on end-of-sequence analysis in case of quickdraw. The benefits of transfer learning are clear.

Strokes analysis Table 6 summarize the results on strokes analysis in case of *Quick-Draw!*. The benefits of transfer learning are clear. We also report the confusion matrix for the strokes, figure 5.8. It is notable that the performance of the system in general decreases, as the diversity in the number of strokes increases. But still, transfer learning provides the better results.

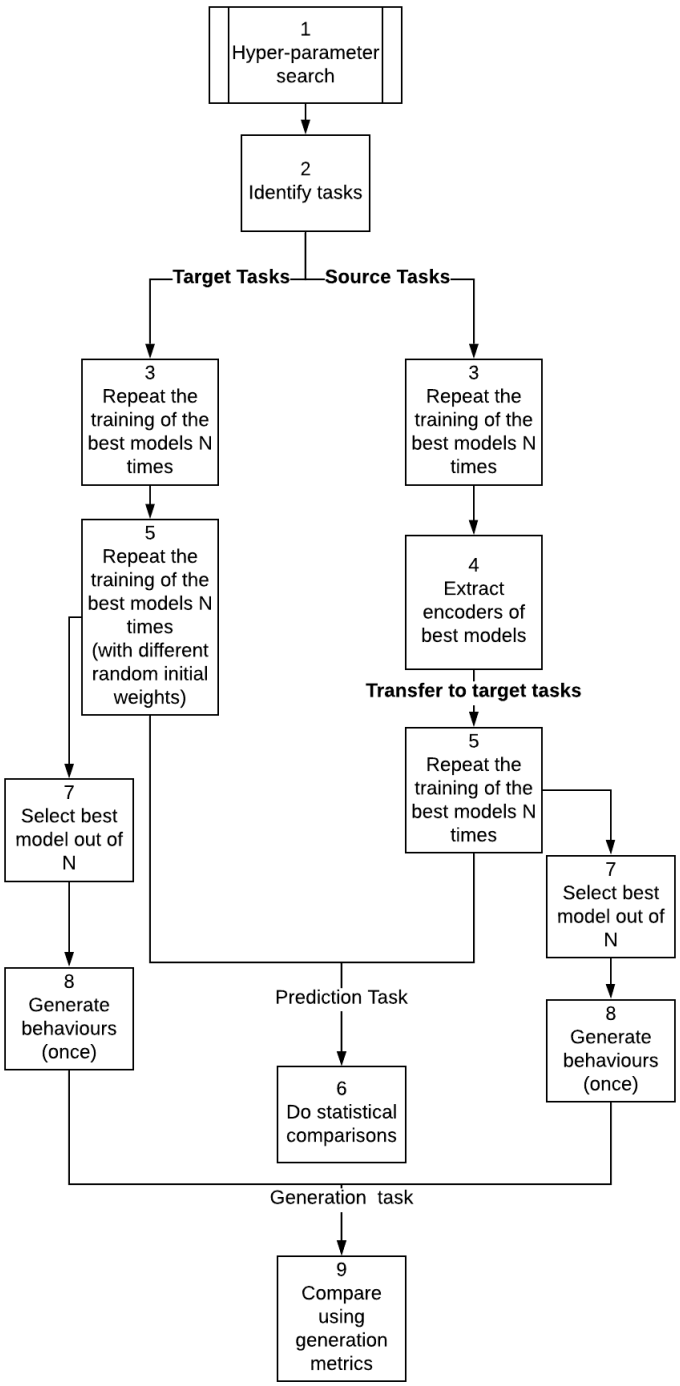


Figure 5.6: Flow chart explaining the experiment protocol used in *QuickDraw!* dataset.

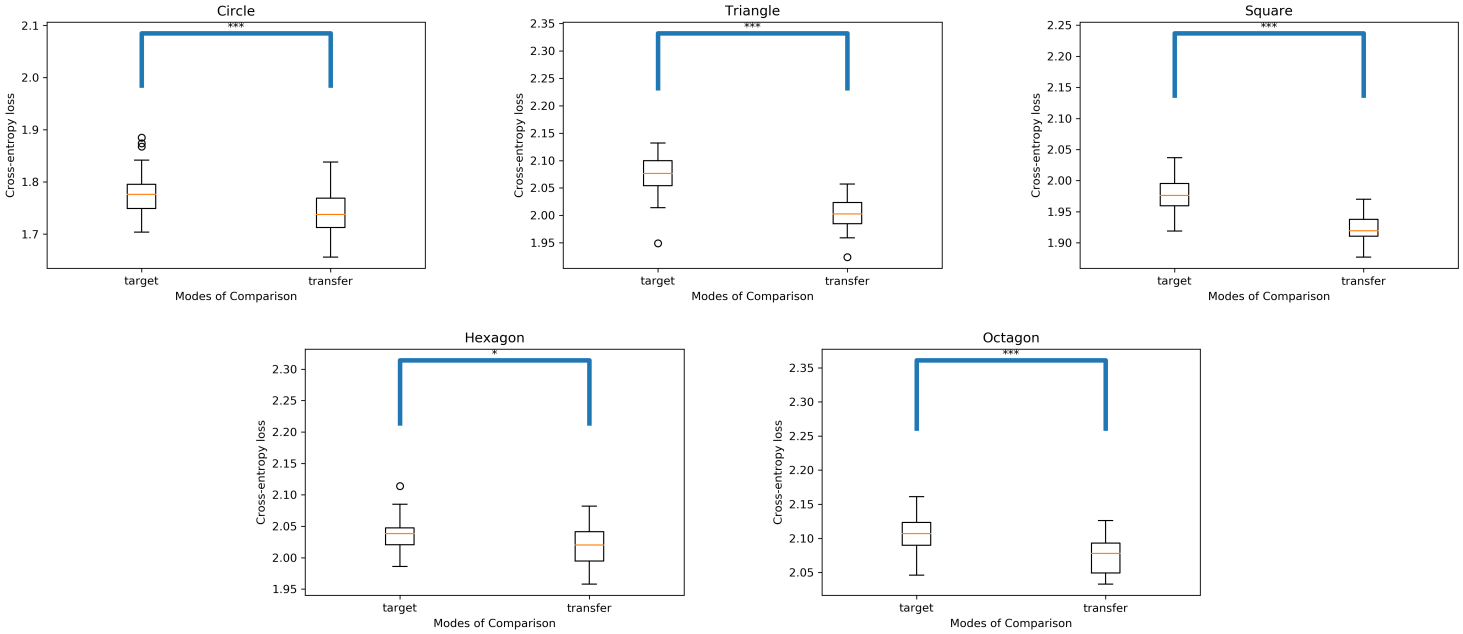


Figure 5.7: *QuickDraw!*: cross-entropy of prediction of test dataset for different combinations of source/target tasks, with 30 repetitions. We can see that, in all possible source/target task combinations, the transfer learning gives better advantage than just learning from scratch on the target task. The stars indicate the statistical significance level (1 for < 0.05, 2 for < 0.01 and 3 for < 0.001).

Aspect/Feature	Speed			Freeman		
Model / B-score	B-1	B-2	B-3	B-1	B-2	B-3
Circle-baseline	59.0	54.5	49.6	60.0	54.7	48.9
Circle-transfer	**70.4**	**65.5**	**60.3**	**65.0**	**58.1**	**50.6**
Triangle-baseline	47.3	40.0	32.6	33.2	28.2	24.0
Triangle-transfer	**61.3**	**52.4**	**44.1**	**50.6**	**44.8**	**39.8**
Square-baseline	46.8	40.1	32.7	44.0	39.1	34.9
Square-transfer	**57.9**	**50.8**	**42.9**	**53.0**	**47.4**	**42.3**
Hexagon-baseline	58.1	50.4	41.4	45.4	40.3	35.9
Hexagon-transfer	**62.0**	**54.0**	**44.8**	**47.6**	**42.3**	**37.8**
Octagon-baseline	55.2	47.1	38.3	43.7	38.7	34.6
Octagon-transfer	**57.3**	**49.3**	**40.5**	**46.1**	**41.1**	**36.7**

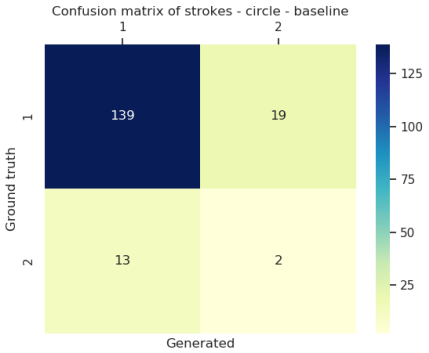
Table 4: *QuickDraw!*: BLEU score results on the generated letters, for the baseline models (trained on the target task only), and the transfer models (the encoder – style extractor – is trained on the source task, while the decoder is trained on the target task). The results show an advantage in using transfer learning.

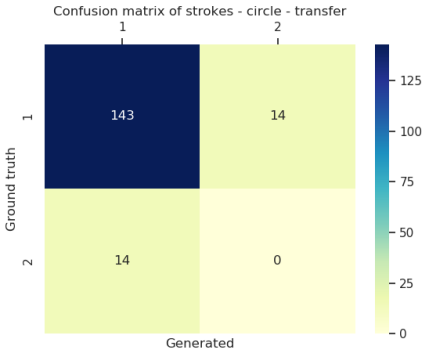
Task/Model	Baseline	Transfer
Circle	0.6	**0.84**
Triangle	-0.05	**0.61**
Square	0.04	**0.35**
Hexagon	0.07	**0.2**
Octagon	0.07	**0.16**

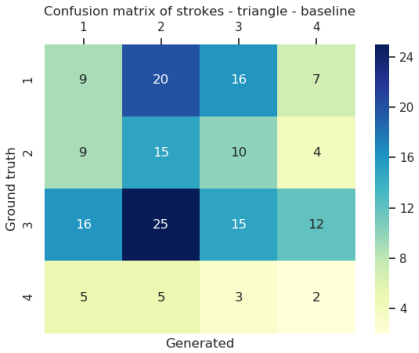
Table 5: *QuickDraw!*: Krippendorff correlation coefficients for the end-of-sequence distributions between the generated letters and the ground truth letters.

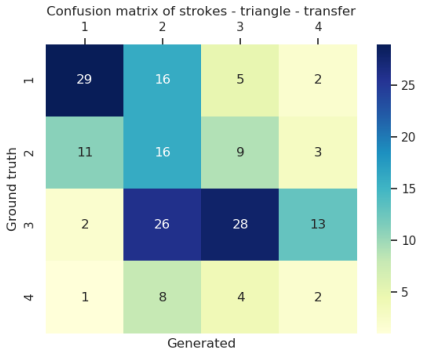
Task/Model	Baseline	Transfer
Circle	-0.04	**0.1**
Triangle	-0.04	**0.42**
Square	0.03	**0.25**
Hexagon	-0.08	**0.23**
Octagon	0.06	**0.18**

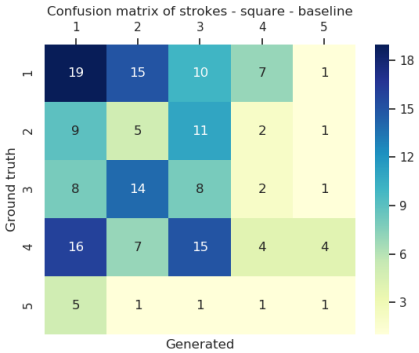
Table 6: *QuickDraw!*: Krippendorff correlation coefficients for the strokes distributions between the generated letters and the ground truth letters. Transfer learning achieves better results than the baseline on all the different tasks.

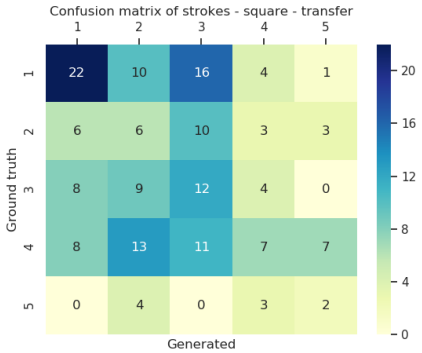

Confusion matrix of strokes - circle - baseline
Ground truth
Generated
139
19
13
2


Confusion matrix of strokes - circle - transfer
Ground truth
Generated
143
14
14
0


Confusion matrix of strokes - triangle - baseline
Ground truth
Generated
9 20 16 7
9 15 10 4
16 25 15 12
5 5 3 2


Confusion matrix of strokes - triangle - transfer
Ground truth
Generated
29 16 5 2
11 16 9 3
2 26 28 13
1 8 4 2


Confusion matrix of strokes - square - baseline
Ground truth
Generated
19 15 10 7 1
9 5 11 2 1
8 14 8 2 1
16 7 15 4 4
5 1 1 1 1


Confusion matrix of strokes - square - transfer
Ground truth
Generated
22 10 16 4 1
6 6 10 3 3
8 9 12 4 0
8 13 11 7 7
0 4 0 3 2

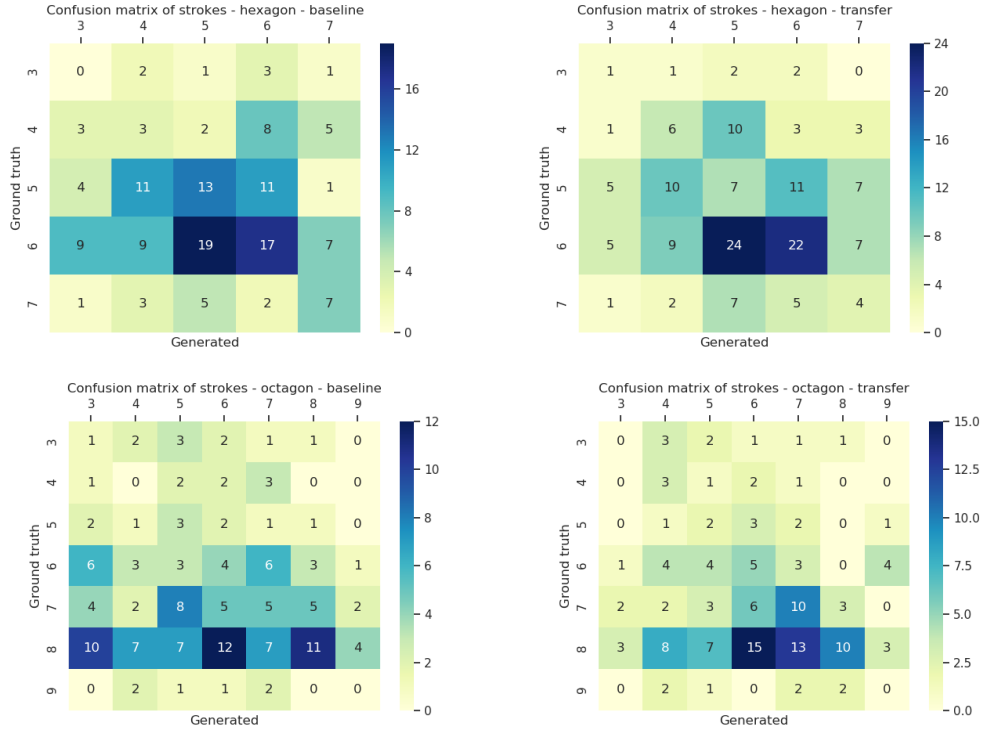


Figure 5.8: *QuickDraw!* Confusion matrix for strokes for both baseline and transfer modes, on the different tasks.

5.3.3 A word of caution about confusion matrix

In the previous two experiment, we presented the confusion matrix for the strokes. However, it is important to understand what these numbers actually mean. In a normal classification problem, the confusion matrix gives important information about the precision and recall of the classification. When comparing two algorithms – for example –, the usage of the values of confusion matrix is an acceptable approach to consider. However, this is not the case when dealing with the generative aspects of machine learning. Bear in mind the asymmetry between the training of the model (using the log-likelihood of prediction) and the usage of the model (by sampling from the model), thus, a one-to-one comparison between what is generated and the ground truth is not possible or meaningful. A logical consequence for this is that we can not use such a method to compare two models together (i.e., compare transfer learning and the baseline). What we would like to achieve with generative model is to *capture* the distribution of the ground truth, thus, for a meaningful comparison, we want to compare the distribution of the generation relative to the distribution of the ground truth. That is a core challenge in the this PhD: to provide tools to capture the distribution, to facilitate the comparison and evaluation of different models/methods.

One way to look at the confusion matrix is the general trends in the matrix, like:

- The deviation around the diagonal: as the shapes gets more complicated, and as the number strokes increases, we expect that the deviation will increase.
- Looking for systematic errors (e.g., the number of strokes for a square is always one instead of being three or four): this could help diagnosing problems with the models design, data processing ... etc.

5.4 Are we actually capturing styles?

In the previous section, we showed that all the metrics indicate the benefits of transfer learning over using the target data only. We argued that this approach captures the styles, and that we are transferring the styles to a new task. But, are really capturing the styles this way?. In section 4.2.4, we tried to motivate this point, by some exploring the bottleneck of the styles. We showed that looks consistent with what we know beforehand (in case of letter 'X' for example), or that it uncovers things we either did not notice beforehand (like in letters 'C' and 'S'), or uncover new things that we did not anticipate (in case of letter 'A').

Testing for something that is ill-defined, and not always known beforehand, like styles, is quite challenging. To add to the challenge, the methods we used to

explore the bottleneck space (PCA and tSNE) are not designed to uncover a 'styles' manifold, so sometimes, basic styles went missing (for example, with this exploration methods, we could not uncover the clockwise/counterclockwise direction in letter 'O' for example, which we know a priori that it exists!).

In this section, we want to shade more light on this problem, from a different and quantifiable angle. We will choose one aspect using *QuickDraw!* dataset this time, we manually annotated circles and octagons into clockwise/counterclockwise categories. Then, we build a classifier on top of the styles bottleneck, to measure the quality of the capturing of this style aspect. If we can not see this style visually with PCA and tSNE, then we should be able to measure its impact.

We annotated 716 drawings (see table 7). We trained a classifier (*Random forests* classifier) on this data directly, randomly separating training and test data, and repeated this 5 times. We did not configure the hyper-parameters of that classifier (we want a general figure, not interested in an exact one). We obtained 92.9 6.0 % on the weighted F1 score. This suggests that this style extraction method does indeed capture this style aspect.

Task/Direction	Clockwise (CW)	Counter-clockwise (CCW)
Circle	339	57
Octagon	115	205

Table 7: Results of manual annotation for CW-CCW on 716 drawings (octagon/circles) in *QuickDraw!* dataset. Sometimes the drawing is not clear, so we did not include. The selected examples are the clear ones only. It can be seen that the data is not balanced.[4]

5.5 Summary and take-away message

In this chapter, we presented our hypotheses for style transfer learning. We presented our proposed approach, and the proposed experimental protocol in order to investigate these hypotheses. To have better support for our conclusions, we carried out the experiments on two different datasets, *IRONOFF* and *QuickDraw!*, both of them presenting different challenges and behaviors: *IRONOFF* has more clear semantics, made by a pen, while in *QuickDraw!* the contributors excerted more freedom on their work, and most of the work is believed to be done by the mouse or maybe a tablet (using a finger or a pen). Also, we chose to perform the study on the worst case scenario, where there is an abundance of data available to the target task. To better understand and compare transfer learning versus the baseline models (models trained only on the target task), we considered multiple performance metrics: the log-likelihood of prediction, the BLEU score, end-of-sequence and strokes analyses for the generated shapes.

The results overwhelmingly point to the benefit of transfer learning compare to the baselines on the different proposed metrics, thus providing a strong evidence to our hypotheses.

Part III

Discussion and Closing Remarks

Chapter 6

Prespective and future work

Contents

This chapter will be a free discussion about what I had done, lessons learned, shortcomings of this work, difficulties in the PhD, and potential areas of development.

Science should always be about honesty, humility and respect, and not just flashy results and wide conclusions. Science can always make use of learning from setbacks – something that is almost missing in the scientific literature –. I will do my best in this chapter to highlight the other side of good results.

Maybe one day someone – maybe a PhD student – will decide to follow on this work. It is important for me that they do not repeat the same mistakes. Instead, everything should be ready for them in order to make new mistakes. That is how we move forward. After all, we our objective in this PhD is transfer learning between different tasks. It is now time to transfer learning between two different colleagues.

6.1 Challenges

In this section, I will discuss the challenges faced during the PhD – from my personal perspective –, in the hope that I will learn from it for the future, and to illuminate the path for others as well.

6.1.1 Choice of how to tackle the topic?

The usage of deep learning in this topic has not been the first or clear option from the beginning. As I mentioned earlier, the objective of the project was to develop components for human-robot interaction domain. In a first glance, deep learning and human-robot interaction do not really mix well. The problem is simply the availability of data.

There is a stretch of imagination in this thesis, that we assume – and I believe rightly so – that the data problem in HRI will be resolved in the future. Better and more reliable hardware is becoming available, and there is a general awareness now in the community about the need to do something concerning the data: a lot of data is being recorded by the research group, but there is no standardization or culture of open sourcing the data, even within the same team, leading to a big waste of efforts and time.

At first, it seems that more data-efficient methods – that depends on well designed priors from humans – are the way to go in such project. However, the advances in deep learning application in areas like speech synthesis, image captioning, text and music generation, and the lucrative possibilities that data-driven approaches provide were hard to ignore. Besides, the current advances in machine learning indicate that computational approaches, even with simple algorithms, are outperforming methods that depends on human knowledge and prior[1].

6.1.2 Determine the scope of interest in the state-of-the art

A major challenge during the PhD was to determine the relevant state of the art. For generative models, the usage of deep learning methods was not the clear choice of the beginning, and once chosen, it took considerable effort to determine the scope of the relevant literature.

The same goes for the state of the art on styles. The word itself, and the range

[1] As nicely noted by Richard Sutton, one of the god fathers of reinforcement learning, in his article *The bitter lesson*, `http://incompleteideas.net/IncIdeas/BitterLesson.html`

of study, is very wide – as noted early in the introduction. By far, the work done in handwriting styles was the least relevant to our work; either because the work is done on offline handwriting (thus not dealing with the dynamics of writing itself) – and this is most of the work done currently –[2]. The problem does not always manifest itself in a technical shape, sometimes – and most annoying – it is mostly that we do not know what we are looking for exactly, and even if we do know, we do not know the exact terminology other people are using to describe it.

I am thankful for the great deal of openness that researchers in machine learning are embarrassing. The discussions through online forums, blogs, tutorials, online courses, and recent books, had definitely made this massive task more tractable.

6.1.3 Lack of Benchmarks, evaluation metrics

Getting around these issues was quite a dilemma, for many reasons:

- I do not believe it is a healthy practice to pick up the benchmarks to use. This choice can be easily biased, and it could be argued that the benchmarks are chosen to be weak enough in order to show progress). In our case, it was mandatory to do so nonetheless, and we tried as much as possible to be fair in making these choices.
- In engineering, it is the right practice to have different teams for design and test of the product. If one team do both, the testing process tend to be biased (i.e., even with the best intention, the team is looking for confirmation of their design, not the problems in it). By analogy, I think this a pitfall of us developing the metrics, the benchmarks, and using them. What if we are mainly looking at the metrics that confirms our hypothesis? It is hard to rule out this possibility.

We tried our best to avoid this when selecting the metrics, and by using multiple metrics to evaluate our hypothesis. However, an independent investigation in this issue will provide is favorable.

6.1.4 Deep learning: theory, hardware and software frameworks

Several excellent frameworks – like *Keras* (Chollet et al., 2015), *PyTorch* (Paszke et al., 2017) and *TensorFlow* (Abadi et al., 2015) – do exist at the moment in order to provide friendly APIs for deep learning, with many online tutorials. This gives the impression that you can just jump in the topic, train a neural network, and now you can harness

[2]Most of the great advances that happened recently in neural networks – especially in generative models – is related to computer vision. Thus, it is more convenient to deal with with handwriting as images than as a dynamic process.

the power of deep learning. However, in my experience, the quick gains of this approach will be lost soon in the face of the first problem. Even for someone experienced with traditional machine learning, deep learning poses an extra challenge: it can be computationally expensive. More thinking is needed about what to do and what not to do.

It is important to understand the fundamental of statistical learning theory (Hastie et al., 2001), machine learning and deep learning before engaging in an endeavor that uses deep learning. This particular strategy was a key factor in any progress done during my PhD. There are plenty of excellent online courses, free books and many resources, that provide a gradual and methodological approach, which a learner can use in order to achieve this.

Another thing to highlight here is the hardware needs. Having access a GPU is a necessity in order to learn, experiment and develop using deep learning. Recently, interesting cloud-based solutions – like *Colab*[3], which is free – provide quick access to hardware suitable for deep learning. Some other solutions that I used – like *Amazon AWS*, which is not free – does exist, with the advantage of being easily configured and scalable, and with a support from the research institution, could provide a good replacement for buying and maintaining expensive hardware in-house. The bottom line is: it is important to keep in mind the hardware available, otherwise, the whole process will be hindered.

Another aspect to consider is the framework to use. *Keras* for example provide interesting high-level APIs, while *PyTorch* and *TensorFlow* provide low-level APIs. *PyTorch* focuses more on being close to the Python language way of thinking, while *TensorFlow* provide a wide variety of interesting functions[4]. It is lucrative to go for *Keras*, but once a low-level development is needed, I find that it adds an unpleasant overhead, requiring a mastery level of the underlying framework. Besides, in my opinion, starting by working on high-level directly encourages bad practices – since everything is done in background, it is easy to bypass important details in the way deep learning works, thus, developing poor debugging skills –. Discovering *PyTorch* was, by far, the unspoken hero in this PhD, and one of the best engineering decisions I have ever made. My point from discussion is illuminate the different trade-offs between the different platform. No one is better than the other. It is important to understand the task in hand, and choose the suitable tool for it.

[3] `https://colab.research.google.com/`

[4] This gap between *PyTorch* and *TensorFlow* is closing, with every new version of both.

6.2 Limitations of the current work

In this section, I discuss what I consider shortcomings for some of the methods used in this work.

6.2.1 Style extraction and exploration using PCA and tSNE methods

In this work, when exploring the latent space of our model, I used either PCA or tSNE projection methods (to project the latent space from the high dimensional space into a smaller one), and tried to use the assumptions behind both methods to extract meaningful information from the latent space.

While this is legit, it really stretches these methods to the breaking point, plus, it hinders further investigation.

PCA assumes orthogonality and linearity in the space being projected. There is no reason however to assume that these assumptions hold for different styles. In the non-linearity aspect, the latent space does not have the clear objective of transferring non-linear style relationship into linear ones (simply, because no such objective can be formulated directly, since the problem of styles is ill-defined), unlike what can be noticed for the last layers of classifiers (where an embedded objective of the network is project the data from their non-linear manifold into a linear one). Finding orthogonality in the style space is an interesting aspect to explore, but this is a strong assumption, and there is no reason to believe that it holds for all aspects of styles.

tSNE provide us with a way to deal with non-linearity, thus allowing another further exploring the latent space, but it is hard to repeat the results (the method is stochastic) and the projection does not always yield clear information about the styles. Changing the *perplexity* parameter leads to different results as well (I didn't explore the relation of that parameter to find a more suitable style manifold, and I am not sure if it is worth the effort).

But what is a good projection criteria in this case? should we let the organization of styles emerge on its own, by constraining the latent space and add regularization to the loss function (i.e., during an end-to-end training of the network)? should a second optimization step be performed on the latent space, in order to disentangle it? I discuss some of these ideas briefly in section 6.3.

6.2.2 Leak in the style module

In a basic autoencoder (no condition on the bottleneck), one can assume that encoder part will learn the identity of the task + the style in the same time. The idea of conditioning is to provide the task description (aka: task identity) as an input to the decoder (the condition), thus, relieving the encoder from learning it, and focus only on learning the styles, thus enhancing the style transfer capability.

Ideally, we expect that the output of the encoder has little to none information about the task identity. However, careful testing shows that this is not the case. There is a considerable leak of information about the task identity into the encoder.

I do not have an explanation at the moment for the reason behind this phenomena. My intuition[5] is that on aspect of the problem lies in the task description. The assumption that a harsh one-hot encoding of the task is sufficient to describe the task correctly is flawed in my opinion.

An analogy for this can been drawn from clustering (hard clustering VS fuzzy clustering). Hard clustering, similar to one-hot encoding, does provide us with which this task is, but nothing about how this task relates to other tasks (i.e., proximity/similarity to other task), which is what fuzzy clustering do.

The influential work done by Geoffrey Hinton in (Hinton et al., 2015) – performed on the MNIST dataset (LeCun and Cortes, 2010) – is a contributing factor in this intuition. I will not dive into details about this article here, since it is outside the scope of this work, I will just mention two interesting results from this study:

- In a classification task, the traditional description of the labels is one-hot encoding. However, using a soft/fuzzy description of the labels reveals much better results (makes sense, since it is more rich in information).
- If you train a classifier on the soft labels of digits 7 and 8 only, the classifier will perform almost 90% accuracy on the other labels [6]!. It means that a better task description may increase the data efficiency of the model.

A similar concept should definitely be explored in the context of this work.

6.3 Future directions

Separate the technical parts (very tactical/low-level) from the research parts (strategic/high-level/long-range)

[5] I did not have the time to perform rigorous testing for this idea unfortunately.
[6] I personally find this particular result fascinating.

- Embedding (robust generalization/maybe style extraction)
- Multi-stage optimization (for style extraction) (research)
- Disentanglement of styles / latent space distributions / loss OR constraint (for knowledge extraction): interesting for extracting new styles AND building behavior generators. (research)
- RL and planning to reduce the complexity/have control over the generation and evaluation process (not sure)
- Memories of the neural network for better task decomposition (useful for transfer learning) – maybe report my experiment with the external memory here. (research)
- Statistical testing for neural network performance and for the inference quality + the effect of the random seed. (engineering/deployment)
- Better task description (research)
- Automatic separation of task and style (research)
- Data efficiency and Model Complexity (engineering/deployment)
- Define data augmentation protocol: while many techniques exists for images, it is not so clear in case of sequences. (research)
- **Treating dynamical system as an image, with coloring gradients representing the dynamics of the system**: this will give us massive power in using the SOTA of generative models developed for static images (based on GANs), while addressing all the characteristics of the dynamical system in the same time.
  This is similar to treating sound waves as images (the ML course I did with Washington University), or spectrum images (like what i did with Marielle).
- Dealing with style extraction as an embedding problem: **NOTE: i am not sure if we didn't do that already. Probably in the shaping of the latent space or the loss function. Or maybe it is already okay now.**
- Modeling the error distribution of the model during the generation phase: in order to enhance the quality of the generation, and reduce the distance between the generation and the prediction (i.e., the training and the usage phases) [7].
- World models paper: having a model of the world, and use it to train a proper generator using RL.
- Re-visiting MDN (even the in discrete case): a useful tool, even for finite-discrete distribution. Mixed with good embedding, and giving 'uncertainty' level, provides important information. This also has a very good potential for optimization.
- Using RL for generation instead of log-likelihood models, and free the metric and the study from the constraints of differentiability.
- The hyper-parameter tuning: Do we need it? And how to do it efficiently? (yes we do need it. the question is how to do it in an effecient manner).
- Evaluate the quality of the content/task generated, not just the style: usually done via subjective testing or using another model – which i strictly refuse –. Subjective testing is too slow.
- What do these numbers actually mean in real life? how much should we care? (Wagstaff,

[7] nothing, quick test

2012) (machine learning that matters paper): we used multiple metrics in order to compare/assess our models. But the main issue is how much this really matter? at how much $X\%$ difference in the performance that humans start to perceive the difference for example? This is very important in order to reach ML that actually matters in real life.

- The experimental protocol and scientific practices:
- Use of CNN instead of RNN for the style extraction (embedding) part: more flexible and faster, many off-the-shelf components to choose from (in order to build the network)
- Changing the name of the BLEU score metric, since it created a confusion with the way we use it, relative to the way to was created for.

Chapter 7

Closing Remarks

- Summarize the vision and the objectives of the PhD
-

Appendix A

Hyper-parameter tuning

Machine learning in its core is a search in the hypothesis space of a particular algorithm, in order to find the suitable parameters/hyper-parameters that best fit the data, while adhering the rules of statistical learning theory (Hastie et al., 2001). For parameters tuning in neural networks (i.e., learning the weights of the networks), back-propagation (Rumelhart et al., 1988) is one of the well established algorithms most commonly used[1].

Yet, when it comes to hyper-parameters, it is not that obvious[2]. There is two main strategies for finding hyper-parameters[3]:

- Babysitting the model: for people with small computational
- Using a search strategy: one of the dominant search strategies in this case is a random search in the hyper-parameter space.

[1] Several other algorithms do exist, like evolutionary algorithms (Eiben et al., 2003) – which recently are achieving remarkable results –.

[2] By not obvious, I am politely meaning it was an utter suffering!

[3] As wonderfully explained by 'Andrew Ng' in his *Deep Learning Specialization* in *Coursera* platform, `https://www.coursera.org/specializations/deep-learning`

Appendix B

List of publications

This is a list of publications done during this PhD

B.1 International Conferences

- O.Mohammed, G.Bailly, D.Pellier. *Style transfer and extraction for the handwritten letters using deep learning*. 2019 ICAART (International Conference on Agents and Artificial Intelligence), Prague, Czech Republic.
- O.Mohammed, G.Bailly, D.Pellier. *Handwriting styles: benchmarks and evaluation metrics*. Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2018. Valencia, Spain.
- Gerazov, B., G. Bailly, O. Mohammed and Y. Xu (2018), *Embedding Context-Dependent Variations of Prosodic Contours using Variational Encoding for Decomposing the Structure of Speech Prosody*, 2018 Workshop on Prosody and Meaning: Information Structure and Beyond, Aix-en-Provence, France
- Marielle MALFANTE, Omar MOHAMMED, Cédric GERVAISE, Mauro DALLA MURA, Jérôme I. MARS, *Use of deep features for the automatic classification of fish sounds*, 2018 OCEANS - MTS/IEEE Kobe Techno-Ocean
- Omar Mohammed, Gerard Bailly, Damien Pellier, *Acquiring Human-Robot Interaction skills with Transfer Learning Techniques*, HRI '17 Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction

B.2 Journal articles

B. Gerazov, G. Bailly, O. Mohammed, Y. Xu and P. Garner, *A Variational Prosody Model for the decomposition and synthesis of speech prosody*, to be submitted.

B.3 Infomal communication

- RHUM Workshop, presentation, 2017
- RHUM Workshop, poster, 2018
- FADEX sur l'AI, presentation, 2018
- RHUM Workshop, presentation, 2019
- JDD of DPC, presentation, 2019

Appendix C

Adversarial evaluation

In this part, I mention an experiment I performed, in order to investigate the problem of using an oracle to evaluate the generator, and use the quality assessment of the oracle as a feedback to the generator, in order to improve it. I will first describe what is the problem

Bibliography

Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., and Shah, M. (2018). Video description: a survey of methods, datasets and evaluation metrics. *arXiv preprint arXiv:1806.00186*.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability.

Bailly, G., Mihoub, A., Wolf, C., and Elisei, F. (2018). Gaze and face-to-face interaction. In Oben, G. B. . B., editor, *Eye-tracking in Interaction. Studies on the role of eye gaze indialogue*, pages 139 – 168. Benjamins.

Bailly, G., Raidt, S., and Elisei, F. (2010). Gaze, conversational agents and face-to-face communication. *Speech Communication*, 52(6):598–612.

Bishop, C. M. (1994). Mixture density networks. *Technical Report. Aston University, Birmingham*.

Blitzer, J., McDonald, R., and Pereira, F. (2006). Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics.

Chang, B., Zhang, Q., Pan, S., and Meng, L. (2018). Generating handwritten chinese characters using cyclegan. *CoRR*, abs/1801.08624.

Chattopadhyay, R., Sun, Q., Fan, W., Davidson, I., Panchanathan, S., and Ye, J. (2012). Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):18.

Cho, K. (2013a). Boltzmann machines and denoising autoencoders for image denoising. *arXiv preprint arXiv:1301.3468*.

Cho, K. (2013b). Simple sparsification improves sparse denoising autoencoders in denoising highly corrupted images. In *International Conference on Machine Learning*, pages 432–440.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chollet, F. (2017). *Deep Learning with Python*. Manning Publications. https://www.manning.com/books/deep-learning-with-python.

Chollet, F. et al. (2015). Keras. https://keras.io.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Churamani, N., Anton, P., Brügger, M., Fließwasser, E., Hummel, T., Mayer, J., Mustafa, W., Ng, H. G., Nguyen, T. L. C., Nguyen, Q., et al. (2017). The impact of personalisation on human-robot interaction in learning scenarios. In *Proceedings of the 5th International Conference on Human Agent Interaction*, pages 171–180. ACM.

De Sa, C. (2017). Non-convex optimization.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Denil, M., Bazzani, L., Larochelle, H., and de Freitas, N. (2012). Learning where to attend with deep architectures for image tracking. *Neural computation*, 24(8):2151–2184.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Diaz, M., Ferrer, M. A., Parziale, A., and Marcelli, A. (2017). Recovering western on-line signatures from image-based specimens. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1204–1209. IEEE.

Dörr, D., Grabengiesser, D., and Gauterin, F. (2014). Online driving style recognition using fuzzy logic. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1021–1026. IEEE.

Eiben, A. E., Smith, J. E., et al. (2003). *Introduction to evolutionary computing*, volume 53. Springer.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Fogel, I. and Sagi, D. (1989). Gabor filters as texture discriminator. *Biological cybernetics*, 61(2):103–113.

Freeman, H. (1961). On the encoding of arbitrary geometric configurations. *IRE Transactions on Electronic Computers*, 2:260–268.

Gallaher, P. E. (1992). Individual differences in nonverbal behavior: Dimensions of style. *Journal of personality and social psychology*, 63(1):133.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2015). Texture synthesis and the controlled generation of natural stimuli using convolutional neural networks. *CoRR*, abs/1505.07376.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1243–1252. JMLR. org.

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.".

Glorot, X., Bordes, A., and Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 513–520.

Gondara, L. (2016). Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 241–246. IEEE.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Google (2017). The quick, draw! dataset.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Ha, D. (2015). Mixture density networks with tensorflow. *blog.otoro.net*.

Ha, D. and Eck, D. (2017). A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*.

Ha, D. and Schmidhuber, J. (2018). World models. *arXiv preprint arXiv:1803.10122*.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, 57(2):243–259.

Hinton, G., Srivastava, N., and Swersky, K. (2012). Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Jain, A. K. and Farrokhnia, F. (1991). Unsupervised texture segmentation using gabor filters. *Pattern recognition*, 24(12):1167–1186.

Johnson, D. A. and Trivedi, M. M. (2011). Driving style recognition using a smartphone as a sensor platform. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1609–1615. IEEE.

Jolliffe, I. (2011). Principal component analysis. In *International encyclopedia of statistical science*, pages 1094–1096. Springer.

Jonas Jongejan, Henry Rowley, T. K. J. K. N. F.-G. w. f. a. G. C. L. and Team, D. A. (2017). The quick, draw! game.

Karpathy, A. and Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Kashi, S. and Levy-Tzedek, S. (2018). Smooth leader or sharp follower? playing the mirror game with a robot. *Restorative neurology and neuroscience*, 36(2):147–159.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klakow, D. and Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28.

Konidaris, G., Kuindersma, S., Grupen, R., and Barto, A. (2012a). Robot learning from demonstration by constructing skill trees. *The International Journal of Robotics Research*, 31(3):360–375.

Konidaris, G., Scheidwasser, I., and Barto, A. G. (2012b). Transfer in reinforcement learning via shared features. *J. Mach. Learn. Res.*, 13:1333–1371.

Kostadinov, S. (2017). How recurrent neural networks work.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Larochelle, H. and Hinton, G. E. (2010). Learning to combine foveal glimpses with a third-order boltzmann machine. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 1243–1251. Curran Associates, Inc.

Lathuiliére, S., Mesejo, P., Alameda-Pineda, X., and Horaud, R. (2019). A comprehensive analysis of deep regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Li, S. and Zong, C. (2008). Multi-domain adaptation for sentiment classification: Using multiple classifier combining methods. In *2008 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1–8. IEEE.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2014). Microsoft COCO: Common Objects in Context. *ArXiv e-prints*.

Long, M., Wang, J., Ding, G., Sun, J., and Yu, P. S. (2013). Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 2200–2207.

Lowe, D. G. et al. (1999). Object recognition from local scale-invariant features.

Malfante, M. (2018). *Automatic classification of natural signals for environmental monitoring*.

Malfante, M., Dalla Mura, M., Mars, J. I., and Gervaise, C. (2016). Automatic fish sounds classification. *The Journal of the Acoustical Society of America*, 139(4):2115–2116.

Malfante, M., Dalla Mura, M., Métaxian, J.-P., Mars, J. I., Macedo, O., and Inza, A. (2018a). Machine learning for volcano-seismic signals: Challenges and perspectives. *IEEE Signal Processing Magazine*, 35(2):20–30.

Malfante, M., Mohammed, O., Gervaise, C., Dalla Mura, M., and Mars, J. I. (2018b). Use of deep features for the automatic classification of fish sounds. In *2018 OCEANS-MTS/IEEE Kobe Techno-Oceans (OTO)*, pages 1–5. IEEE.

Manzoni, V., Corti, A., De Luca, P., and Savaresi, S. M. (2010). Driving style estimation via inertial measurements. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 777–782. IEEE.

Marti, U.-V. and Bunke, H. (1999). A full english sentence database for off-line handwriting recognition. In *Document Analysis and Recognition, 1999. ICDAR'99. Proceedings of the Fifth International Conference on*, pages 705–708. IEEE.

Martinez, C. M., Heucke, M., Wang, F.-Y., Gao, B., and Cao, D. (2017). Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):666–676.

Mihoub, A., Bailly, G., Wolf, C., and Elisei, F. (2016). Graphical models for social behavior modeling in face-to face interaction. *Pattern Recognition Letters*, 74:82–89.

Molnar, C. (2019). *Interpretable Machine Learning*. `https://christophm.github.io/interpretable-ml-book/`.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Narang, S. and Gupta, M. D. (2015). Speech feature extraction techniques: a review.

Neubauer, J. and Wood, E. (2013). Accounting for the variation of driver aggression in the simulation of conventional and advanced vehicles. Technical report, National Renewable Energy Lab.(NREL), Golden, CO (United States).

Nguyen, D. C., Bailly, G., and Elisei, F. (2017). Pattern Recognition Letters Learning Off-line vs. On-line Models of Interactive Multimodal Behaviors with Recurrent Neural Networks. *Pattern Recognition Letters*, 100:29–36.

of ISCA (the International Speech Communication Association), S. S. I. G. (2019). Blizzard challenge 2019.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016a). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Oord, A. v. d., Kalchbrenner, N., and Kavukcuoglu, K. (2016b). Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q., and Chen, Z. (2010a). Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World wide web*, pages 751–760. ACM.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2010b). Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., and Miller, J. (2017). Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*.

Raina, R., Madhavan, A., and Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th annual international conference on machine learning*, pages 873–880. ACM.

Ranzato, M., Chopra, S., Auli, M., and Zaremba, W. (2015). Sequence level training with recurrent neural networks. *CoRR*, abs/1511.06732.

Ribeiro, M., Lazzaretti, A. E., and Lopes, H. S. (2018). A study of deep convolutional auto-encoders for anomaly detection in videos. *Pattern Recognition Letters*, 105:13–22.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1.

Sakurada, M. and Yairi, T. (2014). Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, page 4. ACM.

Salakhutdinov, R. and Hinton, G. (2009). Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978.

Séraphin-Thibon, L., Gerber, S., and Kandel, S. (2019). Analyzing variability in upper-case letter production in adults. In *Spelling and Writing Words*, pages 163–178. BRILL.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., and Saurous, R. A. (2018). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. *CoRR*, abs/1803.09047.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014a). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014b). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Tachibana, M., Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T. (2004). Hmm-based speech synthesis with various speaking styles using model interpolation. In *Speech Prosody 2004, International Conference*.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). A survey on deep transfer learning. *ArXiv*, abs/1808.01974.

Taylor, M. E. and Stone, P. (2007). Cross-domain transfer for reinforcement learning. In *Proceedings of the 24th international conference on Machine learning*, pages 879–886. ACM.

Taylor, P. (2009). *Text-to-speech synthesis*. Cambridge university press.

Thórisson, K. R. (2002). Natural turn-taking needs no manual: Computational theory and model, from perception to action. In *Multimodality in language and speech systems*, pages 173–207. Springer.

van den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315.

van der Maaten, L. and Hinton, G. (2008). Visualizing high-dimensional data using t-sne.

Viard-Gaudin, C., Lallican, P. M., Knerr, S., and Binter, P. (1999). The ireste on/off (ironoff) dual handwriting database. In *Document Analysis and Recognition, 1999. ICDAR '99. Proceedings of the Fifth International Conference on*, pages 455–458.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014). Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Wagstaff, K. (2012). Machine learning that matters. *arXiv preprint arXiv:1206.4656*.

Wang, W., Xu, S., and Xu, B. (2016). Gating recurrent mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5520–5524. IEEE.

Wang, X., Takaki, S., and Yamagishi, J. (2017a). An autoregressive recurrent mixture density network for parametric speech synthesis. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4895–4899.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. (2017b). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., and Saurous, R. A. (2018). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*.

Weiss, K., Khoshgoftaar, T. M., and Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, 3(1):9.

Wikipedia (2019a). Convex funcation.

Wikipedia (2019b). Gradient descent.

Williams, R. J. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1(2):270–280.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., and Bengio, Y. (2015a). Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

Xu, L., Hu, J., Jiang, H., and Meng, W. (2015b). Establishing style-oriented driver models by imitating human driving behaviors. *IEEE Transactions on Intelligent Transportation Systems*, 16(5):2522–2530.

Xu, Y., Du, J., Dai, L.-R., and Lee, C.-H. (2014). Cross-language transfer learning for deep neural network based speech enhancement. In *The 9th International Symposium on Chinese Spoken Language Processing*, pages 336–340. IEEE.

Zen, H. and Senior, A. (2014). Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 3844–3848. IEEE.