The University of Adelaide

Project Proposal

From Script to Movie

Miguel Martin

supervised by

Qinfeng Shi and Dong Gong

August 17, 2018

Contents

1 Introduction

In this research proposal, we will discuss the problem of converting a script to a movie in detail, focusing the challenges one has to consider when developing a solution. At a high level, the problem we are tasked to solve is to convert text, which represents some script or section of a script, to some generated video corresponding to the input text. We will also mention restrictions with respect to data and computational resources we have will have access to, in order develop and evaluate our solution. The utilisation of Machine Learning will be the main focus point of the research.

Currently, there are multiple ways of generating images from nothing (random latent variables), text and other inputs through many different generative models (TODO: references). These methods have amazing results and have been improving rapidly over time. Unfortunately, there are significant limitations to these approaches such as performing well only to one class, instability in training or being computationally expensive. Naively taking some of these solutions and scaling them to produce videos will have some draw-backs, without considering the problem of generating a video further in depth.

In this proposal, I hypothesise that adding specific domain knowledge to influence the structure of the problem will help produce state of the art results. This is hypothesised due to two factors: data and reduction in the complexity of the search space; even large datasets will likely not cover a big proportion of the total size of the input and output space. Specifically, the use of so-called 'Scene Graphs' will be utilised, similar to that in (TODO: ref).

2 Motivation

From the perspective on the current state of research, it should be clear that utilising Machine Learning is key to solve this problem. This is due to the fact that most NLP problems are solved with Machine Learning (TODO: references), and the Computer Vision tasks required for this problem (generation of images) are also currently being tackled with Machine Learning with state of the art results. It is unclear how one would solve this without Machine Learning, with scalability in mind (relative to problem complexity). With Machine Learning, even if a simplified version of the problem is considered, parts of the approach can be improved upon in allow for the problem to be increasingly more complex. This is evident through recent advancements with generative models, e.g. Generative Adversarial Networks (GANS) (TODO: reference) have been iterated and improved upon relative to generative quality, training stability, variations of the problems, etc. (TODO: references)

In order to generate videos from text, one must truly understand (model) the world which he is generating images of. This is due to the implicit nature of natural language, that is we must be able to infer physical relationships between entities (objects, the environment and characters) and other characteristics such as stereotypes/biases. The main root cause of the complexity for the problem stems from data, even with large datasets there is .

This problem is quite complicated, with the biggest challenge stemming from

3 Related Work

Generative Models TODO: GANs + variants/improvements, text to image

Video Generation from Text TODO: (use author?)

Imagine This! Scripts to Composition to Videos TODO: (use author?)

sg2img TODO

TODO: for existing work: go through how they solve the problem, their assumptions, and the limitations with each approach

TODO: for scene graph, go through what this is and how it helps model complicated scenes for text to image. Mention also the following limitations:

1. Only converts scene graph to image. Doesn't convert text to a scene graph
2. Scene graph is synthetic for one of the datasets
3. If we want to utilise this, we must do something about this (predict the scene graph in some way) or just use data that has scene graphs

4 Challenges

Clearly there are a lot of challenges required to overcome in order to solve this problem. To understand what challenges we face, let us analyse the following two sentences: 'Bob swung his golf club. Bob threw his club in frustration.':

1. *Data and Generalisation/Overfitting*
   We must realise that the number of input/output pairs in any real-world (supervised) dataset will be very small in comparison to the possible set of input/output pairs. In fact, if you let the time be unrestricted then there is an infinite possible examples; this is true even for other problems such as audio generation, however, the size of this 'infinity' (TODO: reference sizes of infinity?) is quite large due to the output being a sequence of images rather than a sequence of one number. Thus over-fitting will likely be a big issue that is required to overcome.

2. *Feasibility and computational complexity**
   The generation of the video must be performed in a realistic time frame.

3. *Names/aliases of entities**
   Names and aliases might be a word, small set of words, or a description of the entity being referenced. Ambiguity may be an issue in some pieces of text, which may be done on purpose by the author; in this case, interpretation is required, which is discussed in challenge 7.

   In the example, we must recognise that 'Bob' is a name referencing a human being (character). Also, 'golf club' and 'club' are referring to the same object, due to context.

4. *Visual consistency over time**
   Visually everything should make sense and be consistent throughout the entire video sequence. This ties in with challenge 7.

   For example: Bob cannot suddenly change his appearance from one frame to the next. We must also realise that Bob in the second sentence is the same Bob in the first sentence (ties in with challenge 3).

5. *Layout consistency over time**
   The layout (positioning, size) of entities should make logical sense throughout the entire video sequence. Similar to the previous challenge, this ties in with challenge 7.

   The golf club should remain in Bob's hand(s) throughout the entire sequence (until he throws it), and Bob should remain stuck to the ground of the environment, i.e. he should not spontaneously float or teleport.

6. *Relationships/interactions between entities**
   It is implicitly defined that Bob is holding the golf club in his hand(s) and Bob is on the floor of the environment. It is also implied that when Bob 'swings' his club he does so through his shoulders with some particular form.

   At a high level this interaction can be described as a scene graph, similar to that as done in TODO: ref, for example: Bob $\rightarrow$ club $\rightarrow$ ball $\rightarrow$ environment, where $\rightarrow$ describes an interaction. The nodes of the directed graph describe an entity and a directed edge between two entities describes an interaction (event).

7. *Interpretation of missing entities and details/attributes*
   Interpretation can be thought of as random assignments to variables and/or the most probable scenario, e.g. it is more likely that Bob is an old man due to his name and because he is playing golf. The line between implicit and interpretation is a bit fuzzy.

   Other missing details include: the environment (probably a golf course), if Bob is a new or experienced golf player, Bob's outfit, the material of the golf club, colour of the golf ball, etc.

8. *Quality of the appearance of entities*

5 Hypothesis

I hypothesise that to achieve state of the art results, utilisation of domain knowledge will be required. The recent generative model sg2img (TODO: ref) supports this hypothesis, and as such a similar approach

will be taken.

Another hypothesis to be tested, if time permits, stems from the 'Imagine This!' paper's claim (TODO: ref). The paper specifies that pixel generative models increases the complexity of the problem too much, which I believe to be false.

6 Approach

From this small example we have accumulated a large number of challenges to resolve. This gives us the intuition that scaling up to larger sequences of text has much higher complexity with the existing challenges and potentially other challenges, e.g. a script for a Hollywood or short film. Due to this, this implies a simplification and breakdown of the problem is required, which is done with existing work (TODO: reference papers).

The presented challenges give us a reasonable breakdown of the problem, thus I propose that each of these challenges should be addressed with a given solution. Some of the challenges may require simplification, where others should have more emphasis due to the larger potential impact on the video generated. For example, focusing on the quality of appearance does not show a true understanding of what the text represents. Specifically, the order given in Section 4 show the priority of each challenge. If time permits and each of the main challenges are addressed to a sufficient manner, more focus will be done on less important challenges.

TODO: mention data

6.1 sg2video

In order to convert the scene graphs to videos rather than images, there are two possible ways to approach this:

1. Text to Scene Graph to [Frames]
   TODO: describe
2. Text to [Scene Graph] to Frame TODO: describe

These two approaches are equivalent. The first approach is much more flexible.

One might model the second approach a bit different, where you produce two graphs, the initial layout the scene graph and then

7 Research Methodology

7.1 Metrics For Evaluation

7.2 Issues

8 Conclusion

Appendices

A Definitions

A.1 Entity

An object, character or environment. Each class of entity has four main shared attributes which define them. All attributes depend on time and past events:

- **Position:** The position of where an object is relative to the world it is placed in.
- **Size/Shape:** The size of the entity, relative to the world it is placed in.
- **Visual Appearance:** Closely related to the size/shape of an entity. The visual appearance of the entity is what the object looks like. Text is very limited in what you can describe, which leaves a lot up for interpretation. For example, a 'red ball' doesn't tell you much about the texture of the ball, what shade of red it is, if the ball is a football or a soccer-ball, etc.
- **Name/Aliases**: An entity may go under multiple names/aliases. This is a consequence of verbal and written language, such as English both in it's formal and informal contexts. For example, when referencing a particular person by name, rarely will their last name be used in conjunction with their first. Conversely, if one doesn't know the person, one might describe the appearance of the person, which could be considered an alias.

  Note that aliases may be ambiguous, and in the context of a script (to tell a story), ambiguity may be done so on purpose.

There may be other attributes which define entities, but these additional attributes are likely to be specific to the particular class of the entity, such as the nature of a character (charismatic, boring, etc.).

A.2 Events

Events are entities' interactions (and reactions) between entities.

A.3 World

Where and when the entities live in.

B Machine Learning

TODO

B.1 Generalisation

TODO

C Script to Film

TODO