Capstone Project: Project Proposal

Section 1: Project Overview

We are hoping to undertake a project with a supervisor, Dr. Taylor, from the university of Guelph. Dr. Taylor is a researcher in the software engineering department at Guelph who studies machine learning. Dr. Wen Bo He, from McMaster University, will also help us as a co-supervisor on this project.

One project that Dr. Taylor is involved in is a “text to motion” subcomponent of a larger collaboration between his own university -- the University of Guelph, SRI (a non-profit research organization in the US), and other institutions. This aforementioned collaboration was established with the goal of producing a “Computational Storytelling” system.

The input to such a “Computational Storytelling” system is intended to be a five line basic story in natural language and the output will be an animated movie. The content of the output movie will of course be based on the natural language input. The system will involve the use of machine learning or a statistical model that will perform a mapping between natural language and video.

Section 2: Problem Statement and Our Contribution

The problem that we seek to overcome in this project is that in order to implement the overall “Computational Storytelling”, there is no good existing database linking videos, rich text annotations, and descriptions of motions (e.g. human pose estimations via join positions).

As our team’s contribution to the overall “Computational Storytelling” project, we propose to use video as an intermediate modality in order to harvest motion data with rich text annotations since large video databases do exist (with dialogue, script, or closed-captioning annotations). We will use one (or several-) of such databases, in concert with off-the-shelf pose estimation and tracking methods to capture a large database of text-motion pairs.

Section 3: Motivation

The motivation behind this project is that this ability to move between natural language and video representations of information is an important task the human brain performs when trying to reason about the future or when analyzing the past. Since this ability is an important part of our own reasoning it may then be useful in computer applications that perform analysis to be able to take a natural language problem and then solve it in a video representation. Especially if the problem involves motion or some type of physical experience that relies on data that is stored as video.

Section 4: Background Information and Challenges

Over the last ten years machine learning algorithms have significantly improved as well as the limits of our computer processing power, computer memory and available sources of many types of data. The growth in all these areas has made it possible for computer applications that use machine learning to improve their performance.

The key advancement in the algorithms of machine learning has been the ability to train neural networks that are more than a single layer deep. With this ability neural networks are able to learn structures from long sequences of sequential data(long sequences of inputs such as natural language and video in our case).

Many different researchers have already been able to create a mapping between text and images (Kiros *et al*., 2014a, b; *Socher et al.,* 2010). However these examples simply map from image to text but the problem of performing the mapping using videos is that there is now a new dimension of time and the volume of data grows very quickly as the length of the video increases.

The databases that we intend to use for our project which have dense labelling and for supervised learning algorithms are:

- MovieQA
- Charades
- MSR-VTT

Section 5: Objectives, Constraints and Deliverables

The goal of the entire collaboration is to create a basic text-to-motion tool which can link language concepts to a data structure representing motion (e.g. a series of joint angles). The first step in analyzing motion from a machine learning approach is typically to create a feature space -- which is a lower dimensional representation of some of the important aspects of the data which the machine learning algorithms can then analyze.

The specific feature space we wish to consider is called a pose structure (Wei, 2016). This structure is used to model the motion of people. The feature space essentially consists of a labelling of the regions in an image where a specific limb is located (arm, head eye etc). Modelling the motion of the entire body is made much simpler when the problem can be decomposed into the “independent” motion of these features (applying the constraints that the features are connected in the conventional human form).

Therefore in order to achieve this first step we hierarchically break this step into two substeps.

Deliverable components

1. Database and web-based interface
    a. Allows one to navigate between clips with human motion
    b. Shows skeletons annotated on top of video
    c. Shows text matching clip

    d. Database should work well with our existing Python learning-based workflows
    e. ML web-based interfaces that we will hope to use as examples for our own product:
        i. Hedonometer
        ii. LSTMVis
        iii. TensorFlow playground

2. Exploration of existing pose estimation and tracking methods (2d and 3d) that can find people and extract their motion in the video clips
    a. Deep learning-based methods are preferred, because we may decide to integrate these techniques into our text-to-motion pipeline1
    b. References we will use as starting points:
        i. (Pfister, Charles, and Zisserman 2015)
        ii. (Wei, Ramakrishna, and Kanade 2016)
        iii. (Belagiannis and Zisserman 2016)1
        iv. (Tekin et al. 2016)
        v. (Zhou et al. 2015)
        vi.

These sub-goals reflect the requirement that we wish to create an interactive database that contains thousands of examples of our data structure (video motion that is labelled with the poses of the objects in the video). These examples will then be used in the larger collaboration for a supervised learning algorithm that will perform the natural language to video mapping.

Section 6: Bibliography

Belagiannis, V., and A. Zisserman. 2016. “Recurrent Human Pose Estimation.” *arXiv Preprint arXiv:1605.02914*. arxiv.org. http://arxiv.org/abs/1605.02914.

Jain, Ashesh, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena. 2016. “Structural-RNN: Deep Learning on Spatio-Temporal Graphs.” In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014a). Multimodal neural language models. In *ICML’2014*. 102

Kiros, R., Salakhutdinov, R., and Zemel, R. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv*:1411.2539 [cs.LG]. 102, 410

Pfister, Tomas, James Charles, and Andrew Zisserman. 2015. “Flowing ConvNets for Human Pose Estimation in Videos.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1506.02897.

Socher Richard, Karpathy Andrej, Le Ve Quoc, Manning D. Christopher, Ng Y. “Connecting modalities: Semi-supervised segnmentation and annotation of images using un-aligned text corpora.” CVPR (2010).

Taylor, Graham W., Geoffrey E. Hinton, and Sam T. Roweis. 2011. “Two Distributed-State

Models For Generating High-Dimensional Time Series.” *Journal of Machine Learning Research: JMLR* 12 (Mar): 1025–68.

Tekin, Bugra, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. 2016. “Structured Prediction of 3D Human Pose with Deep Neural Networks.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1605.05180.

Wei, S. E., V. Ramakrishna, and T. Kanade. 2016. “Convolutional Pose Machines.” *arXiv Preprint arXiv:* arxiv.org. http://arxiv.org/abs/1602.00134.

Zhou, Xiaowei, Menglong Zhu, Spyridon Leonardos, Kosta Derpanis, and Kostas Daniilidis. 2015. “Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1511.09439.