NADiA - Towards Neural Network Driven Virtual Human Conversation Agents

Socially Interactive Agents Track

Jason Wu
School of Interactive Computing,
Georgia Institute of Technology
Atlanta, GA
jasonwu@gatech.edu

Sayan Ghosh
Institute for Creative Technologies,
University of Southern California
Playa Vista, CA
sghosh@ict.usc.edu

Mathieu Chollet
Institute for Creative Technologies,
University of Southern California
Playa Vista, CA
mchollet@ict.usc.edu

Steven Ly
Institute for Creative Technologies,
University of Southern California
Playa Vista, CA
sly@ict.usc.edu

Sharon Mozgai
Institute for Creative Technologies,
University of Southern California
Playa Vista, CA
smozgai@post.harvard.edu

Stefan Scherer
Institute for Creative Technologies,
University of Southern California
Playa Vista, CA
scherer@ict.usc.edu

ABSTRACT

Advances in artificial intelligence and machine learning - in particular neural networks - have given rise to a new generation of virtual assistants and chatbots. Within this work, we describe the motivation and architecture of *NADiA* - Neurally Animated Dialog Agent - which leverages both the user's verbal input and facial expressions for multi-modal conversation. NADiA combines a neural language model that generates conversational responses, a convolutional neural network for facial expression analysis, and virtual human technology that is deployed on a mobile phone.

KEYWORDS

Virtual Agent; Chatbot; Neural Language Model; Convolutional Neural Network; Animation

ACM Reference Format:
Jason Wu, Sayan Ghosh, Mathieu Chollet, Steven Ly, Sharon Mozgai, and Stefan Scherer. 2018. NADiA - Towards Neural Network Driven Virtual Human Conversation Agents. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 INTRODUCTION

There has been a growing interest in conversational agents and natural language-based assistant technologies. Indeed technologies such as Apple's Siri, Google Home, and Amazon's Alexa have made their way into people's everyday lives. Early computer conversational agents based on pattern matching and sentence transformation rules showed that computer chatbots could facilitate believable conversation without any semantic and contextual understanding [17]. ELIZA, a 1960s chatbot which simulated conversation with a therapist, was an example of a chatbot able to create the illusion of understanding, leading many early users of ELIZA to attribute some human-like qualities to the program [16]. Since ELIZA, conversational agent technology has matured considerably due to advances in artificial intelligence, natural language processing, and the increased availability of cloud computing and Software as a Service (SaaS) technologies [4, 5]. These more complex and accessible conversational agents, such as Microsoft's Zo[1], a chatbot with a teenager persona, and Woebot[2] a chatbot for cognitive behavior therapy, have found use-cases in entertainment and medicine.

In contrast to conversational agents that rely mostly on text or language based technologies, human face-to-face communication relies on additional modalities that include facial expressions, paralinguistic aspects of the voice (e.g., prosody or voice quality), and gestures [9, 12]. Using computer graphics and speech synthesis technologies, researchers have recently focused on multimodal conversational interfaces to create Embodied Conversational Agents (ECAs), or virtual agents [2], that utilize additional communicative modalities such as natural language and nonverbal behavior (e.g., gestures, facial expressions, postures), to interact with users. These virtual agents possess a number of advantages over standard textual interfaces and have shown to increase rapport and trust of the system [1]; for instance, they were proven to be useful for the screening and treatment of depression and PTSD [11, 13].

The conversational agent architecture proposed in this work, NADiA, relies on neural network research shown to provide state-of-the-art behavior understanding, recognition, and generation [3, 6–8, 14, 19]. In addition, a key motivation of using neural networks for NADiA is the limited computational power needed for deployment, allowing the system to deliver high precision and state of the art performance in low resource environments such as mobile phones or embedded robotic systems. In our work, NADiA was tested and developed on a Samsung Galaxy 7 mobile phone. We leverage the mobile phone's microphone for the automatic speech recognition and camera for the facial expression analysis and facial expression mimicry.

2 NADIA ARCHITECTURE

The NADiA architecture consists of three main parts: (1) NADiA's ability to generate natural language is based on a novel neural

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

[1]https://www.zo.ai/
[2]https://woebot.io/

language model named Affect-LM [6]. (2) NADiA is further capable of mimicking the human user's facial expressions. This capability is enabled through a convolutional neural network that detects the user's face and analyzes his/her facial expressions and renders the same expression on NADiA's virtual face. (3) The appearance of NADiA is enabled through the use of the Smartbody architecture [15]. NADiA is deployed on a mobile phone and is able to respond to human user prompts in near realtime.

NADiA's conversational responses are generated using Affect-LM, a completely data-driven affective language model capable of generating emotionally-colored responses by inferring the affective context from conversation history [6]. Affect-LM has been shown to improve language modeling performance over a state-of-the-art baseline neural language model [18]. Specifically, Affect-LM achieved lower perplexity than a baseline LSTM model when the affect category is obtained from the words in the context. Further, by leveraging MTurk perception studies it was shown that the model can generate expressive text at varying degrees of emotional strength without affecting grammatical correctness across a number of affective categories (e.g. positive and negative affect).

The NADiA virtual human application allows users to engage in affective conversations by making various facial expressions. This is facilitated by a facial expression convolutional neural network (CNN) that processes the video stream from the device's front-facing camera. The facial expression CNN extracts the activations of 18 Action Units (AUs) as defined by the Facial Action Coding System (FACS). These 18 AUs can be used to infer the user's affective state and influence the conversation dialog, or it can be used to perform facial mimicry by directly controlling the virtual human's facial expression.

Finally, the virtual character rendering system is implemented using the Smartbody character animation platform. Smartbody is an open-source project written in portable C++ and is usable on many different platforms, including Android. Smartbody Mobile (SB Mobile) provides a lightweight platform specifically for developing conversational virtual humans. SB Mobile is easily imported into a standard Android application. The behavior generation commands for NADiA are Smartbody scripts written in Python that communicate via Behavior Markup Language (BML), a language for describing verbal and non-verbal character animation behaviors [10]. The character animation system is controlled by the language generation component to generate lip-syncing behavior and the facial expression CNN for facial mimicry and visual affective feedback.

3 SAMPLE CONVERSATIONAL OUTPUT

Figure 1 shows screenshots of the NADiA virtual human application showing the facial expression CNN and virtual character animation system. The virtual character's facial expression is set to mimic the user's by mirroring the input AU activations.

Examples of conversational interactions are shown in Figure 2 as text transcripts. However, the virtual human application uses text-to-speech (TTS) software for a more natural conversational experience. Two conversations are shown, one using the topic of choosing a life partner and one on the topic of cooking as a hobby. Although the language model responsible for response generation

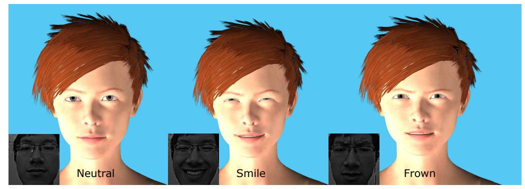

Figure 1: Mobile application screenshots showing facial expressions generated by NADiA, which mimics the detected emotion of the human interacting with it. Detection of Action Units is achieved through a CNN. The CNN input is shown in the left corners as a gray scale image.

Figure 2: Example Conversation Transcripts

```
USER: so what is the most important thing to
look for in a life partner
NADIA: i think that um i would say let's say
if you could you could have a best friend
USER: i think friends are important
NADIA: yeah i do too
USER: are you married
NADIA: i am married
USER: do you have any kids
NADIA: so i i have a young son and so he's
gonna be one of my closest friends
...
USER: do you like to cook
NADIA: yes i do
USER: what kinds of things do you cook
NADIA: i do i do a lot of italian food
USER: i wish i knew how to cook
NADIA: i got i got a little bit of education
here now
```

is completely data driven and lacks any semantic understanding of conversational topics, it is able to produce coherent and believable responses. Indeed the interested reader might notice that the language model produces language that comprises *mistakes*, such as repetitions. This is a result of the training material which was solely conversational text rather than clean written language.

ACKNOWLEDGMENTS

We'd like to thank Ari Shapiro and Arno Hartholt of the USC Institute for Creative Technologies for their support and help developing the mobile prototype of NADiA. This material is based upon work supported by the U.S. Army Research Laboratory under contract number W911NF-14-D-0005. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Government, and no official endorsement should be inferred.

REFERENCES

[1] Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 2 (2005), 293–327.

[2] Justine Cassell. 2000. *Embodied conversational agents*. MIT press.

[3] Jonathan Chang and Stefan Scherer. 2017. Learning Representations of Emotional Speech with Deep Convolutional Generative Adversarial Networks. *arXiv preprint arXiv:1705.02394* (2017).

[4] Robert Dale. 2016. The return of the chatbots. *Natural Language Engineering* 22, 5 (2016), 811–817.

[5] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *interactions* 24, 4 (2017), 38–42.

[6] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. *arXiv preprint arXiv:1704.06851* (2017).

[7] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016. Representation Learning for Speech Emotion Recognition.. In *INTERSPEECH*. 3603–3607.

[8] Sayan Ghosh, Eugene Laksana, Stefan Scherer, and Louis-Philippe Morency. 2015. A multi-label convolutional neural network approach to cross-domain action unit detection. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 609–615.

[9] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal communication in human interaction*. Cengage Learning.

[10] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*. Springer, 205–217.

[11] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.

[12] Albert Mehrabian. 1972. *Nonverbal communication*. Transaction Publishers.

[13] Sharon Mozgai, Gale Lucas, and Jonathan Gratch. 2017. To Tell the Truth: Virtual Agents and Morning Morality. In *International Conference on Intelligent Virtual Agents*. Springer, 283–286.

[14] Najmeh Sadoughi and Carlos Busso. 2017. Joint learning of speech-driven facial motion with bidirectional long-short term memory. In *International Conference on Intelligent Virtual Agents*. Springer, 389–402.

[15] Ari Shapiro. 2011. Building a character animation system. *Motion in Games* (2011), 98–109.

[16] Joseph Weizenbaum. 1966. ELIZA - a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

[17] Joseph Weizenbaum. 1976. Computer power and human reason: From judgment to calculation. (1976).

[18] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* (2014).

[19] Geoffrey Zweig, Chengzhu Yu, Jasha Droppo, and Andreas Stolcke. 2017. Advances in all-neural speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 4805–4809.