Speaking Rationally by Gestures: Information Theoretic Insights from Multi-Modal Language Models

Anonymous ACL submission

Abstract

The multi-modality nature of human communication can be utilized to enhance the performance of computational language models. However, few studies have explored the non-verbal channels with finer theoretical lens. We use multi-modal language models trained against monologue video data to study how the non-verbal expression contributes to communication, by examining two aspects: first, whether incorporating gesture representations can improve the language model's performance (perplexity), and second, whether the gesture channel demonstrates the similar pattern of entropy rate constancy (ERC) found in verbal language, which is governed by Information Theory. We have positive results to support both assumptions. The conclusion is that speakers indeed use simple gestures to convey information that enhances verbal communication, and how this information is organized is a rational process.

1 Introduction

Communication is a multi-modal process, in which information from verbal and non-verbal modalities are mixed into one channel. It has already been revealed in empirical studies that speakers' expression in visual modality, including gestures, body poses, eye contacts and other types of non-verbal behaviors, play critical roles in face-to-face communication, as they add subtle information that is hard to convey in verbal language. However, it remains an untested idea to view these sparse and random non-verbal signals as a formal communication channel that transmits "serious" information, which has seldom been validated by computational studies. A key missing step is to explore whether the non-verbal information can be quantified.

The questions that are worth further investigation include (but are not limited to): How rich is the information contained in these non-verbal channels? What are their relationships to verbal information? Can we understand the meanings of different gestures, poses, and motions embedded in spontaneous language in a similar way to understanding word meanings? The goal of this study is to propose a simple but straight-forward framework to approach the above questions, under the guidance of Information Theory. Some preliminary, yet prospective results are presented.

2 Related Work

2.1 Non-verbal communication in natural language

The recent advances of deep neural network-based machine learning techniques provide new methods to understand the non-verbal components of human communication. Many existing works primarily focus on using multi-modal features as clues for a variety of inference tasks, including video content understanding and summarization (Li et al., 2020; Bertasius et al., 2021), as well as more specific ones such as predicting the shared attention among speakers (Fan et al., 2018) and semantic-aware action segmentation (Gavrilyuk et al., 2018; Xu et al., 2019). More recently, models that include multiple channels have been developed to characterize context-situated human interactions (Fan et al., 2021). Advances in representation learning have enabled researchers to study theoretical questions with the tools of multi-modal language models.

2.2 Insights from cognitive science studies

In laboratory-based studies of interactions between verbal and non-verbal communication, it has been found the multiple layers of visual and vocal signals can add semantic and pragmatic information in face-to-face communication (Holler and Levinson, 2019). Visible gestures are more powerful form of communication than vocalization in dialogue object description tasks (Macuch Silva et al., 2020). In these studies, gestures from human subjects are usually encoded by the hands' spacial loca-

tions, which provide insights to the gesture extraction method used in this study. Also, their results strongly indicate the potentials of building more comprehensive computational language models by including simple non-verbal features. However, so far, few computational studies have attempted to directly model spontaneous language.

2.3 Information theories

Information theory (Shannon, 1948) has been broadly applied in computational linguistics as the theoretic background for the probabilistic models of language. This also provides philosophical explanations to a broad spectrum of linguistic phenomena. One example that interests researchers the most is the assumption/principle of *entropy rate constancy* (ERC). Under this assumption, human communication in any form (written, spoken, etc.) should optimize the rate of information transmission rate by keeping the overall entropy rate constant.

In natural language, *entropy* refers to the predictability of words (tokens, syllables) estimated with probabilistic language models. Genzel and Charniak (2002, 2003) first formulated a method to examine ERC for written language, by decomposing the entropy term into *local* and *global* entropy:

$$H(s|context) = H(s|L) - I(s, C|L) \quad (1)$$

in which s can be any symbol whose probability can be estimated, such as a word, punctuation, or sentence. C and L refer to the global and local contexts for s, among which C is purely conceptual and only L can be operationally defined. By ERC, the left term in eq. (1) should remain an invariant against the position of s. It results in an expectation that the first term on the right $H(s|L)$ should *increase* with the position of s, because the second term $I(s, C|L)$, i.e., the mutual information between s and itself global context should always decrease (see Genzel and Charniak (2003)'s paper for more examples). While they have confirmed the increase of local entropy in written language, Xu and Reitter (2016, 2018) also confirmed the pattern in spoken language, relating it to the success of task-oriented dialogues (Xu and Reitter, 2017).

Now, the goal of this study is to extend the application scope of ERC to the non-verbal realm. More specifically, if the s in eq. (1) represents any symbol that carries information, for example, a gesture or pose, then the same *increase* pattern should be observed within a sequence of gestures. ERC can be interpreted as a "rational" strategy for the information sender (speaker) because it requires less predictable content (higher local entropy) to occur at a later position within the message, which maximizes the likelihood for the receiver (listener) to successfully decode information with the least effort. The question explored here is whether we "speak" rationally by gestures.

3 Questions and Hypotheses

In this study, we focus on two specific hypotheses:
Hypothesis 1: Incorporating non-verbal representations as input will enhance the performance of language modeling task.

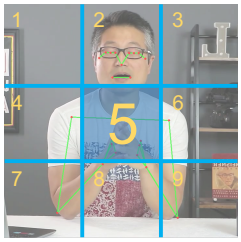
To test Hypothesis 1, we carry out experiments with data-wise and model-wise manipulations. In the former manipulation, non-verbal tokens are inserted into word sequences and form a hybrid type input data for the language model. As for the latter manipulation, the language model is modified to take in non-verbal and verbal input sequences simultaneously and compute a fused internal representation. In both conditions, we expect the inclusion of non-verbal information will increase the performance of language models measured by perplexity.
Hypothesis 2: Non-verbal communication conforms to the principle of Entropy Rate Constancy. To test Hypothesis 2, we will approximate the local entropy ($H(s|L)$) of non-verbal sequences using the perplexity scores from language models, and correlate it with the utterances' relative positions within the monologue data. We expect to discover an increasing pattern of local entropy that is similar to verbal language.

4 Methods

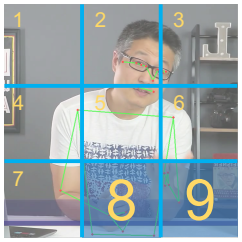
4.1 Data collection and pre-processing

The video data that we use is collected from several YouTube channels. They are manually selected based on the standards that each video must contain only one speaker who faces in front of the camera, and whose hands must be visible. 12 videos from 5 hosts are collected, and the mean duration is 15.0 minutes ($SD = 7.0$).

The pre-processing step is to extract the full-body landmark points of the speaker, in preparation for the next gesture representation step. For this task, we use BlazePose (Bazarevsky et al., 2020), which is a lightweight convolutional neural



(a) Both hands in region 5 → label <5+5>.



(b) Right hand in region 9, left in 8 → label <9+8>.

Figure 1: Create discrete gesture labels based on landmark positions of both hands.

network-based pose estimation model provided in `MediaPipe`[1]. It outputs 33 pose landmarks of the human body detected in each frame.

4.2 Discretization of gestures

The next step is to represent gestures so that they can be embedded into language data. There are various ways of creating *continuous* representations for gestures/poses, such as the pose embedding technique (Mori et al., 2015). However, it is difficult to obtain a set of gestures that are *universal* across speakers using such continuous representations. Thus, for the exploratory purpose of this study, we use a simpler way to represent gestures with *discrete* labels, using the relative positions of hand landmarks.

On each frame, we first split the area containing the body into 9 rectangular regions of the same size, indicated by integer numbers from 1 to 9. Each hand is assigned an integer based on which region it falls into. Then, we use the combination of both hands to create a unique gesture label for that frame. For example, as shown in fig. 1b, the speaker's left and right hands fall into region 9 and 8, so the gesture label is <9+8>. Because there are 9 possible positions for each hand, the total number of gesture labels is $9 \times 9 = 81$. For convenience, we use one integer ID (instead of the merged ID connected by a hyphen) to denote each of these 81 gestures: <1>, <2>, ..., <81>. Note that 81 is the maximum number, and the actual count of unique gesture labels depends on the data.

4.3 Multi-modal language models

We designed two types of LSTM-based language models tailored for the multi-modal training task. All LSTM models are bi-directional with hidden layers of 200, and trained with batch size of 20.

Baseline LSTM model

We implemented an LSTM-based language model to serve as the uni-modal baseline. This model is trained against three types of data: pure word sequence (S_w), pure gesture sequence (S_g), and the mixed sequence (S_{mix}), with a minimum amount of gesture information injected. S_g is generated with the following procedure: First, the audio track of each video is processed with a speech-to-text API[2] that returns a sequence of words, with the *start* and *end* timestamps of each word also annotated. Next, the gesture sequence is generated by sampling one static frame that lies between the duration $[start, end]$ for each word, and then applying the gesture extraction in section 4.2.

The "minimum injection" mentioned above means that S_{mix} is created by appending *one* majority gesture label from S_g to the beginning of S_w. For example, as shown in fig. 2, for a sentence S_w = {*There*, *is*, *one*, *thing* ...}, its corresponding gesture sequence S_g = {<43>,43, ...}, and the majority label is <70>. Thus, the mixed sequence is {<70>, *There*, *is*, ...}. This way of creating S_{mix} is inspired by classical imaging captioning tasks, in which input image is used as the first time step for sentence generation.

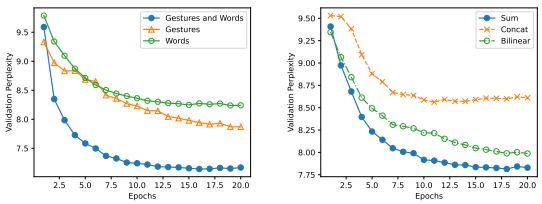
To verify Hypothesis 1, the perplexity scores of S_w and S_{mix} will be compared. The perplexity scores of S_g will be used to verify Hypothesis 2.

Mixed-modal LSTM model

We implemented a mixed-modal LSTM-based language model, which includes gesture inputs of finer granularity. A pair of sequences, S_w (words) and S_g (gestures) are the input, which is then fed into a modality fusion module, where the embedding representation for words and gestures at each time step, i.e., w_i and g_i, are fused by *sum*, *concat*, or a *bilinear* modality fusion component. Finally, the resulting mixed embeddings are inputted into the LSTM encoder to be trained for the next-word prediction task. The model's architecture is shown in fig. 2. Detailed hyper-parameters will be presented in the Appendix. The purpose of this model is to further verify Hypothesis 1, and to explore the optimal modality fusion method.

[1] https://google.github.io/mediapipe/

[2] https://github.com/googleapis/python-speech



(a) Baseline LSTM (b) Mixed-modal LSTM

Figure 3: Validation perplexity vs. training epochs for the models.

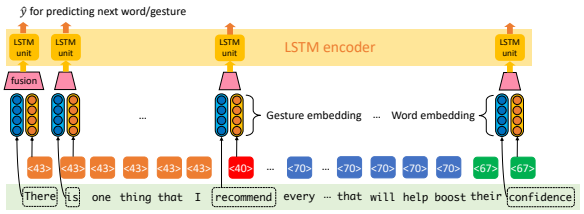


Figure 2: Architecture of the mixed-modal LSTM language model.

5 Results

5.1 Examining Hypothesis 1: Comparing model performance in perplexity

The plots of validation perplexity scores against training epochs are shown in fig. 3. For the baseline LSTM plot (fig. 3a), it can be clearly seen that the mixed input sequence (S_mix) has lower perplexity than the word (S_w) or gesture (S_g) sequences, which supports Hypothesis 1: Gestures do contain useful information that can improve the language model's performance. The mixed-modal LSTM plot (fig. 3b) shows that among all three modality fusion methods, *sum* yields the best performance.

When comparing the perplexity scores of the baseline (base-) and mixed-modal (mm-) LSTM models, we have two major findings: First, mm-LSTM has lower perplexity than base-LSTM with words input (significant by t-test, $t = 29.9$, $p < 0.01$), which is expected because the former has richer inputs (S_g and S_w) than the latter (S_w only). Second, however, mm-LSTM has higher perplexity than base-LSTM with mixed input ($t = -95.8$, $p < 0.01$). This is somewhat counter-intuitive because the mixture of S_g and S_w should encode more information than S_mix, which merely contains one gesture token at the sequence head. We conjecture that this may be due to the lack of data, which needs be re-verified with finer hyper-parameter tuning in future work.

5.2 Examining Hypothesis 2: Local entropy and utterance position relationship

We use linear models to examine the correlations between the local entropy of sequences (S_g or S_w) and the relative position of utterances. For gesture sequences, utterance position is a significant predictor of local entropy with positive coefficient ($F(1, 74) = 4.481$, adjusted $R^2 = .044$, $p < .05$), which supports the Hypothesis 2. A visible increasing trend of local entropy is shown in fig. 4. Surprisingly, word sequences yield no significant models, which contradicts with previous findings. However, this is likely due to the small data size used and the inaccurate sentence tokenization results, which could also be because of the randomness in the data.

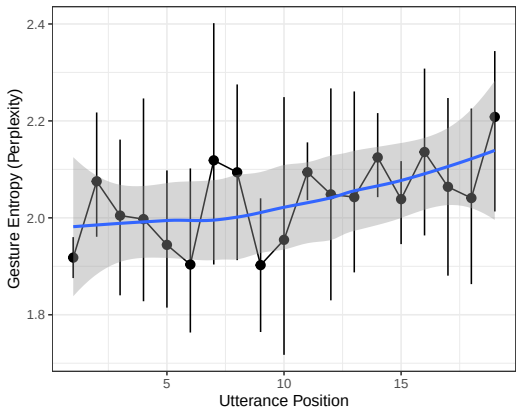


Figure 4: Local entropy of gesture sequences increases with utterance position.

6 Conclusions

Based on our results, we conclude that gestures carry information that can **enhance** verbal communication. More importantly, speakers use gestures in a **rational** way that conforms with the principle of Entropy Rate Constancy in Information Theory. This work is exploratory but the evidence is promising, as only a small data-set is used and a simplistic gesture representation method is applied.

For future work, we plan to work with a larger and more diverse dataset with a higher variety in genres (public speech, etc.) and examine more advanced representation methods, such as continuous embedding and clustering. Another direction to pursue is to interpret the semantic meanings of gestures and other non-verbal features by examining their semantic distance from words/utterances in vector space.

References

Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*.

Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. 2018. Inferring shared attention in social scene videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6460–6468.

Lifeng Fan, Shuwen Qiu, Zilong Zheng, Tao Gao, Song-Chun Zhu, and Yixin Zhu. 2021. Learning triadic belief dynamics in nonverbal communication from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7312–7321.

Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. 2018. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5958–5966.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 199–206, Philadelphia, PA.

Dmitriy Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 65–72, Sapporo, Japan.

Judith Holler and Stephen C Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.

Vinicius Macuch Silva, Judith Holler, Asli Ozyurek, and Seán G Roberts. 2020. Multimodality and the origin of a novel communication system in face-to-face interaction. *Royal Society open science*, 7(1):182056.

Greg Mori, Caroline Pantofaru, Nisarg Kothari, Thomas Leung, George Toderici, Alexander Toshev, and Weilong Yang. 2015. Pose embeddings: A deep architecture for learning to match human poses. *arXiv preprint arXiv:1507.00302*.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Fei Xu, Kenny Davila, Srirangaraj Setlur, and Venu Govindaraju. 2019. Content extraction from lecture video via speaker action classification based on pose information. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1047–1054. IEEE.

Yang Xu and David Reitter. 2016. Entropy converges between dialogue participants: explanations from an information-theoretic perspective. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 537–546, Berlin, Germany.

Yang Xu and David Reitter. 2017. Spectral analysis of information density in dialogue predicts collaborative task performance. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 623–633, Vancouver, Canada. Association for Computational Linguistics.

Yang Xu and David Reitter. 2018. Information density converges in dialogue: Towards an information-theoretic model. *Cognition*, 170:147–163.