Gestures Are Used Rationally: Information Theoretic Evidence from Neural Sequential Models

Yang Xu
Department of Computer Science
San Diego State University
yang.xu@sdsu.edu

Yang Cheng
University of Southern California
ycheng04@usc.edu

Riya Bhatia
Carnegie Mellon University
riyabhat@andrew.cmu.edu

Abstract

Verbal communication is companied by rich non-verbal signals. The usage of gestures, poses, and facial expressions facilitates the information transmission in verbal channel. However, few computational studies have explored the non-verbal channels with finer theoretical lens. We extract gesture representations from monologue video data and train neural sequential models, in order to study the degree to which non-verbal signals can effectively transmit information. We focus on examining whether the gestures demonstrate the similar pattern of entropy rate constancy (ERC) found in words, as predicted by Information Theory. Positive results are shown to support the assumption, which leads to the conclusion that speakers indeed use simple gestures to convey information that enhances verbal communication, and the production of non-verbal information is rationally organized.

1 Introduction

Communication is a multi-modal process, in which information from verbal and non-verbal modalities are mixed into one channel. It has been revealed from a long history of 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 (Pease and Pease, 2008; Krauss et al., 1996). 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: 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 take a simple yet necessary first step approaching the above questions, by examining a basic Information Theoretic property of gestures that comes along with verbal language. Some preliminary but prospective results are presented.

2 Related Work

2.1 Gestures as non-verbal communication

There is vast literature on the connection between gesture and language in human communication. Gestures, defined as "the spontaneous hand movements produced in rhythm with speech" (Clough and Duff, 2020) naturally co-occur with spoken language. According to the thorough survey from (Clough and Duff, 2020), the *communication* function of gestures is one of the main focus of early studies. McNeill (1992) has classified gestures into two categories, representative and non-representative, in which the former has clearer semantic meanings (e.g., depicting objects and describing locations), while the latter refers to the brief, repetitive movements that has little substantive meanings.

2.2 Non-verbal communication in natural language processing

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.,

Proceedings of the 29th International Conference on Computational Linguistics, pages 134–140
October 12–17, 2022.

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.

2.3 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, communication in any form (written or spoken) 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). Xu and Reitter (2016, 2017, 2018) has confirmed the pattern in spoken language.

Now, the goal of this study is to extend the application scope of ERC to the non-verbal realm. If the s in eq. (1) represents any symbol that carries information, for example, a gesture, 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 here is to examine whether we "speak" rationally by gestures.

3 Question and Hypothesis

Our hypothesis is: non-verbal communication also conforms to the principle of ERC. To test it, we approximate the local entropy ($H(s|L)$) of non-verbal "tokens" using the perplexity scores obtained from neural sequential models, and correlate it with the utterances' relative positions within the monologue data. If we can find that $H(s|L)$ increases with utterance position, is similar to verbal language, then it supports the hypothesis.

4 Methods

4.1 Data collection and pre-processing

The video data that we use is collected from several YouTube channels. All the videos are carefully 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 network-based pose estimation model provided in `MediaPipe`[1]. It outputs 33 pose landmarks of the human body detected in each frame.

4.2 Extract gesture labels

The next step is to represent gestures so that they can be encoded by the neural sequential model. 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 purpose of this study, we extract *discrete* gesture labels, by categorizing the hands positions into grids. We divide the front space of speaker into 3×3 regions, i.e., indicated by integer numbers from 1 to 9. Each *hand* is assigned a number based on which region it falls into. Next, we use the combination of both hands

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

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, which determines its gesture label as <72>. 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>. The total number of gesture labels is $9 \times 9 = 81$. Note that 81 is the theoretical maximum number, and the actual count depends on the size of data.

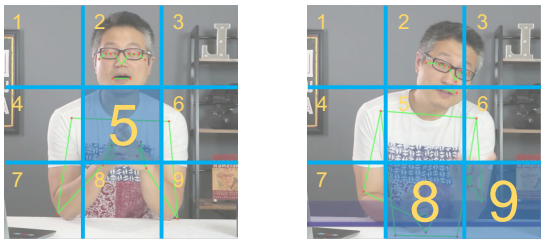
4.3 Prepare gesture sequences

After obtaining the discrete gesture labels for all video frames, we prepare the gesture sequences based on the time stamped text transcript for each video. We use the automatically generated text transcript in `.vtt` format, which contains the *start* and *end* time stamps for each word token in the subtitle. See the following example:

```
<00:00:00.510><c> let's</c>
<00:00:00.780><c> talk</c>
<00:00:01.020><c> about</c>
```

in which each the start time stamp is appended to the head, and the start time for a token is the end time for the previous token. In this example, the token *talk* elapses from 0.780 to 1.020 in seconds. Multiplying the time stamps with frame rate (24 FPS) returns that the word elapses from the 19th frame to the 24th. Then, for each frame within the range of $[19, 24]$, we extract a gesture label using the method described in Section 4.2, resulting in a sequence of gesture labels, $[g_{19}, g_{20}, \ldots, g_{24}]$. This sequence represents a continuous change of gestures during the articulation of the corresponding word, which in most cases, consists of identical gesture labels. Therefore, we select the median label g among $[g_{19}, \ldots, g_{24}]$ as a compact representation.

For a sentence of N wordss, we obtain the median gesture label for each token, $\{g_1, g_2, \ldots, g_N\}$. Despite the down sampling effect of using the median label, there is still large amount of repetition in the resulted sequence. For example, in the first row of table 1, the median gesture label is the same <24> for the first 6 tokens, which means that the speaker did not move his/her hands during that period of time. It makes sense that we treat these repeated gesture labels just as one label. By merging the 6 repeats of <24> and 2 repeats of <36>, we get a compressed gesture sequence, {<36>, <24>}, which means the speaker has made two distinct gestures during the utterance. For each median gesture sequence of length N, we obtain its compressed version $\{\hat{g}_1, \hat{g}_2, \ldots, \hat{g}_{N'}\}$, where $N' \leq N$. See table 1 for examples.



(a) Both hands in region 5 → label <25>.

(b) Right hand in region 9, left in 8 → label <72>.

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

4.4 Sequential models for gesture input

We implement two neural network-based models for the sequential modeling tasks, using LSTM (Hochreiter and Schmidhuber, 1997) and Transformer (Vaswani et al., 2017) encoders. The model takes as input a sequence of gesture labels (median g or compressed $\hat{g}$) and convert them to the embedding space. Then the gesture embeddings are fed to the LSTM/Transformer encoders to capture the temporal dependency between gestures, which compute a dense representation for gestures at each time step. Lastly, the dense representation at the previous time step is used to predict the gesture label at the next time step using a softmax output. The model architecture is shown in fig. 2.

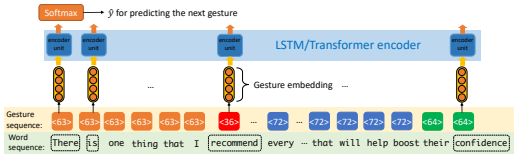


Figure 2: Architecture of the sequential model for encoding gesture input and next time-step prediction.

The learning task is to predict the next gesture label, i.e., minimizing the negative log probability:

$$NLL = -\sum_{t=1}^{T} \log P(g_t | g_1, g_2, \ldots, g_{t-1}) \quad (2)$$

in which $g_1, \ldots, g_{t-1}$ is all the gesture tokens before g_t within the same utterance. An exponential conversion of eq. (2) leads to the local entropy term,

Word tokens in utterance	Median gesture g of each token	Compressed gesture sequence $\hat{g}$
going to give you a flatter look glossy	<24> <24> <24> <24> <24> <24> <36> <36> ($N = 8$)	<24> <36> ($N' = 2$)
now this is really your preference	<40> <72> <64> <64> <40> <40> ($N = 6$)	<40> <72> <64> <40> ($N' = 4$)
I think most of us can get on board	<63> <63> <63> <63> <63> <63> <63> <63> <63> ($N = 9$)	<63> ($N' = 1$)

Table 1: Examples of gesture sequences. Integers wrapped by "<>" are gesture labels. For each sequence, its compressed version is shorter in length: $N' < N$

.

$H(g|L) = \exp(NLL)$, which is the target variable of our interest. This learning task is no different from conventional language modeling tasks, except that the input here is non-verbal tokens. Detailed model hyper-parameters and training procedures are included in appendix A.1.

5 Results

5.1 Summary of data

53 videos are collected from 4 YouTube channels (i.e., 4 distinct speakers). The average length of videos is 723.7 seconds (*SD* = 438.1). There are 17.9K lines of automatically generated subtitles consisting of 121.5K word tokens in total. 81 distinct gesture labels are extracted. The total count of the median gesture label is the same as that of the word tokens (121.5K). The compressed gesture labels has a much smaller total count 26120.

The top 5 most frequent gesture labels are <63>, <56>, <64>, <72> and <36>. The frequency distribution of gesture labels roughly follows the Zipf's law, which is a common distribution pattern in natural language data (Zipf, 2013; Piantadosi, 2014) (See fig. 3). Gesture label <63> is the dominant gesture throughout the data. It is gestural position where the speaker's right hand (from his/her perspective) is in region 7, and left hand region 9.

5.2 Examine hypothesis: local entropy increases with utterance position

The local entropy of each gesture sequence (median and compressed, respectively) is plotted against the corresponding utterance's position in fig. 4, which shows a visible increasing trend.

We use linear models to verify the correlations between local entropy and utterance position. It is confirmed that utterance position is a significant predictor of local entropy with positive β coefficients. For raw gestures, the *beta*s are smaller: $\beta_{\text{LSTM}} = 1.6 \times 10^{-3}$ ($p < .05$), $\beta_{\text{Trm}} = 2.3 \times 10^{-3}$ ($p < .01$); for compressed gestures: $\beta_{\text{LSTM}} = 0.097$, $\beta_{\text{Trm}} = 0.093$ ($p < .001$). Therefore, the increase of local entropy is statistically significant. It supports our hypothesis.

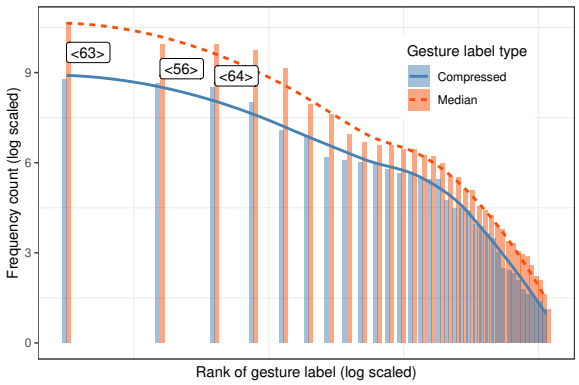


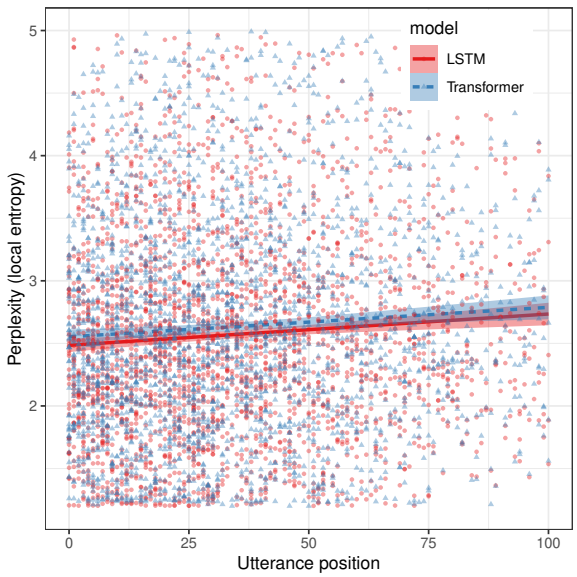
Figure 3: Frequency count against the rank gesture labels in logarithm transformed scales. Top three most frequent gesture labels annotated.

5.3 Analysis of typical gesture

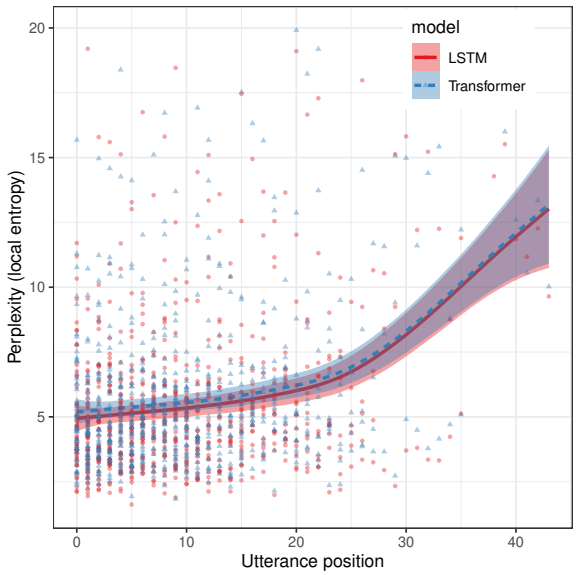
We select three highly frequent gesture labels <63>, <56> and <72>, and show some typical screenshots in fig. 5. In these gestures, the positions of both hands are at the mid-lower position in front of the body. Gesture <63> has two hands evenly distant from the center, while gesture <56> captures a movement to the right and gesture <72> to the left. In general, these are very commonly seen patterns in daily communication.

6 Discussion and Conclusions

Our results confirms that the way gestures are used as a complementary non-verbal communication side-channel follows the principle of entropy rate constancy (ERC) in Information Theory. It means that the information encoded in hand gestures, albeit subtle, is actually organized in a *rational* way that enhances the decoding/understanding of information from a receiver's perspective. The main contribution is that we extend the scope of ERC to realm of non-verbal communication.



(a) Raw gesture

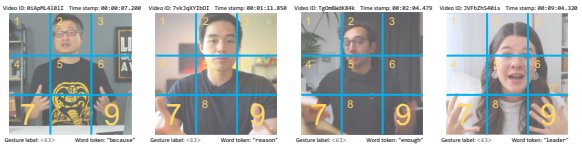


(b) Compressed gesture

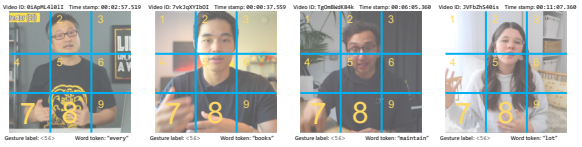
Figure 4: Local entropy of gesture sequences increases with utterance position. 95% CIs are shown.

There are two explanations for what causes the observed entropy increasing pattern: First, more rare gestures (higher entropy) near the later stage of communication; Second, the entropy for the same gesture also increases during the communication. While the latter indicates a more sophisticated and interesting theory about gesture usage, both explanations requires further investigation.

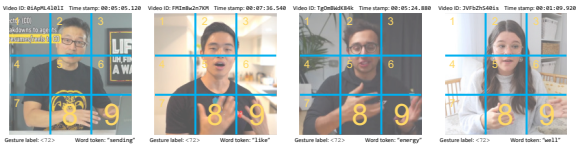
While the motivation of this study is theoretical, but we believe the idea of extracting discrete gesture labels from spontaneous monologue/dialogue also has potentials in application. For instance, into better analysis of speaker intensions, sentiments,



(a) Gesture label <63>



(b) Gesture label <56>



(c) Gesture label <72>

Figure 5: Typical screenshots for the top frequent gesture labels <63>, <56> and <72>.

and other implicit messages. For future work, we plan to use a larger dataset with a higher variety in genres (public speech, etc.) and examine more advanced representation method. such as continuous embedding and clustering. It is also interesting to interpret the semantic meanings of gestures and other non-verbal features by examining their semantic distance from words/utterances in vector space. More specifically, non-parametric clustering algorithms can be used to identify distinct the actions or poses of a person, which provides a way to extract more general gesture/action/pose labels for training.

Acknowledgements

This work is supported by National Science Foundation (CRII-HCC: 2105192). We thank Chen Song for providing valuable ideas on modeling.

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*.

Sharice Clough and Melissa C Duff. 2020. The role of gesture in communication and cognition: Implications for understanding and treating neurogenic communication disorders. *Frontiers in Human Neuroscience*, page 323.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Robert M Krauss, Yihsiu Chen, and Purnima Chawla. 1996. Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In *Advances in experimental social psychology*, volume 28, pages 389–450. Elsevier.

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*.

David McNeill. 1992. Hand and mind1. *Advances in Visual Semiotics*, page 351.

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*.

Barbara Pease and Allan Pease. 2008. *The definitive book of body language: The hidden meaning behind people's gestures and expressions*. Bantam.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin and Review*, 21(5):1112–1130.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

George Kingsley Zipf. 2013. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Routledge.

A Appendix

A.1 Hyper-parameters and training procedures

For the LSTM-based encoder, embedding size is 300, hidden size is 200, number of layers is 2; a fully connected layer is used as the decoder connecting the encoder output and the softmax; dropout layers of probability 0.2 are applied to the outputs of both the encoder and decoder. For the Transformer-based encoder, model size is 20, hidden size is 100, number of layers is 2; same fully connected linear decoder is used; dropout layers of probability 0.5 are used at the position encoding, and each transformer encoder layer. To enable the one-direction (left to right) modeling effect, a mask matrix (of 0 and 1s) in an upper-triangular shape is used together with each input sequence.

Model parameters are randomly initialized. Training is done within 40 epochs, with batch size of 20, at and initial learning rate $lr = 0.05$. SGD optimizer with default momentum is used for training the LSTM model; Adam optimizer is used for training the Transformer model. Data are split to 80% for training and 20% for testing. After each training epoch, evaluation is done over the test set, and the model with lowest perplexity scores is saved as the best one.

Models are implemented with PyTorch. `torch.nn.CrossEntropyLoss` module is used as the loss function. The mathematical meaning of the output from this function is the negative logarithm likelihood (*NLL* in eq. (2)), and thus we compute the exponential values of the output to get the local entropy scores. The entropy scores used in the plot and statistical analysis are obtained from both train and test sets.