Logographic Information Aids Learning Better Representations for Natural Language Inference

Zijian Jin
Tandon School of Engineering
New York University
zj2076@nyu.edu

Duygu Ataman
Courant Institute of Mathematical Sciences
New York University
ataman@nyu.edu

Abstract

Statistical language models conventionally implement representation learning based on the contextual distribution of words or other formal units, whereas any information related to the logographic features of written text are often ignored, assuming they should be retrieved relying on the cooccurence statistics. On the other hand, as language models become larger and require more data to learn reliable representations, such assumptions may start to fall back, especially under conditions of data sparsity. Many languages, including Chinese and Vietnamese, use logographic writing systems where surface forms are represented as a visual organization of smaller graphemic units, which often contain many semantic cues. In this paper, we present a novel study which explores the benefits of providing language models with logographic information in learning better semantic representations. We test our hypothesis in the natural language inference (NLI) task by evaluating the benefit of computing multi-modal representations that combine contextual information with glyph information. Our evaluation results in six languages with different typology and writing systems suggest significant benefits of using multi-modal embeddings in languages with logograhic systems, especially for words with less occurence statistics.

1 Introduction

The essential idea in statistical language modeling is to represent the meaning of a word as a function of its context. The function, modeled via the conditional probability of observing a word in a given utterance, has most efficiently been approximated with a neural network based architecture (Mikolov et al., 2013a,b; Bengio et al., 2003; Mikolov et al., 2010; Sundermeyer et al., 2012). The outstanding performance of neural methods in language modeling and their recent development (Peters et al., 2018; Tenney et al., 2019) have them a preliminary component in various downstream NLP tasks.

One of the main limitations in the formulation of language models lies however in the choice of ortographic units in calculating the contextual distribution, which is usually convenient in English and other languages using phonetic scripts. On the other hand, many languages rely on logographic writing systems, where surface forms are represented as a visual organization of smaller graphemic units and the word meaning can be changed through compositional variations of these units. Although a direct segmentation of these units has been found quite challenging due to visual compositions in the final form of the grapheme, previous studies have found potential benefits of using visual information to aid NLP models in sentence representation (Liu et al., 2017a; Meng et al., 2019; Dai and Cai, 2017; Salesky et al., 2021). On the other hand, none of these studies have focused on isolating the effects of different linguistic features in relation to their correlation to visual features.

As shown in Figure 1, logographic information often contain important features related to the word meaning. In this paper, we perform the first focused analysis to measure the significance of logographic features specifically to the semantic information encoded in token or character-level language representations by evaluating the performance of multi-modal embeddings in the NLI task. In particular, we aim to answer the following research questions:

1) How important may logographic information be to for an accurate representation of semantic information in word or character-level language units

2) Whether the contribution of logographic information to semantic representations may depend on the language typology and writing system

In order to answer these questions we implement a multi-modal representation learning model where

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 268–273
November 20–23, 2022.

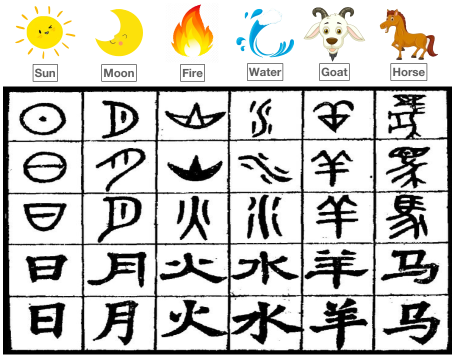

Figure 1: Logographic information in Chinese.

Layer	Visual encoding model
1	`Spatial Conv.` (3,3) $\rightarrow$ 32
2	`ReLu`
3	`MaxPool` (2,2)
4	`Spatial Conv.` (3,3) $\rightarrow$ 32
5	`ReLu`
6	`MaxPool` (2,2)
7	`Spatial Conv.` (3,3) $\rightarrow$ 32
8	`ReLu`
9	`Linear` (800,128)
10	`ReLu`
11	`Linear` (128,128)
12	`ReLu`

each written text segment is representation as a combination of visual embeddings obtained from prominent convolutional neural network (CNN) based models (Liu et al., 2017a; Meng et al., 2019; Salesky et al., 2021), and contextual representations obtained from multilingual pre-trained language models (Devlin et al., 2019; Conneau et al., 2020a). We evaluate the contribution of visual information to the performance in the NLI task under few-shot learning settings in six languages with varying typology and writing systems: English, Spanish, Hindi, Urdu, Vietnamese and Chinese. We also study the optimal representation granularity for semantic information by comparing word or character-level multi-modal representations in our experiments.

In conclusion, we find that taking into account the visual information improves the performance in NLI tasks especially in logographic languages like Chinese and that the improvements are correlated with the factors that determine the quality of token representations, such as the occurence of the tokens in training data as well as language model capacity and hyperparameters. Our findings suggest multi-modal processing is a promising direction, especially for processing languages where conditions of data sparsity may create fall backs in assumptions undertaken in statistical formulations.

2 Computing Visual Glyph Embeddings

Our multi-modal embedding model is composed of two components: *(i) the visual encoder*, which computes embeddings based on the input images representing each text segment, and *(ii) the pre-trained language model* providing the text-based embeddings.

Image conversion Text segments consisting of complete sentences are split into words (or characters) and then converted into images. Sentences are split into 30 x 60 pixel word images using the Jieba[1] tool. All graphemes are centralized to the middle of the image.

Visual Embeddings In order to extract the glyph information from text images, we use the CNN model developed by (Liu et al., 2017b; Sutskever et al., 2014) to generate visual embeddings. The model consists of a three-layer CNN, augmented with a two-layer feed-forward network. The full details of the network is given below.

The visual features extracted by the CNN are further encoded in a long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) network to learn the glyph embeddings.

For a sequence consisting of t tokens $x_0, x_1, ..., x_t$, the *visual embedding* v is computed by concatenating (Su et al., 2020; Lu et al., 2019) the hidden states of the LSTM and averaging them as

$$v = \text{mean}([h_0; h_1; ...; h_t])$$

Embedding composition In order to isolate the learning of representations from two modalities and measure their effect on the learning task in a controlled setting, we deploy late fusion in combining the visual embeddings with the text embeddings obtained by the pre-trained model for prediction in the down-stream task. The two embeddings are linearly composed through a simple affine projection and then concatenated. For the down-stream prediction task we use a multi-layer perceptron classifier.

[1] https://github.com/fxsjy/jieba

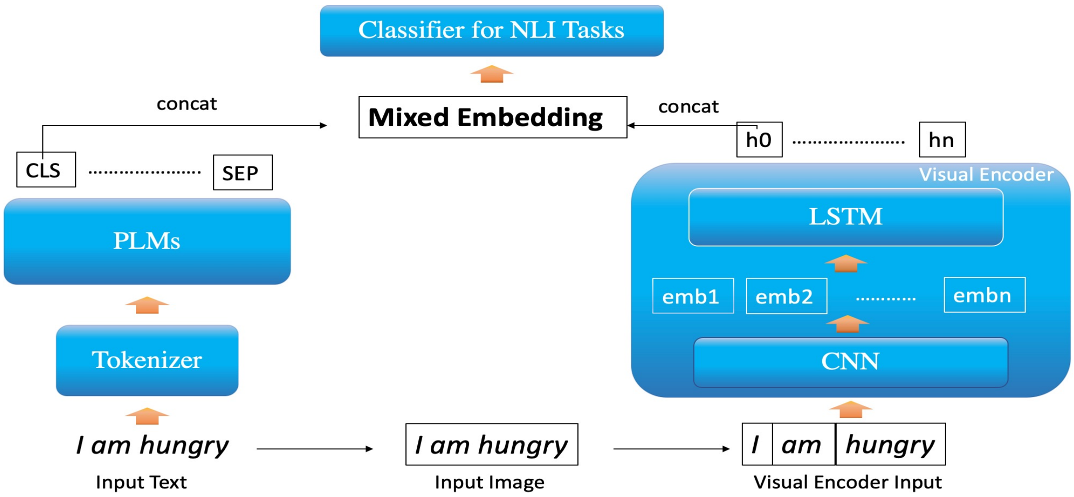

Figure 2: Method overview.

3 Experiments

3.1 Character recognition

As an initial verification, we implement the visual encoder and evaluate it individually in the character recognition task. We use the CASIA Chinese Handwriting Database (Liu et al., 2011) and obtain competitive results (93.23% accuracy) on this task, confirming the visual encoder works sufficiently in extracting character features from input images.

3.2 NLI

Data We evaluate our model under few-shot learning settings using the XNLI dataset (Conneau et al., 2020b), using only a small portion of the testing data for training and development, and test the effect of logographic information to contribute to resolve the high level of semantic ambiguity.

Datasets	Number of Sentences
Training	4509
Development	501
Test	2490

Table 1: Data statistics for training, development and test sets.

Model settings and hyper-parameters In training the multi-modal models, the learning rates of both XLM-R and mBERT based pre-trained models are set to 1e-6. The visual encoder is trained on the images captured from the training sentences, either at word or character-level resolution, with a learning rate of 4e-6 (for XLM-Roberta) and 1e-6 (for mBERT). The hidden size of the LSTMs used is 128 and we use dropout of with 0.3 in this layer. All hyper-parameters are tuned with grid-search. For each task we train 30 epochs and always choose the results with smallest validation loss.

Languages We pick six languages with varying typology and writing systems, including English, Spanish, Urdu, Vietnamese, Chinese and Hindi. **English** and **Spanish** use the Latin script; **Urdu** is written with the Arabic alphabet, whereas **Hindi** uses Devanagari, all of which are phonetic writing systems. **Chinese** uses logographic writing. **Vietnamese**, although traditionally have used logographic writing, recently and in the XNLI data set is written with the Latin script.

Contextual representations We verify the significance of logographic information for contributing to enrich the language representations by testing our multi-modal approach with two different pre-trained language models, including the `mBert-base` and the `XLM-R-base` both available from Huggingface[2]. We also investigate the effects of different segmentation methods for processing sentence images either at the level of words or characters.

4 Results and Discussion

Our experiment results are given in Table 2. At a first glance, we observe the performance of the models are much lower than reported in (Conneau et al., 2020b), since we have significantly less training and development data available. Under these challenging evaluation settings with high amount of sparsity, we observe that the logographic information improves the performance obtained using the `mBert-base` model in all languages that do not deploy the Latin script, including Chinese, Urdu, Hindi and Vietnamese. In case

[2]https://huggingface.co

Table 2: Results in the XNLI benchmark. `base` models represent baseline pre-trained language model performance in the down-stream task. `base-CNN` models represent the multi-modal system performance. `(C)` denotes character and `(W)` denotes word level input representations. `Random` stands for comparisons to multi-modal systems where random images were input to the visual encoder to verify the effect of visual information on the overall performance.

Languages	English	Chinese	Urdu	Hindi	Vietnamese	Spanish
mBERT-base	**65.86**	55.28	51.29	56.48	57.08	**62.27**
mBERT-base-CNN (W)	62.87	58.88	53.49	**57.68**	59.88	60.47
mBERT-base-CNN (C)	64.07	**59.08**	**53.69**	57.48	**60.07**	61.67
mBERT-base-CNN (C) — Random	-	54.33	-	-	-	-
XLM-Roberta-base	**69.86**	64.27	**59.88**	**63.87**	63.07	**65.66**
XLM-Roberta-base-CNN (W)	69.26	**66.66**	57.88	62.87	61.67	65.46
XLM-Roberta-base-CNN (C)	68.26	62.07	56.28	61.67	**63.07**	65.26
XLM-Roberta-base-CNN (C) — Random	-	61.36	-	-	-	-

of `XLM-Roberta-base`, which had better optimization on a larger corpus, the overall performance are consistently better than `mBert-base` and the improvements remain consistent, especially in Chinese and Vietnamese. We hypothesize that the slightly higher amount of improvements in `mBERT-base` might be due to better quality of representations provided with the optimized training regime of `XLM-Roberta-base`. Using the `mBERT-base` model, we find more advantage of embedding logographic information at the character level in Chinese, Urdu and Hindi, however, in Hindi, the results are comparable. When using the `XLM-Roberta-base`, we observe improvements in Chinese with word-level glyph embeddings and in Vietnamese using character-level glyph embeddings. While character-level embeddings might be suitable for a phonetic language like Vietnamese, the logograhic writing system in Chinese might make word-level visual embeddings more convenient, since the intra-graphemic dependencies can be captured at the visual level.

Although the improvements highly correlate with the logograhic nature of the writing system, the fact that they apply to most languages, even Urdu and Hindi with phonetic alphabets, point to the suboptimal effects in tokenization or segmentation and their potential harms to correctly model the contextual distribution. We also see in high-resource language representations like English, our fusion method may be harmful to the downstream task, which we anticipate that could be resolved with higher amount of fine-tuning and development data. In light of all these considerations, the findings suggest that multi-modality is a promising direction for overcoming problems related to data sparsity, and eventually tokenization or segmentation-free language modeling.

Models	Accuracy	# of UNK
mBERT-base (C)	45.31	128
mBERT-base-CNN (C)	52.43	

Table 3: Results for targeted evaluation, UNK represents unknown tokens.

As an additional analysis investigating the effects of token frequency on the positive effects of logographic information integrated in the language model, we sample sentences in the test set that have unknown words in the model vocabulary and compute the targeted accuracy on this sample of sentences. The results shows in table 3 further illustrate the boosted performance on the sample test, suggesting that data sparsity is an important obstacle to learning high-quality contextual representations, and such conditions can be the ideal place where logographic information might be useful to improve the semantic features embedded in representations.

5 Conclusion

In this paper, we evaluated the benefits of using logographic information in language modeling by implementing a multi-modal representation learning model which combines contextual language representations with visual embeddings. Our experiments in the NLI task in six languages confirmed the benefits of logograhic information in obtaining more reliable semantic representations, especially under sparse learning settings. As future work we hope to contribute to the development of larger multilingual benchmarks to evaluate the effect of visual information on more languages and linguistic phenomena. Our software and the experimental data will be available upon publication.

Acknowledgements

The authors would like to express deep gratitude to Chinmay Hedge and Aishwarya Kamath for discussions and their useful feedback. This work was supported by Samsung Advanced Institute of Technology (under the project Next Generation Deep Learning: From Pattern Recognition to AI) and NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Holger Schwenk, Ves Stoyanov, Adina Williams, and Samuel R Bowman. 2020b. Xnli: Evaluating cross-lingual sentence representations. In *2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2475–2485. Association for Computational Linguistics.

Falcon Z. Dai and Zheng Cai. 2017. Glyph-aware embedding of chinese characters. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP, Copenhagen, Denmark, September 7, 2017*, pages 64–69. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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

Cheng-Lin Liu, Fei Yin, Da-Han Wang, and Qiu-Feng Wang. 2011. Casia online and offline chinese handwriting databases. In *2011 International Conference on Document Analysis and Recognition*, pages 37–41. IEEE.

Frederick Liu, Han Lu, Chieh Lo, and Graham Neubig. 2017a. Learning character-level compositionality with visual features. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2059–2068.

Frederick Liu, Han Lu, Chieh Lo, and Graham Neubig. 2017b. Learning character-level compositionality with visual features. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 2059–2068. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. *Advances in Neural Information Processing Systems*, 32.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations, Workshop Track Proceedings*.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. volume 2, pages 1045–1048.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26:3111–3119.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Elizabeth Salesky, David Etter, and Matt Post. 2021. Robust open-vocabulary translation from visual text representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7235–7252.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *13th Annual Conference of the International Speech Communication Association*, pages 194–197.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.