Text-Based Prediction of Visual Complexity: How Does What We See Influence What We Say?

Eleanor Lin,[1] Ziyan Yang,[2] Vicente Ordóñez[2]

[1]Department of Computer Science and Program of Linguistics, Columbia University; [2]Vision, Language, and Learning Lab, Department of Computer Science, Rice University

e.lin2@columbia.edu, zy47@rice.edu, vicenteor@rice.edu

Introduction

Visual complexity has applications to art, web design, advertisement, psychology, and computer science (Saraee et al., 2020).

Do humans describe images differently based on their complexity?

Can machines predict visual complexity from verbal image descriptions?

"There are **many people** walking on the **busy** downtown street."

Materials and methods

BERT: a machine learning model "pre-trained" to understand human language (Kenton et al., 2019)

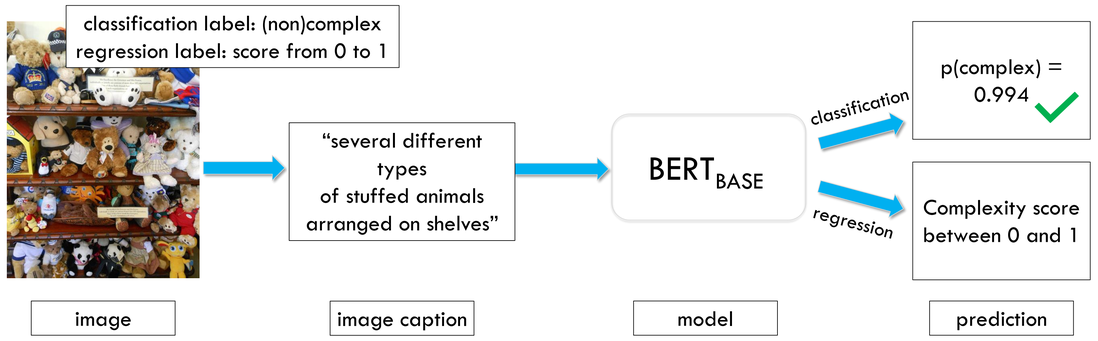
SAVOIAS dataset: images with complexity scores generated by humans (Saraee et al., 2020) → We developed a complexity metric, **distinct number of regions**, using this data

COCO dataset: images with captions written by humans (Lin et al., 2014) → We scored images' complexity and "fine-tuned" BERT using this data

Results

- New visual complexity metric: **distinct number of regions**
- Pearson's correlation of **r = 0.62** ($p < 0.001$) between number of distinct regions and human-generated complexity scores for images from the Objects, Interior design, and Scenes subsets of the SAVOIAS dataset

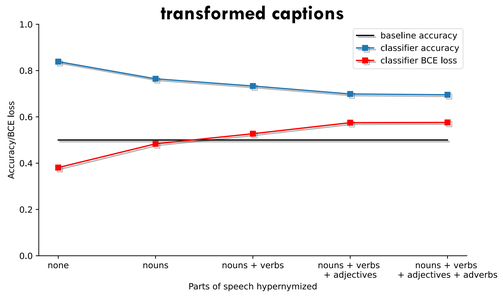
Model setup with example input and output



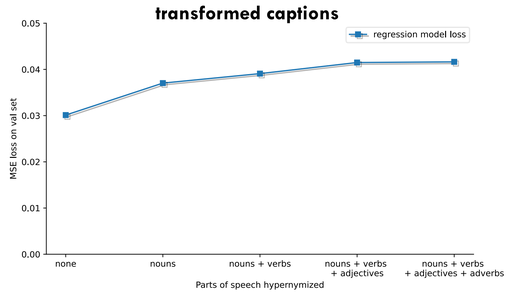
- **Classification model: 83.9% accuracy**
- **Regression model:** Pearson's correlation **r = 0.659** ($p < 0.001$) between predicted and actual complexity scores
- **Problem:** How to prevent the model from learning dataset-specific biases?
  - **Solution:** Caption transformation

Shelves of stuffed animals of various color and shapes. → objects of plain objects of plain object and objects.

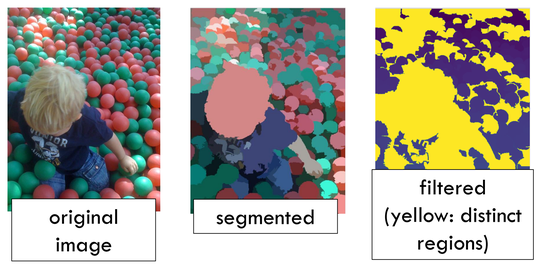
Classification model performance with transformed captions



Regression model performance with transformed captions



Results of complexity scoring algorithm



Conclusions

In this work, we:

- Defined a new visual complexity metric: number of distinct regions
- Identified the complexity of images from the COCO Dataset (Lin et al., 2014) using our metric
- Provided classification and regression models to predict image complexity from captions

Our work suggests that **visual and linguistic complexity are related,** and that we can use this relationship to better identify complex images and improve algorithms for tasks that involve both vision and language, such as automatic caption generation.

References

Lee Kenton, Jacob Devlin, Ming-Wei Chang, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Elham Saraee, Mona Jalal, and Margrit Betke. 2020. Visual complexity analysis using deep intermediate-layer features. *Computer Vision and Image Understanding*, article 102949.

Acknowledgments

This work was supported by the Computing Research Association's Distributed Research Experiences for Undergraduates program, which receives funding from the National Science Foundation, and by the National Science Foundation under Grant No. 2221943. 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 National Science Foundation.

Further information

Please **see** https://emlinking.github.io/ for more on this project, or reach out via email at e.lin2@columbia.edu.

Scan me!