Reconsidering the definition of noise: Multimodal discourse analysis of computer vision datasets

Keywords: Noise, Multimodal discourse analysis, Computer vision, Gender representation, Dataset bias

Extended Abstract

Introduction

Since Shannon and Weaver (1949) defined "noise" as an unintentional interruption in the way of message transmission, the concept "noise" has been treated as something that must be removed for effective communication. More recently, in the process of constructing flawless and organized datasets for machine learning, the term "noise" is employed to refer to irrelevant or misleading data as well as obvious errors, which can lead to a tremendous amount of discarded data (Jeatrakul et al., 2010). To create a dataset containing image-text tuples for face classifications, for example, if there is only female and male dichotomy within the classification schemes used, transgenders' faces are more likely to be mislabeled by annotators and thus removed to avoid inconsistency and misclassification (i.e., class noise). In this process, existing classification systems, which has defined what noise is, are considered self-evident and natural (Star & Bowker, 2007). However, noise is an essentially relational, subjective, and unstable concept (Ballard, 2010; Hainge, 2013), and it was also rightly pointed out in Shannon and Weaver (1949) that the decision as to what constitutes noise or information is necessarily a subjective matter. As such, noise can be important as they expand magnitude and diversity within an alternative classification system (e.g., transgenders' faces within a gender spectrum system) in dataset construction.

Materials and Methods

The aim of this study was to tackle the existing classification systems by analyzing multimodal (i.e., verbal and visual) discourse comparing Google AI's two datasets: *Conceptual Captions 3M* (CC3M) and *Conceptual 12M* (CC12M) focusing on gender binarism. The discarded data was revaluated to examine what data were discarded or ignored within the datasets CC3M (Sharma et al. 2018), which contains 3.3 million images with annotations selected from 5 billion candidates through three steps of thorough filtering for a 'cleaned' dataset. Then, by relaxing the filtering processes – not excluding annotations with minor grammatical errors or images with more varied ratios – CC12M (Changpinyo et al., 2021) was released, which contains 12 million image-text pairs. This CC12M dataset not only contains four times larger image-text pairs than CC3M, but also preserve a higher diversity degree of the concepts captured. Assuming that various potential data conflicting gender binarism was removed more through the filtering for CC3M than CC12M for the sake of noise cleaning, three analyses of comparison were conducted: 1) Types of discarded explanations about images were examined through semantic network analysis of the annotations of the datasets; 2) The images contained in each dataset represent were analyzed through socio-semiotic analysis of the images; 3) Based on the results of 1) and 2), what was represented in the selected annotations and what was ignored were examined.

Results

The results show that the annotations and images of CC3M are more compatible with gender binarism than CC12M. That is, while there were more about 'feminine' behaviors such as wearing makeup and cooking in CC3M, the ways females and males were represented in CC12M were more diverse. Likewise, the annotations used in CC3M were congruent with traditionally recognized gender roles. Another interesting aspect found in the analyses is that when the annotations containing gender neutral personal (pro)nouns (e.g., *they*, *person*, or *student*) described males in the images, more diversified verbs and adjectives were used in the descriptions.

Implications

This study offers meaningful implications for reconsidering the concept of "noise" from a feminist viewpoint and also for understanding the computational construction of meanings using multimodal aspects of discourses.

References

Ballard, S. (2011). Information, noise, et al.. In M. Nunes (Ed.), *Error: Glitch, noise, and jam in new media cultures* (pp. 59–79). Bloomsbury Publishing.

Changpinyo, S., Sharma, P., Ding, N., & Soricut, R. (2021). Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568. https://ieeexplore.ieee.org/abstract/document/9578388

Hainge, G. (2013). *Noise matters: Towards an ontology of noise*. Bloomsbury Publishing.

Jeatrakul, P., Wong, K. W., & Fung, C. C. (2010). Data cleaning for classification using misclassification analysis. Journal of Advanced Computational Intelligence and Intelligent Informatics, *14*(3), 297–302. https://doi.org/10.20965/jaciii.2010.p0297

Shannon, C. E., & Weaver, W. W. (1949). *The mathematical theory of communication*. University of Illinois Press.

Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018, July 15–20). Conceptual captions: A cleaned, hypernymed, image Alt-text dataset for automatic image captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, *1*, pp. 2556–2565. https://aclanthology.org/P18-1238/

Star, S. L., & Bowker, G. C. (2007). Enacting silence: Residual categories as a challenge for ethics, information systems, and communication. *Ethics and Information Technology, 9*(4), 273–280.

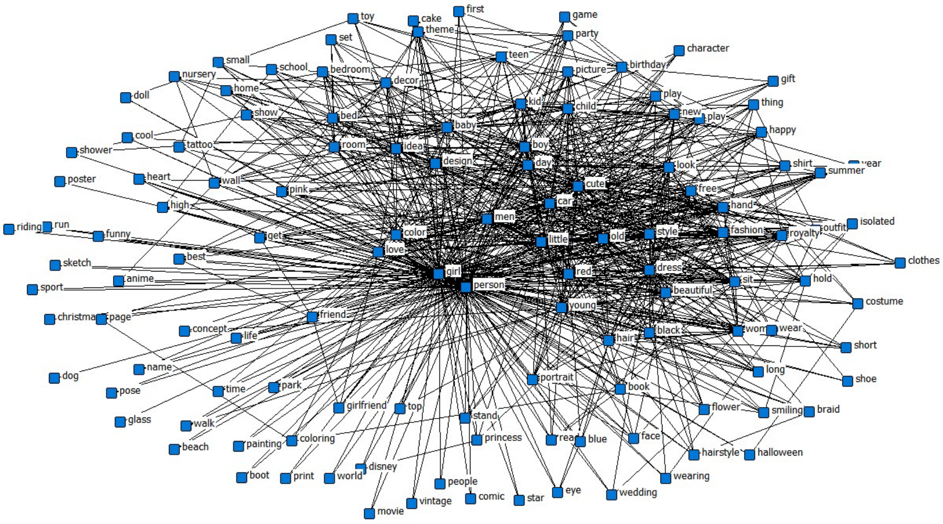

Figure 1. An ego network focusing on the word *girl* from CC3M

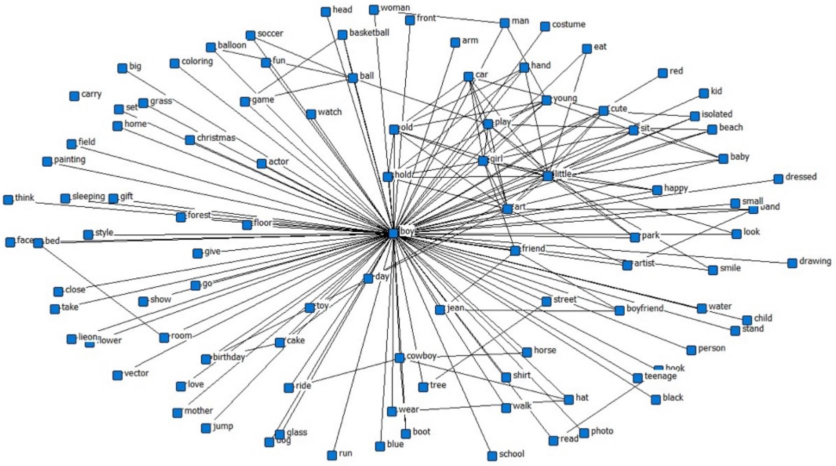

Figure 2. An ego network focusing on the word *boy* from CC3M

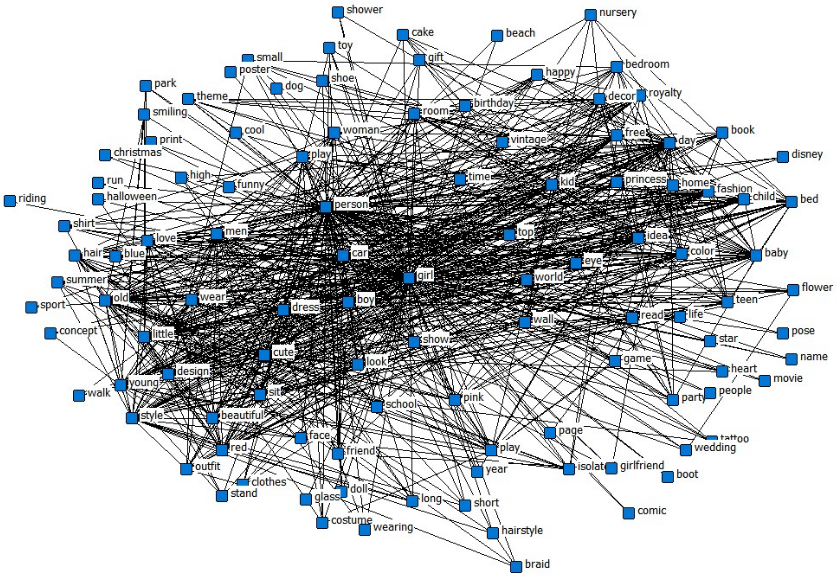

Figure 3. An ego network focusing on the word *girl* from CC12M