Human similarity judgments of emojis support alignment of conceptual systems across modalities

Bryor Snefjella[1]

Yiling Yun[1]
yiling.yun@g.ucla.edu

Shuhao Fu[1]
fushuhao@g.ucla.edu

Hongjing Lu[1,2]
hongjing@ucla.edu

[1]Department of Psychology, [2]Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

Humans can readily generalize their learning to new visual concepts, and infer their associated meanings. How do people align the different conceptual systems learned from different modalities? In the present paper, we examine *emojis*—pictographs uniquely situated between visual and linguistic modalities—to explore the role of alignment and multimodality in visual and linguistic semantics. Simulation experiments show that relational structures of emojis captured in visual and linguistic conceptual systems can be aligned, and that the ease of alignment increases as the number of emojis increases. We also found that emojis with subjective impressions of high popularity are easier to align between their visual and linguistic representations. A behavioral experiment was conducted to measure similarity patterns between 48 emojis, and to compare human similarity judgments with three models based on visual, semantic and multimodal-joint representations of emojis. We found that the model trained with multimodal data by aligning visual and semantic spaces best accounts for human judgments.

Keywords: multimodal representation, alignment, similarity, emoji, visual symbol

Introduction

Two enduring questions in cognitive science concern the relationship between visual and linguistic semantics, and how humans quickly generalize their learning to new visual concepts and their meanings. The concept of *alignment* of conceptual systems across modalities has been proposed as a mechanism that links visual and linguistic systems, and also explains the strong human capacity for generalization (Aho, Roads & Love, 2022; Roads & Love, 2020). Concepts in two modalities can be aligned if the relational structure underlying these concepts in each modality are shared, regardless of the actual features or format of the representations in each modality. Alignment can be contrasted to multimodality, where representational features themselves (but not necessarily relational structure) are shared between modalities. Whereas alignment between modalities has been claimed to be advantageous in representing a common underlying reality (Roads & Love, 2020), modalities also need to represent the unique aspects of concepts within that modality. In the present paper, we examine *emojis*— pictographs uniquely situated between visual and linguistic modalities—to explore the role of alignment and multimodality in visual and linguistic semantics.

Emojis are ideal stimuli for exploring links between vision and language. Emojis are symbols assigned codepoints in the Unicode system, allowing them to be directly embedded in text like traditional orthography, with their visual appearance rendered on computer platforms according to a specific font. In computer-mediated written language use, emojis (unlike most other kinds of images) can thus be directly embedded into local linguistic contexts. Emojis often function as a written approximation to paralanguage—i.e., they function similarly to gesture, pitch contours, intensity and facial expressions in spoken language (James, 2017; McCulloch, 2020). Emojis can also sometimes directly replace some words, serving as content and function words (Na'aman et al., 2017). This embedding of emojis into written language creates shared co-occurrence with words and other emojis in written discourse, enabling the derivation of semantic vectors for emojis from corpora of online language use through distributional semantic models. Hence, their semantic features can be extracted based on how emojis are used in linguistic contexts along with other words and tokens. At the same time, emojis are images with visual properties such as color, texture and object shapes, and can serve as multimodal affective makers (Na'aman et al., 2017). The visual features of emojis reveal their distinctive characteristics that make them appear interesting, engaging and novel. The expressive power of emojis as visual symbols continues to grow with their popularity in social media across languages and cultures, and as new emojis or new variations of emojis are added to Unicode and fonts for Unicode symbols. The versatility of emojis increases as their meanings evolve and use in online platforms expands.

For many emojis, semantic meaning and visual appearance are well aligned (e.g., the smiley-face emoji 😀, and thumbs up sign 👍). But some other emojis may be confusing in terms of alignment with their originally-intended meanings. Whereas some emojis are visually similar, they may be semantically distinct. For example, sad-face emojis 🙁 and happy-face emojis 🙂 are visually similar, but have contrasting affect. Similarly, emojis that have different visual appearance may share similar meanings. For example, the emojis of "person facepalming" 🤦‍♀️ and "face with rolling eyes" 🙄 look quite different, but align with similar meaning

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

and affect. Emojis thus can be used to investigate how different concept systems are aligned under ambiguity.

We first ran simulations to examine the conditions in which visual and linguistic representations for emojis can be aligned easily. Previous work found that visual and linguistic conceptual systems can be aligned, and that the ease of alignment increases as the number of concepts increases (Roads & Love, 2020). However, this previous study used images of real-world objects and their linguistic labels; it is important to determine whether comparable findings can be obtained with human-invented concepts such as emojis. We also investigated whether the popularity of emojis influenced their degree to which the emojis can be aligned. Next, we conducted an experiment to measure human similarity judgments of emojis using an "odd-one-out" task (Hebart et al., 2019). We then compared human judgments with predictions derived from models based only on visual similarity, only on linguistic semantics, or on a joint visual-semantic representation obtained after aligning the visual and semantic spaces.

Alignment between semantic and visual representations of emojis

Roads and Love (2020) performed a computational-level analysis to examine how well different representation spaces can be aligned to reveal correspondences between conceptual systems derived from different sources of input. The intuitive idea is that, despite being from different modalities, inputs based on the same objects come from the same underlying reality. Hence, derived conceptual systems (either visually or linguistically) are constrained to reflect this underlying consistency. For visual images and linguistic texts, similar co-occurrence statistics are likely to be found across the two modalities: functionally similar objects will tend to look alike, and also be described in similar linguistic contexts. Roads and Love indeed found evidence that with a sufficient number of objects, structural relations among objects in one representation space (e.g., visual) can be captured in another space (e.g., semantic). Specifically, when the visual and semantic spaces are aligned with systematic correspondences between visual images and semantic labels, similarity derived from visual embeddings will show the highest correlation with similarity derived from semantic embeddings.

We adapted the same type of analysis to study emojis by quantifying the alignment between visual and semantic embedding spaces. First, we trained a distributional semantic model for emoji use in language via **fastText**, and a visual model for emoji images based on an **auto-encoder (AE)**, to extract semantic and visual embeddings for each emoji respectively.

For the language model fastText, we trained the model on emojis used in posts on Reddit. We found no existing sets of pre-trained word embeddings with large numbers of emojis; accordingly, we collected a text corpus and trained our own purely linguistic emoji embeddings. We queried the Pushshift.io Reddit corpus (Baumgartner, Zannettou, Keegan, Squire & Blackburn, 2020) for all Reddit comments containing emojis. From this text-emoji corpus we eliminated the top 1% of posters who most frequently use emojis (any Reddit user with more than 57 posts containing emojis), as these seemed to be primarily bot accounts. This left 17,082,678 Reddit posts in the Reddit text-emoji corpus. Using this corpus, we trained skip-gram with negative sub-sampling model to create fastText word embeddings (Joulin, Grave, Bojanowski & Mikolov, 2016), with hyperparameters: learning rate 0.05, 300 dimensions, a window size of 5, a minimum frequency count of 3, subword characters from 2 to 4 ngrams, and for 20 epochs. The objective function is to best predict the next token (words or emojis) in the input passages. Emoji embeddings are 300-dimensional latent vectors from the fastText model.

We then trained a visual model auto-encoder (AE) only using emoji images, taken from the most-used emoji image fonts. This autoencoder is trained to reconstruct pixel-level emoji images. The training data included 8665 emoji images and we used another 3775 images for testing. All emojis images were drawn from emoji fonts used on the most popular social media platforms. The encoder is a deep convolutional network including 5 convolutional layers with 3 by 3 kernels, stride 2 and [32, 64, 64, 64, 64] filters in each layer, and leaky relu activations, feeding into a 300-dimensional latent vector. The decoder consists of 5 2d transposed convolutional layers again with 3 by 3 kernels and stride 2, and [64, 64, 64, 64, 32] filters, followed by a final 3 by 3 kernel convolution to reconstruct the emoji images. The object function is to minimize pixel-level deviations between reconstructed emoji images and the input images. Emoji visual embeddings are 300-dimensional latent vectors.

With the language and vision models described above, we also explore relations between our semantic and visual embeddings and human judgments of emoji semantics and appearance. Our source of human judgments of emojis properties is taken from Ferre et al. (2022). These researchers collected subjective ratings for 1031 emojis, using 7-point Likert scales, along six dimensions: visual complexity, familiarity, frequency of use, clarity, valence, and arousal. From the 1031 emojis, we further narrowed down the set to 995 emojis that overlap with use in the Reddit text-emoji dataset. These 995 emojis were input to the pre-trained fastText and AE models to extract their semantic and visual embeddings.

Simulation procedure

Since we have a ground truth of correspondence between semantic and visual embeddings of emojis, we can manipulate the number of emojis that are correctly matched between semantic and visual embeddings, yielding *mapping accuracy*. When mapping accuracy is 1, all emojis have correct correspondences between semantic and visual embeddings. When mapping accuracy is 0.5, semantic embeddings for half of emojis have correct correspondences to their visual embeddings, and the other half of the emojis have the mismatches between their semantic and visual embeddings. We examined 51 levels of mapping accuracy in the range of 0 and 1 with the stepsize of 0.02. For each level of mapping accuracy, 10,000 unique mappings were sampled. For each sample, *alignment correlation* is computed as the Spearman correlation between two similarity matrices, in which one similarity matrix is computed by using visual embeddings and the other from semantic embeddings. The correlation values averaged across 10,000 samples are defined as the alignment correlation for the level of mapping accuracy.

Note that entities with mismatch between semantic and visual embeddings (less than perfect mapping accuracy) could yield higher alignment correlation between uni-modal similarity matrices. Figure 1 illustrates a toy example.

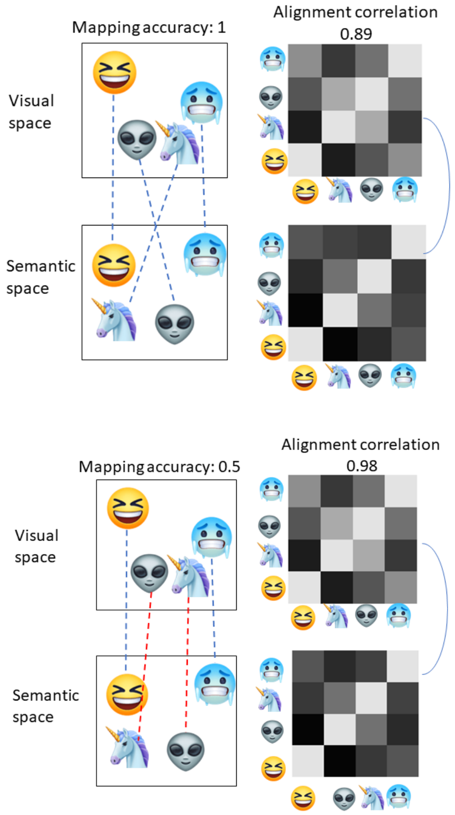

Figure 1. An illustration shows that imperfect mapping with mapping accuracy of 0.5 (bottom panel) yields higher alignment correction between the two similarity matrices than does perfect mapping (top panel). The dash lines indicate mapping of emojis. The red dash lines indicate the mismatched emoji embeddings between the visual and semantic representations.

Results

To examine the impact of the number of emojis on the alignment performance, we ran a spectral clustering algorithm to select emojis closest to the centroid embeddings for 10, 20, 50, 100, 300, 600, and 900 clusters. We found that the correlation between mapping accuracy and alignment correlation increased with more emojis, with correlation ranging from 0.17 for 10 emojis to 0.97 for 900 emojis, replicating the finding reported by Roads and Love (2020). As shown in Figure 2, the fewer emojis are considered, the more likely that systems with misaligned emojis yield high (spurious) correlations between similarity patterns derived from visual and semantic embeddings. In other words, a large number of emojis are likely to exhibit similarity relations shared between visual and semantic representations, which enables easy alignment of the two conceptual systems. In addition, with increased numbers of emojis, the region of *misleading mapping*—showing higher alignment correlation of similarity from visual and semantic matrices than the correct mapping— is reduced significantly. This result shows that maximizing alignment correlation based on similarity between visual and semantic embeddings does not warrant a perfect mapping between the two systems. But as the number of emojis increases, such inconsistency is reduced.

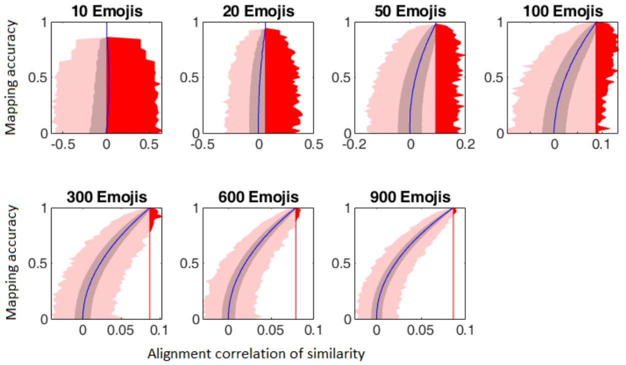

Figure 2. Distribution of alignment correlations between similarity of visual and semantic embeddings conditional on mapping accuracy. Each plot shows the mean alignment correlation (blue line), a one standard deviation envelope (blue shading), the range envelope (pink shading). The pink straight line marks the value of alignment correlation of similarity for the correct mapping. The red regions on the right side of the pink line indicate misleading mappings, with imperfect matches that yield higher alignment correlation of similarity than does the correct mapping.

As defined by Roads and Love (2020), we use *alignment strength* to quantify the prevalence of alignable mappings: the probability that maximum alignment correlation of similarity reveals the best and correct mappings between visual and semantic spaces. The alignment strength is 1 if there is no misleading mapping. When half of incorrect mappings are misleading mappings that show higher alignment correlation than does the correct mapping, the alignment strength is 0.5. The alignment strength corresponds to the pink regions in Figure 2 plots.

Next, we examined the impact of specific sets of emojis on the alignment strength of visual and semantic representations. We focused on the comparison between emojis that are rated as having high familiarity by human participants (subjective high frequency), and emojis with high frequency of usage as determined by objective frequency data from their usage in Reddit (objective high frequency). As shown in Figure 3, we found that the alignment strength for emojis based on subjective frequency ratings of familiarity are higher than that for emojis based on objective frequency, especially when the number of emojis is small (such as the top 20 emojis). This result suggests that emojis that people judge to be used more frequently may be easier to align between semantic and visual embeddings. This ease of alignment is probably due to the visual expressiveness of this subset of emojis, which also may enhance the subjective impression of their high frequency of usage. We acknowledge that both subjective frequency ratings and objective frequency data are aggregated across participants, which may be not the subsets that are most frequently used by an individual Reddit user.

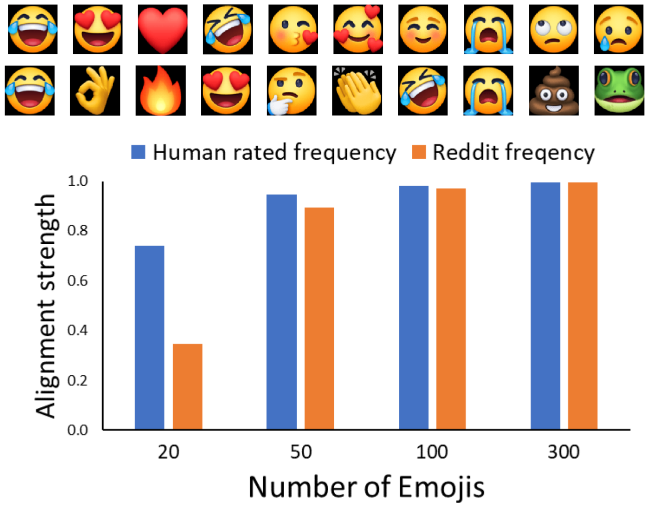

Figure 3. Top row, top ten emojis with high subjective frequency according to Ferre et al. (2022). Bottom row, top ten emojis with high objective frequency based on their usage in Reddit. Bar graph, alignment strength for emojis with subjectively-rated high frequency versus emojis with high usage frequency in reddit.

Human similarity experiment

Method

Participants Fifty-six undergraduate students in the Psychology department at University of California, Los Angeles participated in the online experiment. We excluded five participants who self-reported not being serious throughout the experiment, and two additional participants who failed the practice trials (which involved very easy odd-one-out judgments) more than once. We analyzed data from the remaining 49 participants.

Stimuli We used emojis created by Facebook for the behavioral experiment and modeling. We applied the spectral clustering algorithm to find 50 clusters among a total of 1669 emojis based on their semantic embeddings from the fastText model. We further divided six large clusters that contained 60 or more emojis into smaller clusters using the spectral clustering algorithm, based on their visual embeddings. In each cluster, we selected one emoji that was the most representative (i.e., closest to the centroid). We used these 48 emojis for the behavioral experiment and for modeling.

The experiment was programmed in HTML, JavaScript, CSS, and PHP. On each trial, we displayed three emojis side by side in the center of the computer screen. Each emoji was 100 px * 100 px with a 200 px gap between each two of them.

Design We used the odd-one-out paradigm in the behavioral experiment to assess the similarity between each pair of the 48 emojis. On each trial, we displayed three emojis and asked participants to select the odd-one-out. When a participant selected one emoji as the odd one, their response implied that they considered the two unselected emojis to be more similar to each other than to the selected emoji. On each trial, each emoji thus serves as a context for the other two emojis.

To test each pair of emojis against all remaining 46 emojis, we created the full combination of 17,296 unique trials and randomly assigned them to 46 different versions of the experiment. In each version, we also included six easy trials for which the odd one was obvious, so as to identify and exclude participants who were making unsystematic decisions. Each participant received one version. The order of the trials for each participant was randomized. The position of the three emojis was also randomized for each trial. In total, each participant completed 382 trials.

Procedure Participants accessed the experiment from their personal computers. They first read the instructions about the task and were shown example emojis including faces, animals, objects, and symbols. They then familiarized themselves with the task through three example trials. After an instruction quiz question that tested their understanding of the task, they gave consent to start the experiment. There was no time limit for their decisions. No feedback was given, so they were not guided to make judgments in a particular way. There was a progress bar at the top of the screen. They could only proceed to the next trial after they had clicked on an emoji to select it. After completing all the trials, we administered some survey questions to ask if they were serious throughout the experiment, had any comments about the study, or had encountered any technical issues. The experiment lasted about 30 minutes.

Models

To compare with human similarity judgments, we used the fastText semantic model and AE vision model, and added one more model based on joint representations from two modalities. **CLIP** (Contrastive Language Image Pretraining, Radford et al., 2019) is a deep neural network-based model to create joint visual and linguistic embeddings. The model consists of an image encoder and text encoder which are trained to align visual and semantic representations by projecting image and text to a joint embedding space. The model is trained to discriminate between true and false pairs of image and image caption using the dot product between these representations. We used a pre-trained CLIP model of the clip-vit-large version by OpenAI via the huggingface/transformers python library. This model is trained with a corpus of 400 million captioned images. We then ran the CLIP model with emoji inputs to derive three types of embeddings. For CLIP Language Embeddings of emojis, we took the top layer of the CLIP text encoder, using the Unicode symbol name for the emoji as the textual input; we could not use the emoji symbol itself as the CLIP language tokenizer's dictionary contains no emojis. The CLIP linguistic embeddings contained 768 dimensions. For CLIP Visual Embeddings, we took the top layer of the CLIP image encoder, using a rasterized emoji image as input, to generate 768-dimensional visual embeddings. For CLIP Vision and Language Embeddings, we concatenated the two embeddings.

Results

We computed the similarity matrix from human responses in the odd-one-out task (Figure 4). To be specific, each grid represented the proportion of trials in which the two corresponding emojis were judged as similar (i.e., not selected as the odd one out) among all trials with the two emojis. We used the split-half method to calculate the noise ceiling of human responses. We randomly splitted the human results to two groups of equal size and calculated the correlation between these two groups. After repeating this process for 50 times, we calculated the mean of the correlations and found that the noise ceiling of human similarity judgments was 0.85 ($p < .001$; CI = [0.836, 0.860]). A strong model would show correlation to human judgments closer to the noise ceiling.

To compute similarity matrices predicted by the models, we calculated the pairwise cosine similarity using emoji embeddings extracted from the three models, vision embeddings from AE model, semantic embeddings from fastText model, and joint embeddings from CLIP model. The model similarity matrices are shown in Figure 5.

We then compared the human similarity judgments with the modeling results (Figure 6), by computing the Spearman correlation between human similarity judgments and model-predicted similarity. CLIP showed the highest correlation (ρ = .38). The fastText model generated the second-highest correlation (ρ = .36). The AE model showed the lowest correlations (ρ = .17). We conducted the Mantel test to show that all the correlation coefficients were significantly greater than zero (*ps* < .001).

We next examined the semi-partial correlation between the CLIP model and human similarity judgments, controlling for semantic fastText on human similarity judgments. We found that the semi-partial correlation maintained significant (*sr* = .34, $p < .001$). Controlling for visual model (AE) on human similarity judgments, the CLIP model showed a significant semi-partial correlation with the human similarity judgments (*sr* = .38, $p < .001$). To test if the difference was simply because CLIP embeddings contains more information involving better text or visual inputs, we concatenated fastText embeddings and AE embeddings are found a lower correlation to human similarity judgments (ρ = 0.17). Hence, the aligned representations in CLIP showed a better account to human similarity judgments than merging the visual and semantic embeddings derived from independent models via simple concatenation.

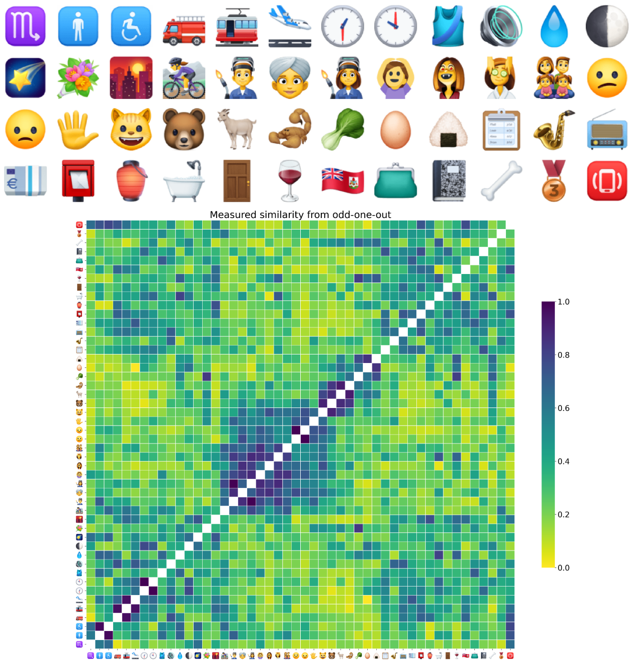

Figure 4. Similarity matrix from 48 emojis (shown in the top panel) derived from human responses in the odd-one-out task. Darker blue indicates higher similarity, and yellow indicates lower similarity.

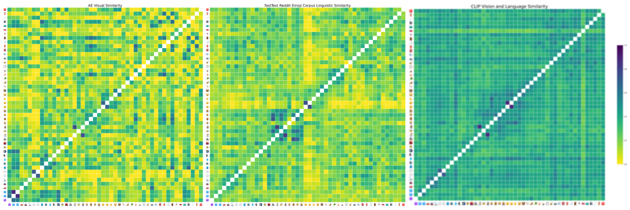

Figure 5. Similarity matrices predicted from visual model (AE), semantic model (fastText), and joint vision-language model (CLIP). We derived them from the calculated distance matrices for easier interpretation in graphs.

We also compared our results with similarity matrix derived from human ratings of emoji features (Ferre et al., 2022). Using 7-point Likert scales, each emoji was rated across six dimensions: visual complexity, familiarity, frequency of use, clarity, valence, and arousal. We used rating responses to create a 6-dim vector for each emoji, and then computed the pairwise cosine similarity for 47 emojis used in our experiment to generate the similarity matrix.

We compared the similarity calculated from the odd-one-out task in our experiment with similarity derived from six ratings collected by Ferre et al. (2022). We found that correlation between odd-one-out similarity and subjective ratings was the lowest (ratings: ρ = .12). We then compared similarity derived by ratings and model-predicted similarity. Only the similarity matrix by fastText model significantly correlated with similarity derived from human ratings of six dimensions (ρ = .18, p < .001). Other models did not show significant correlations with similarities derived from human ratings. We then expanded the analysis to 995 emojis in the ratings dataset. The fastText model still showed the highest correlation with the ratings data (ρ = .29). The CLIP model showed low correlation (ρ = .12). The visual model AE model did not show significant correlation with ratings (p = .27). Note that CLIP similarity showed the highest correction with human similarity derived from the odd-one-out task, but relatively weak relation to similarity derived from human rating data. This difference suggests the impact of task on emoji representation: the task of asking people to provide ratings for single emoji image recruits more of semantic representations; whereas the odd-one-out task elicits both visual and semantic representation through comparisons.

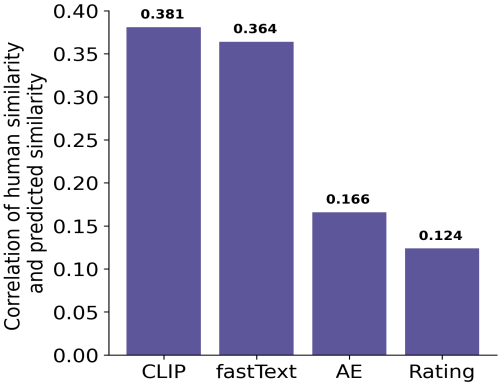

Figure 6. Results of correlation between human similarity from the responses in the odd-one-out task and similarity derived by the models and human ratings. The noise ceiling of human responses is 0.85.

General discussion

Through simulations, we replicated earlier work with emojis showing that visual and linguistic conceptual systems for emojis can be aligned, and that the ease of alignment increases as the number of concepts increases (Roads & Love, 2020). We also found that subjectively familiar emojis were easier to align between visual and linguistic embeddings than were emojis with high objective frequency of usage on Reddit. It is possible that people use the ease of aligning the visual appearances and intended meanings of emojis to estimate their popularity.

To further examine the contributions of visual and semantic information in emoji representations, we performed a representational similarity analysis using an "odd-one-out" task (Hebart et al., 2019). We compared human similarity judgments with pairwise similarities predicted by a vision model, a language model, and a multimodal model jointly based on visual and linguistic semantics. We found that the model trained with multimodal data produced the strongest correlation with human similarity judgments. This finding suggests that humans rely on a joint representation that captures visual appearance of emojis and their usage in linguistic contexts. When the language model was compared with the visual model, the former proved to be the better predictor of human similarity judgments. This finding is consistent with the primary role of emojis as an effective means of communication. Emojis can be considered as a representative example of symbols. In the words of Saint Augustine, "symbols are powerful because they are the visual signs of invisible realities". The expressive power of emojis arises from the semantic and visual representations aligned in their conceptual system.

We found that emojis high on an objective measure of emoji frequency were less alignable than emojis with high subjective familiarity. One possible explanation for the superiority of subjective ratings of familiarity over our objective measure of emoji frequency from Reddit could be that emoji use on Reddit is different from other social media platforms, texting, or other genres of text where emojis are used. Reddit also differs from other platforms in the demographic composition of its users (Amaya, Bach, Keusch & Kreuter, 2021) which in turn may affect emoji use. Emoji corpora constructed from other sources and fastText embedding trained on them could clarify this issue.

The present study also illustrates how different task demands can elicit different aspects of representations for multimodal stimuli. Asking human participants to explicitly rate specific dimensions (e.g., familiarity, valence, and arousal) is a common method in psychology. We found, however, that using these ratings tasks to compute similarity led to a shift in emphasis, such that the language model alone was a better predictor of human similarity than was the multimodal model. Thus, these rating tasks elicited a more purely semantic (rather than also visual) representation of emojis. When the task was changed to making comparisons of multiple emojis, multimodal representations were more likely to be recruited. The odd-one-out task used in the current experiment did not include any explicit instruction regarding how to compare the emojis to make a judgment; thus, it appears that joint vision-language embeddings correspond to the default mode elicited by the task. Future research should systematically examine the influence of task demands on the flexible use of multimodal representations. In addition, because the present experiment tested a relatively small set of emojis, future work should examine larger sets of emojis to discover their interpretable latent representations. Probing the psychological representation of emojis will be an important tool to advance our understanding of human learning in multimodal environments.

Acknowledgements. We dedicate this paper to the memory of our co-author Dr. Bryor Snefjella, whose creativity, brilliance, insight and generosity make this project possible. Preparation of this paper was supported by NSF Grant IIS-1956441 awarded to H.L.

References

Aho, K., Roads, B. D., & Love, B. C. (2022). System alignment supports cross-domain learning and zero-shot generalization. *Cognition*, *227*, 105200. https://doi.org/10.1016/j.cognition.2022.105200

Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2021). New data sources in social science research: Things to know before working with Reddit data. *Social science computer review*, *39*(5), 943-960.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The pushshift reddit dataset. *Proceedings of the international AAAI conference on web and social media, 14*(1), 830-839. https://doi.org/10.1609/icwsm.v14i1.7347

Ferré, P., Haro, J., Pérez-Sánchez, M. Á., Moreno, I., & Hinojosa, J. A. (2022). Emoji-SP, the Spanish emoji database: Visual complexity, familiarity, frequency of use, clarity, and emotional valence and arousal norms for 1031 emojis. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01893-6

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2019). Revealing the behaviorally-relevant dimensions underlying mental representations of objects. *Journal of Vision*, *19*(10), 170b. https://doi.org/10.1167/19.10.170b.

James, A. (2017). Prosody and paralanguage in speech and the social media: The vocal and graphic realisation of affective meaning. *Linguistica*, *57*(1), 137-149. https://doi.org/10.4312/linguistica.57.1.137-149

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759.*

McCulloch, G. (2020). *Because internet: Understanding the new rules of language*. Penguin.

Na'aman, N., Provenza, H., & Montoya, O. (2017). Varying linguistic purposes of emoji in (Twitter) context. In *Proceedings of ACL 2017, student research workshop*, 136-141.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research, 139*, 8748-8763.

Roads, B. D., & Love, B. C. (2020). Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, *2*(1), 76-82. https://doi.org/10.1038/s42256-019-0132-2

Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. *Proceedings of the Neural Information Processing Systems*, *30*.