Uncovering Visual-Semantic Psycholinguistic Properties from the Distributional Structure of Text Embedding Space

Si Wu[*]
Northeastern University
siwu@ccs.neu.edu

Sebastian Bruch[*]
Northeastern University
s.bruch@northeastern.edu

Abstract

Imageability (potential of text to evoke a mental image) and concreteness (perceptibility of text) are two psycholinguistic properties that link visual and semantic spaces. It is little surprise that computational methods that estimate them do so using parallel visual and semantic spaces, such as collections of image-caption pairs or multi-modal models. In this paper, we work on the supposition that text itself in an image-caption dataset offers sufficient signals to accurately estimate these properties. We hypothesize, in particular, that the peakedness of the neighborhood of a word in the semantic embedding space reflects its degree of imageability and concreteness. We then propose an unsupervised, distribution-free measure, which we call *Neighborhood Stability Measure (*NSM*)*, that quantifies the sharpness of peaks. Extensive experiments show that NSM correlates more strongly with ground-truth ratings than existing unsupervised methods, and is a strong predictor of these properties for classification. Our code and data are available on GitHub.[1]

1 Introduction

The psycholinguistic concept of *imageability* is a measure of how easily text evokes a clear mental image (Paivio et al., 1968). "Beach" and "banana" are examples of highly-*imageable* words. By contrast, "gratitude" and "criterion" have low imageability. Other words fall somewhere in between.

Imageability has been shown to correlate with another psycholinguistic property called *concreteness* (Paivio et al., 1968; Richardson, 1976; Ellis, 1991). It measures the degree to which a piece of text denotes a perceptible entity (Brysbaert et al., 2014). Studying and measuring these properties can help us understand how we process text cognitively and further our understanding of perception.

Imageability and concreteness are traditionally measured by surveying individuals—an expensive and laborious process. The costs involved have motivated researchers to develop computational methods to estimate these measures more economically. As a direct consequence, these methods can help us expand the vocabulary for which imageability and concreteness ratings are available.

Among the *unsupervised* methods, which is the focus of this work, one recent study is the work by Wu and Smith (2023) which estimates imageability using a text-to-image model. The authors prompt the model with a query and ask it to generate images. They show that the average CLIP scores (Radford et al., 2021) of generated images correlate with ground-truth imageability ratings.

In an earlier study, Hessel et al. (2018) propose an unsupervised method to estimate concreteness. They use a dataset of image-caption pairs and measure how "clustered" a word's associated images are. As they demonstrate, this measure correlates reasonably with ground-truth concreteness ratings.

While these unsupervised methods have lowered the costs of estimating psycholinguistic measures, they each have their own limitations: The method of Wu and Smith (2023) is computationally expensive as it involves sampling, per input text, a number of images from a large, generative model. Sampling too few images can also lead to a high variance in the estimated measurements.

The method of Hessel et al. (2018) presents its own challenges. It requires access to a parallel image-caption dataset, which can be expensive to acquire.[2] Additionally, because their method filters images by their associated words, it is insensitive to word senses and is further susceptible to the *vocabulary mismatch* problem: Small variations in spellings can affect the measurements. Indeed, as

---

[*]Both authors contributed equally to this work.
[1]https://github.com/Artificial-Memory-Lab/imageability

---

[2]Many large image-caption collections require researchers to crawl the images independently from the web.

we will explain later, this method is able to predict the concreteness of roughly 2% of words for which ground-truth ratings exist.

We set out to address these limitations and present a unified, *unsupervised* approach for estimating the imageability and concreteness of words. First, we make the case for *single* modality and aim to quantify the psycholinguistic properties of interest using only the captions of an image-caption dataset. Our hypothesis is that, whether or not words paint a picture can be determined by the contexts in which they appear; access to the visual modality is not strictly necessary.

Second, we remedy the vocabulary mismatch problem by operating in the *semantic embedding* space, similar to what was done by Wu and Smith (2023). That is unlike the method by Hessel et al. (2018), which uses lexical signals only.

Finally, we hypothesize that, given a sufficiently large collection of semantic embeddings of text, the structure of the embedding space itself contains clues on the concreteness and imageability of individual words. Working directly with the vector collection obviates the need to sample from the underlying distribution through a generative process.

Importantly, such a vector collection can be constructed once, and reused and recycled indefinitely. That is in direct contrast to Wu and Smith (2023), whose method requires executing multiple rounds of inference over a large, multi-modal model to quantify the imageability of a single input.

With those objectives in mind, we put forward the following contributions:

- We hypothesize that, in a collection of text embeddings, the neighborhoods around the embeddings of concrete (imageable) words are *structurally different* from those of abstract (non-imageable) words. To support our hypothesis, we present an empirical observation that suggests the former group is more *stable* than the latter.

- We present a novel, unsupervised, and distribution-free measure to quantify the degree of stability of neighborhoods in an embedding collection, which we call the *Neighborhood Stability Measure* (NSM). As we later show empirically, a piece of text with a more "stable" neighborhood in the embedding space is more imageable or concrete. Our method, which is straightforward to implement and fast to execute, draws from the literature on approximate nearest neighbor (ANN) search (Bruch, 2024).

- We conduct a comprehensive experimental evaluation of our proposed measure, through which we confirm the validity of our hypothesis above. Furthermore, a comparative analysis shows that our algorithm surpasses the performance of existing methods by a substantial margin.

The remainder of this work is organized as follows: We review the relevant literature in Section 2. We describe the datasets and embedding models used in this work in Section 3 to facilitate discussions in subsequent sections. Section 4 presents our empirical observation and proposed algorithm. We evaluate our method and compare it with baselines in Section 5. We discuss our findings in Section 6, before concluding in Section 7.

2 Related Work

A word's imageability rating is usually obtained by interviewing human subjects, where ratings are on a 7-point Likert scale (Toglia and Battig, 1978; Gilhooly and Logie, 1980; Coltheart, 1981; Wilson, 1988). The process is costly and laborious. Consequently, the resulting database has a limited vocabulary size. The psycholinguistics database where we obtained our ground-truth imageability ratings, MRC Psycholinguistics Database, combines three different large psycholinguistics databases, yet has a vocabulary of 9,240 words (Coltheart, 1981), of which only 4,848 have imageability ratings.

There have been attempts to expand MRC ratings: Schock et al. (2012) added ratings for 3,000 disyllabic words by interviewing human subjects; Liu et al. (2014) use synonyms and hyponyms identified in WordNet. The limited success of these attempts presents an opportunity to develop unsupervised computational methods to estimate ratings and expand the vocabulary at a lower cost.

Wu and Smith (2023), for instance, predict the imageability of an input text by prompting a text-to-image model and calculating the average CLIP scores (Radford et al., 2021) of the generated images. While this method is unsupervised and can estimate the imageability of arbitrary pieces of text, it is computationally expensive as it requires multiple forward passes over a large model per input. Unlike this method, we operate in single modality without the help of a generative model.

Separately, the MRC Psycholinguistics Database also contains concreteness ratings of individual words, which have been shown to correlate with imageability ratings (Paivio et al., 1968; Richardson,

1976; Ellis, 1991). Later, its vocabulary was expanded to 37,058 words by Brysbaert et al. (2014) using Amazon Mechanical Turk.

Like imageability, a number of computational methods have been developed to predict concreteness. Among these, Hessel et al. (2018) rely on an image-caption collection to quantify how clustered a word's associated images are, serving as a proxy to its concreteness.

The focus of our work is on unsupervised methods, and as such, in our experiments, we focus exclusively on other unsupervised algorithms. However, we would be remiss if we did not note the body of literature on supervised approaches to the task of predicting imageability and concreteness.

Tater et al. (2024), for example, explore the role of *visual* features in determining abstractness versus concreteness, and present a supervised classifier trained on such features. Our method, in contrast, is unsupervised and relies on *textual* features only.

Other works have used lexical and semantic statistics to determine concreteness. Naumann et al. (2018), for instance, are concerned with the distributional differences between abstract and concrete words, but where the distributions are based on part-of-speech tags, and the predictive feature is entropy. We, on the other hand, explore the space of embeddings generated by neural networks, with the predictive feature being the peakedness of distributions in the high-dimensional vector space.

Similarly, the works by Frassinelli et al. (2017) and Schulte im Walde and Frassinelli (2022) all consider distributional properties of text where the distribution is based on co-occurrences (involving part-of-speech tags or other semantic features of individual words). In contrast, the distributions we are concerned with are based on learned representations of text, often trained for determining similarity between pairs.

Other methods include: Lenci et al. (2018) study how Distributional Affective Scores (DAS)—a quantity estimated by a supervised method—differ between concrete and abstract words. The supervised method by Charbonnier and Wartena (2019) trains a regression model over word embeddings and morphological features. Similarly, Kastner et al. (2020) use visual features such as color distribution to estimate imageability.

Finally, the recent study by Martínez et al. (2024) explores the ability of (private) Large Language Models to predict concreteness. Besides reproducibility limitations due to its dependence on closed-source (and ever-changing) models, this method is also entirely distinct from our unsupervised method that leverages publicly-available datasets and models.

3 Datasets and Embeddings

Throughout this work, we ground our discussions in specific datasets and embeddings so as to provide illustrative examples and evaluate our proposal. To facilitate that, we introduce these early.

Datasets: As we will explain in subsequent sections, our method estimates the imageability and concreteness of English words using a (single-modal) text dataset. In particular, we use the text portion of the datasets listed below:

- **COCO Captions** (MS COCO): A dataset that contains 1.5 million captions describing over $330,000$ images from the Microsoft Common Objects in Context dataset (Lin et al., 2015). The image captions were collected using Amazon Mechanical Turk (Chen et al., 2015).

- **Conceptual Captions** 3**M** (CC3M): A dataset that contains 3.3M filtered image-caption pairs harvested from the web. The description for each image was obtained from the alt-text in the HTML attribute (Sharma et al., 2018).

- **Conceptual Captions** 12**M** (CC12M): Collected by relaxing the pipeline used to construct CC3M: the captions are less precise, but the dataset is $4\times$ larger (Changpinyo et al., 2021).

Embedding models: Our method operates in the semantic embedding space associated with a (text) dataset; not in the lexical space. As such, we must transform the datasets described above into vector collections using text embedding models. We do so using the following publicly-available models:

- ALLMINILM: This model has been trained on a large number of datasets, exceeding a total of one billion pairs of text.[3] It produces 384-dimensional embeddings with cosine similarity as distance function. The model is compact in size (with 33M parameters) and efficient to infer, yet offers competitive quality for semantic retrieval tasks.[4]

---

[3]Checkpoint available at `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

[4]Visit `https://sbert.net` for additional details.

- GTE-BASE and GTE-LARGE: The GTE-BASE-EN-V1.5 model and its larger variant GTE-LARGE-EN-V1.5 are embedding models with 137 million and 434 million parameters respectively (Li et al., 2023; Zhang et al., 2024).[5,6] The smaller model produces 768-dimensional embeddings and the larger 1024-dimensional vectors.

The reason we chose these models is two-fold. First, the three models cover a reasonable range of parameter sizes. Second, the dimensionality of their embedding space progresses from 384, to 768, and finally 1024. This variety in embedding sizes allows us to examine the effect of the number of dimensions on the performance of our method.

Finally, we adopt the following naming convention: When dataset $\mathcal{T}$ is embedded with model ϕ, we denote the resulting vector collection by $\phi(\mathcal{T})$. For example, the collection of GTE-BASE embeddings of CC3M is denoted by GTE-BASE(CC3M).

4 Methodology

Consider the imageable word "beach" and contrast it with the non-imageable word "thing." Intuitively, the set of contexts in which the former can be used should be more limited than the contexts in which the latter could appear: "thing" can be used in virtually any context. The same can be said about concrete versus abstract words. In this section, we develop that intuition into a practical method.

4.1 Structural clues in the embedding space

Suppose that the intuition above holds, so that the variability of the contexts a word appears in correlates with its imageability or concreteness. The first step in formalizing that intuition is to translate it into the "semantic space": A space governed by a distribution from which natural text is sampled.

Translated thus, we arrive at the following statement, which is our hypothesis:

Hypothesis 1. *The distribution of contexts around an imageable or concrete word forms a sharper peak in the semantic space.*

That line of reasoning implies that the neighborhoods of words in the semantic space encode structural clues about the degree of imageability or concreteness. An imageable word should fall in a more concentrated region of the distribution of text
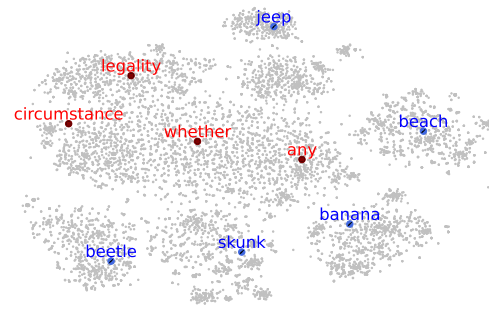

Figure 1: Visualization of the neighborhoods of a set of imageable words (in blue) and non-imageable words (in red) in the embedding collection of ALLMINILM(CC3M). We use tSNE (van der Maaten and Hinton, 2008) for dimensionality reduction. Neighborhoods of non-imageable words have a wider spread and overlap with each other, whereas imageable words have comparatively more separable neighborhoods.

than a non-imageable word. Indeed, without explicitly noting it, that is what Wu and Smith (2023) attempt to model, but in a parallel "image space."

Unfortunately, we do not have full knowledge of the structure of the semantic space. Testing the hypothesis above directly is therefore out of the question. However, there is no shortage of large text collections made up of rich content, "sampled" from a broad region of the semantic space. Furthermore, with semantic embedding models, we can transform pieces of text from such collections into vectors. Such a large collection of embedding vectors can indeed serve as a surrogate to the semantic space itself, closely capturing its structure.

The method developed and presented next gives us a tool to test Hypothesis 1. Before we proceed, however, we visualize an illustrative example that hints at the validity of the hypothesis. In Figure 1, we render the neighborhoods of select imageable and non-imageable words in the embedding space of ALLMINILM(CC3M), using tSNE (van der Maaten and Hinton, 2008) for dimensionality reduction. The neighborhoods of imageable words appear to be more separable than the neighborhoods of non-imageable words, respectively reflecting the sharpness and flatness of the peak of the distribution around these words.

4.2 The Neighborhood Stability Measure

We design a method to quantify the peakedness of the distribution *around any given neighborhood* in the embedding space of a text collection. Because

we wish for this quantity to serve as a surrogate estimate of the peakedness of the (unknown) distribution that underlies the semantic space, we limit our attention to distribution-free measures. To that end, we draw insights from the literature on approximate nearest neighbors (Bruch, 2024).

Preliminaries: Consider a semantic embedding space in $\mathbb{R}^d$ equipped with the similarity measure $\delta : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$. Typically, $\delta(\cdot, \cdot)$ is cosine similarity or inner product. Denote a *query* point by $q \in \mathbb{R}^d$, representing the embedding of the text whose imageability and concreteness we wish to quantify. Denote further a collection of semantic embeddings, $\mathcal{X} \subset \mathbb{R}^d$, as projections of a large text corpus into the embedding space.

For a fixed k, denote by $\mathcal{N}_k(q)$ the subset of k points from $\mathcal{X}$ that are closest to q according to δ:

$$\mathcal{N}_k(q) = \arg\max_{u \in \mathcal{X}}^{(k)} \delta(q, u), \qquad (1)$$

where the superscript (k) indicates that the argmax operator returns the k maximizers of its argument. In practice, we relax $\mathcal{N}_k(q)$ to be the *approximate* set of nearest neighbors to facilitate scalability.

Concepts: Take a neighborhood $\mathcal{P}$ in some collection $\mathcal{X} \subset \mathbb{R}^d$. We say that $\mathcal{P}$ is α-stable, for some $\alpha \in [0, 1]$, if the proportion of points in $\mathcal{P}$ whose nearest neighbor also belongs to $\mathcal{P}$ is α. Formally:

Definition 1 (α-Stable Neighborhood). *A neighborhood $\mathcal{P}$ in $\mathcal{X} \subset \mathbb{R}^d$ is α-stable for $\alpha \in [0, 1]$ calculated as follows:*

$$\alpha = \frac{1}{|\mathcal{P}|} \Big| \{ u \in \mathcal{P} \mid \mathcal{N}_1(u) \in \mathcal{P} \} \Big|, \qquad (2)$$

where $\mathcal{N}_1(\cdot)$ is defined as in Equation (1) over $\mathcal{X}$, and $|\cdot|$ returns the cardinality of its argument.

This definition is inspired by the concept of "natural neighbors" (Zhu et al., 2016): Two points are natural neighbors if they are each others' nearest neighbors. In effect, we extend that concept to sets of points and quantify it as a continuous measure.

Let us take a moment and understand what α-stability says about a neighborhood visually. In Figure 2, we illustrate two neighborhoods. One with perfect stability (i.e., $\alpha = 1$) and another with stability factor $\alpha = 0.4$. The former is separable from the rest of the collection, while the latter is surrounded by other points. Assuming that these points are drawn from a distribution, the separability or inseparability of a neighborhood is telling
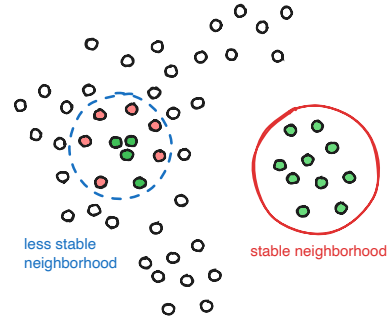

Figure 2: Illustration of a stable neighborhood (encircled by solid red boundary) versus a less stable neighborhood (inside the dashed blue region). The nearest neighbor of every point in the former belongs to the same neighborhood, whereas only 4 out of 10 points in the latter have that property. If these points are sampled from an unknown distribution, the *stability* of these neighborhoods reflects the sharpness or flatness of the regions around them in the underlying distribution.

of the sharpness or flatness of the corresponding regions in the distribution.

Taking k as a hyperparameter and assuming access to an approximate nearest neighbor (ANN) search algorithm, we can measure the α-stability of the neighborhood in a collection $\mathcal{X}$ around any arbitrary query point q. We call this quantity q's Neighborhood Stability Measure with radius k, defined formally as follows:

Definition 2 (Neighborhood Stability Measure for point q and radius k). *Suppose $\mathcal{N}_k(q)$ is an α-stable neighborhood. Then the Neighborhood Stability Measure (NSM) of q given radius k is α.*

Algorithm: We use the NSM of the query point q with radius k as an estimate of the imageability and concreteness of the text represented by q. We leave k as a hyperparameter to be tuned using a validation set of queries given a text collection. Note that, an NSM closer to 1 indicates a higher imageability and concreteness, as per Hypothesis 1.

The algorithm to compute the NSM of a query is described in Algorithm 1. The complexity of a naïve implementation of that is dominated by $\mathcal{O}(kT)$ where T is the complexity of the ANN search subroutine. In a practical implementation, one can populate a table mapping each point in the collection to its (approximate) nearest neighbor upfront; the call to the ANN function on Line 4 of Algorithm 1 can then be replaced with a table lookup, reducing the complexity of the algorithm to $\mathcal{O}(T)$.

Algorithm 1 Neighborhood Stability Measure

Input: Query $q \in \mathbb{R}^d$; embedding collection $\mathcal{X} \in \mathbb{R}^d$; radius, k; subroutine $\text{ANN}_\mathcal{X}(q, k)$ that returns the k approximate nearest neighbors of q from $\mathcal{X}$.
Output: NSM of q with radius k.

1: $\mathcal{N}_k(q) \leftarrow \text{ANN}_\mathcal{X}(q, k)$
2: $\alpha \leftarrow 0$
3: **for** $u \in \mathcal{N}_k(q)$ **do**
4:     **if** $\text{ANN}_\mathcal{X}(u, 1) \in \mathcal{N}_k(q)$ **then**
5:         $\alpha \leftarrow \alpha + \frac{1}{k}$
   return α

We note that, state-of-the-art ANN algorithms are highly efficient and can easily scale to collections of hundreds of millions of points (Bruch, 2024).

5 Empirical Evaluation

We evaluate our proposal in this section. We explain our setup and present our results.

5.1 Setup

We have already introduced the datasets and embedding models used in this work in Section 3. Here, we explain the dataset of ground-truth imageability and concreteness ratings; our evaluation protocol; and, the methods that will be put to the test.

Ground-truth ratings: We use imageability ground-truth ratings for 4,848 words from the MRC Psycholinguistics Database (Coltheart, 1981) collected by interviewing human subjects. The concreteness ground-truth ratings by Brysbaert et al. (2014) were crowd-sourced with Amazon Mechanical Turk, with a vocabulary size of 37,058.

Evaluation protocol: We randomly split each ratings dataset into non-overlapping validation (20%) and test (80%) subsets. If a method requires hyperparameter tuning, we do so on the validation split. We evaluate all methods on the test set only. The measures reported in this work are the average of 10 trials, where each trial uses its own seed for the pseudo-random generator that produces the splits.

We evaluate all methods on the imageability axis as follows: For every word from the imageability ratings dataset, we compute its score according to each method. We then measure the Spearman's correlation coefficient between the ratings and predictions. Evaluation on the concreteness axis follows the exact same logic, except that the set of words come from the concreteness ratings dataset.

Methods: We compare our proposed measure with a number of existing methods from the literature. The complete list of methods is as follows:

- FREQ$_\mathcal{T}$: This method works with the lexicon only. If we are interested in estimating the imageability or concreteness of the word w, we simply count the number of times w appears in the text collection $\mathcal{T}$ and use that as the prediction. This naïve baseline mirrors what was presented in prior work (Wu and Smith, 2023; Hessel et al., 2018). Here, $\mathcal{T}$ is the collection of captions from MS COCO, CC3M, and CC12M.

- HML$_\mathcal{D}$: This denotes the measure developed by Hessel et al. (2018) to predict the *concreteness* of words; the acronym represents the initial letters of the authors' surname. The subscript denotes an image-caption dataset, over which the measure is calculated. In our experiments, we use MS COCO as $\mathcal{D}$ as per the original work. The concreteness score of a word is computed by counting how often the nearest neighbors of an associated image are also associated with the same word. We use the implementation recommended by the author to conduct our experiments.[7]

- AVGCLIP and COSINESIM: Measures proposed by Wu and Smith (2023) to predict the *imageability* of words and phrases. To estimate that quantity for word w, we use the code provided by the authors to present w to a text-to-image model, DALL•E mini, and ask it to generate 16 images.[8] COSINESIM computes the average pairwise cosine similarity between the embeddings of the generated images, while AVGCLIP reports the mean CLIP scores (Radford et al., 2021).

- NSM$_\mathcal{X}$: This is our proposal. When we measure NSM on the vector collection $\mathcal{X}$, we denote it by NSM$_\mathcal{X}$, where $\mathcal{X}$ is the embeddings of MS COCO, CC3M, or CC12M. We use the Faiss library (Douze et al., 2024) to perform ANN search using an Inverted File (IVF) index, where the number of clusters is set to $8\sqrt{|\mathcal{X}|}$ where $|\cdot|$ is the size of its argument. During search, the parameter nprobe is set to 128; this results in a near exact nearest neighbor search over the datasets in our experiments. Finally, we tune the radius k of Algorithm 1 on the validation set, by sweeping

---

[7]For details, visit https://jmhessel.com/projects/concreteness/concreteness.html.
[8]https://github.com/swsiwu/composition_and_deformance

5643

MEASURE	IMAG.	CONC.
FREQ_{MS COCO}	0.26	0.25
FREQ_{CC3M}	0.23	0.30
FREQ_{CC12M}	0.34	0.35
HML_{MS COCO}	0.49	0.45
COSINESIM	0.45	0.40
AVGCLIP	0.56	0.45
NSM_{ALLMINILM(MS COCO)}	0.57	0.53
NSM_{ALLMINILM(CC3M)}	0.61	0.55
NSM_{ALLMINILM(CC12M)}	**0.66**	**0.58**
NSM_{GTE-BASE(MS COCO)}	0.54	0.54
NSM_{GTE-BASE(CC3M)}	0.57	0.58
NSM_{GTE-BASE(CC12M)}	0.58	0.58
NSM_{GTE-LARGE(MS COCO)}	0.45	0.44
NSM_{GTE-LARGE(CC3M)}	0.56	0.51
NSM_{GTE-LARGE(CC12M)}	0.56	0.53

Table 1: Spearman correlation coefficient between ratings and predictions for imageability (Imag.) and concreteness (Conc.). The Neighborhood Stability Measure (NSM) is ours, with the subscript denoting the collection it was measured on. Baselines in the first block include Hessel et al. (2018) (HML); Wu and Smith (2023) (COSINESIM and AVGCLIP); and, word frequency (FREQ).

MEASURE	IMAG.	CONC.
FREQ_{MS COCO}	67.6%	28.7%
FREQ_{CC3M}	84.5%	47.6%
FREQ_{CC12M}	98.0%	71.6%
HML_{MS COCO}	12.9%	2.7%

Table 2: Percentage of words with ground-truth ratings for which each method is able to predict imageability and concreteness. COSINESIM, AVGCLIP, and NSM offer a coverage of 100% by design.

the interval $[64, 4096]$ in increments of 64—we examine the effect of k in Appendix A.

5.2 Results

We present our main results in Table 1, where we report the correlation between the ground-truth imageability and concreteness ratings with the predictions of each method. We note that, even though AVGCLIP and COSINESIM are designed to measure imageability, we measure their correlation on the concreteness dataset as well for completeness. Similarly, we measure the performance of HML on the imageability dataset, despite the fact that it was intended for estimating concreteness.

We find that FREQ yields a stronger-than-expected correlation with both imageability and concreteness. This was not the case in the experiments of Wu and Smith (2023), who report no correlation. That is because, in their experiments, Wu and Smith (2023) use word frequency from the Brown Corpus (Francis and Kucera, 1979), which is a general corpus of natural texts. Here, however, we measure the correlation between the frequency of words in a collection of captions, resulting in a higher correlation and a stronger baseline. We note that, unsurprisingly, larger collections lead to stronger correlation coefficients, with one exception (FREQ_{CC3M} versus FREQ_{MS COCO}).

HML appears to perform better than FREQ;

However, the method is able to predict the imageability of only about 13% of words in the ratings dataset. Similarly, it only predicts concreteness scores for 2.7% of total words. That is due to the vocabulary mismatch problem: the vast majority of words in the ratings dataset do not appear in the captions. As we explained, that is a major limitation of the method of Hessel et al. (2018). Table 2 reports the coverage for all methods.

Finally, AVGCLIP performs better on the imageability dataset, confirming the findings of Wu and Smith (2023). Its performance on the concreteness prediction task is also relatively strong. This method, however, is far more computationally expensive than the other baselines. Using two GeForce RTX 3080 Ti GPUs, it took approximately 120 hours to compute all concreteness scores.

We now turn to our method. Every configuration of NSM outperforms the strongest baseline on both imageability and concreteness tasks, with two exceptions. That goes some way to validate Hypothesis 1, lending credit to the assertion that the embedding space contains structural clues on imageability and concreteness.

In addition to the overall performance, we highlight two important trends. First, the size of the embedding collection matters: regardless of what embedding model is used, larger collections often outperform smaller collections. That is easy to explain: A larger collection represents more granular samples from the underlying distribution, painting a more accurate picture of its structure.

Second, interestingly, the dimensionality of the space plays an important role: predictions correlate more strongly as the dimensionality of the space becomes smaller. The reason for this phenomenon is simple and has to do with the curse of dimensionality: As dimensionality increases, the pairwise distances between points concentrate more heavily around a constant (Bruch, 2024), distorting the

MEASURE	IMAG.	CONC.
NSM$_{\text{ALLMINILM(CC12M)}}$	**0.66**	**0.58**
NSM$_{\text{ALLMINILM(QUORA)}}$	0.40	0.31
NSM$_{\text{ALLMINILM(FIQA)}}$	0.29	0.33
NSM$_{\text{ALLMINILM(MS MARCO)}}$	0.48	0.39

Table 3: Spearman correlation between ground-truth imageability (Imag.) and concreteness (Conc.) ratings and NSM using non-caption text datasets. QUORA (question answering, 523K documents) and FIQA (financial, 57K) are from the BEIR collection (Thakur et al., 2021), and MS MARCO (Nguyen et al., 2016) is a dataset of about 9 million passages designed for retrieval tasks.

picture painted by sample points (i.e., the embedding collection) of the sharpness and flatness of the peaks of the local distribution. As such, an ever larger collection is needed to accurately discover distributional structures of the embedding space.

We conclude our analysis of the results with one last point: as with FREQ, it matters to our method that the text collections from which NSM is calculated are not general natural texts, and instead contain text fragments that describe visual phenomena. That is for good reason: imageability and concreteness are visual-semantic properties. It is therefore reasonable that estimating these properties would require the "visual regions" of the semantic space.

For completeness, we verify the above by repeating our experiments on general-purpose text collections. As Table 3 shows, NSM's correlation degrades when the text collection is not primarily made up of image captions. This agrees with the findings of Wu and Smith (2023) where the highest correlation with imageability ratings was achieved by image captions, not news sentences or poems.

5.3 Analysis

We observed that NSM correlates with ground-truth ratings more strongly than existing unsupervised methods, often with a substantial margin. However, the correlation coefficient itself is less than ideal, implying perhaps that today's unsupervised methods cannot fully replace the manual human judgment process. The reality, however, is more nuanced, as our analysis in this section suggests.

Consider a classification task where we place a word into imageable (concrete) or non-imageable (abstract) categories. As in prior studies (Schulte im Walde and Frassinelli, 2022; Tater et al., 2024), we define ground-truth labels by thresholding the ratings. For instance, words whose
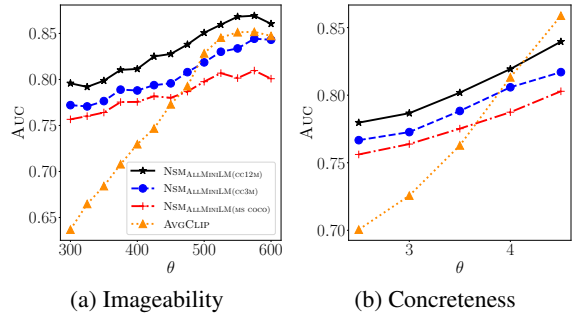

(a) Imageability   (b) Concreteness

Figure 3: Performance in classification tasks. The threshold θ determines class labels: A word whose ground-truth rating is less than θ is non-imageable (abstract). For each θ, we evaluate the effectiveness of select methods as class predictors, quantified using AUC.

imageability ratings are greater than θ would be considered imageable, and those below it non-imageable. NSM scores can therefore be interpreted as class membership probabilities so that for any given θ, we can calculate the Area Under the Receiver Operating Characteristic Curve (AUC).

Figure 3 visualizes AUC for select methods and different values of θ. The plots suggest that NSM serves as a strong predictor of imageability and concreteness in classification tasks. Furthermore, its predictive power is relatively robust to the choice of θ: Even though the correlation coefficients of AVGCLIP and NSM$_{\text{ALLMINILM(MS COCO)}}$ are similar for imageability, the former's predictions are far more sensitive to how we derive class labels.

Let us unpack the predictions by NSM$_{\text{ALLMINILM(CC12M)}}$ further and study the reliability of the NSM scores. We choose a margin, $\omega \in (0, 0.5]$, and exclude words whose NSM score is in the interval $(\omega, 1 - \omega)$. We then compute AUC as before. In this way, we seek to understand if the extreme values offer more predictive power and, thus, are more meaningful.

Figure 4 shows this analysis visually for select margins—margins smaller than 0.35 result in too few samples. Indeed, as the margin becomes smaller, classification accuracy improves. That trend suggests that, if the NSM score for a word is very small (large), the odds of it being non-imageable (imageable) are stronger.

6 Discussion

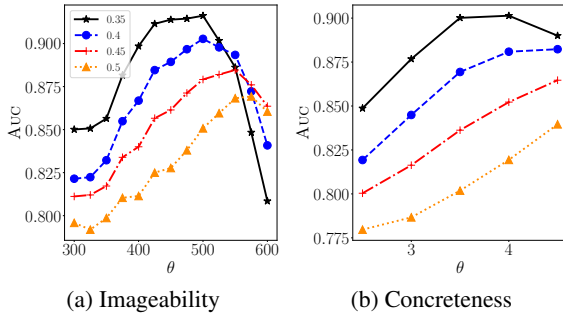We have shown that Algorithm 1 serves as a tool for testing Hypothesis 1, and that its predictions corre-

5645

(a) Imageability   (b) Concreteness

Figure 4: Classification by $\text{NSM}_{\text{ALLMINILM(CC12M)}}$. Words whose NSM scores are in $[0, \omega]$ are negative (i.e., non-imageable or abstract), and $[1 - \omega, 1]$ positive, repeated for $\omega \in \{0.35, 0.4, 0.45, 0.5\}$. There is more confidence in predictions if the NSM score is very small or very large.

late with and are predictive of our visual-semantic psycholinguistic properties of interest. Our measure, NSM, also operates considerably more efficiently than existing unsupervised methods.

Let us expand on efficiency. NSM is *data efficient* as it requires only the text portion of an image-caption dataset, while HML requires a corpus of image-caption pairs. NSM is *computationally efficient* as it operates on a (reusable) collection of text embeddings, performing only ANN search over it. In contrast, the AVGCLIP or COSINESIM methods of Wu and Smith (2023) execute multiple forward passes over an expensive text-to-image model to predict the imageability of a single input text.

Moreover, NSM is *space efficient* as it needs 1) a table mapping the identifier of each point in the collection $\mathcal{X}$ to the identifier of its nearest neighbor, which can be stored in an array of length $|\mathcal{X}|$; and 2) an ANN index, which is typically a fraction of the size of the collection. For example, the size of the ANN index for ALLMINILM(CC12M) is 938MB while the embedding collection itself is 9.5GB in size. GTE-LARGE(CC12M) is 25.4GB in size, while its ANN index is 1GB.

It is worth noting that we use contextualized embeddings of a token, where a token can have diverse contexts and part-of-speech (POS) tags in an image-caption dataset. In contrast, context and POS tags were not indicated when collecting ratings from surveyed humans. Nevertheless our method estimates ratings effectively, suggesting perhaps that a large image-caption dataset captures the everyday context of a word to some extent. Though we do not offer an answer in this work, that raises an interest-

ing psycholinguistic question: do humans consider diverse contexts when rating imageability?

Lastly, we note that our method works in the embedding space and as such is agnostic to the characteristics of the input prior to embedding. That entails that Algorithm 1 can in theory be applied to a sentence, after it is projected into the embedding space. We are, however, unable to demonstrate this extension for two reasons. First, to the best of our knowledge, no dataset exists that provides the imageability rating of sentences. Second, how a sentence is perceived is more complex as it involves compositionality. We believe an extensive qualitative study involving surveys for sentence-level imageability makes for an exciting future direction.

7 Conclusion

We hypothesized that text collections provide sufficient clues on visual-semantic psycholinguistic properties of words. In particular, we claimed that the range of contexts in which a word appears correlates with imageability and concreteness. We then formalized that claim in distributional terms in Hypothesis 1, stating that the structure of the embedding space of a sufficiently large text collection encodes imageability and concreteness.

We subsequently developed a method to quantify the "sharpness" of peaks in the distribution—the Neighborhood Stability Measure. A perfectly-stable neighborhood, according to Definition 1, is one where the nearest neighbor of every point belongs to the same neighborhood. More stable neighborhoods correspond with sharper peaks. Empirically, NSM correlates more strongly with imageability and concreteness ratings than all other unsupervised baselines. Additionally, NSM is a strong predictor of class membership in classification tasks. Finally, NSM is comparatively efficient, and addresses the shortcomings of prior methods.

Even though we have demonstrated that single-modality proves adequate, we ask whether bringing visual features into a joint embedding space can enrich the distributional structure, and lead to better predictions. We have also left unanswered the question of *why* NSM performs better on image-caption datasets, and *whether* its performance can be improved on general natural text collections. Finally, while we claim that our method is applicable to text fragments (as opposed to single words), we have not tested that claim. We leave an exploration of these directions as future work.

Limitations

1. Our experiments were done for English words only. We cannot definitely conclude whether similar observations can be made when applying our method to languages that are morphologically different.

2. We have shown in our experiments that our method relies on text collections that describe visual phenomena such as image captions. Such datasets are not as prevalent as general-purpose natural text collections, especially in languages other than English. As such, extending our findings to other languages can face practical challenges.

Potential risks

To obtain AVGCLIP and COSINESIM, we replicated the experiments by Wu and Smith (2023) which utilize a text-to-image model. As Wu and Smith (2023) claimed, DALL•E mini can potentially generate suggestive images even when given a text that does not contain suggestive content. While we did not notice any images generated that were suggestive, one should be cautious about the output of a text-to-image model, especially if they intend to release the images.

Acknowledgments

References

Sebastian Bruch. 2024. *Foundations of Vector Retrieval*. Springer Nature Switzerland.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911.

Soravit Changpinyo, Piyush Kumar Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3557–3567.

Jean Charbonnier and Christian Wartena. 2019. Predicting word concreteness and imagery. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 176–187, Gothenburg, Sweden.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server.

Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Nick Ellis. 1991. Chapter 21 word meaning and the links between the verbal system and modalities of perception and imagery or in verbal memory the eyes see vividly, but ears only faintly hear, fingers barely feel and the nose doesn't know. In Robert H. Logie and Michel Denis, editors, *Mental Images in Human Cognition*, volume 80 of *Advances in Psychology*, pages 313–329. North-Holland.

W. N. Francis and H. Kucera. 1979. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US.

Diego Frassinelli, Daniela Naumann, Jason Utt, and Sabine Schulte m Walde. 2017. Contextual characteristics of concrete and abstract words. In *Proceedings of the 12th International Conference on Computational Semantics*.

Ken J Gilhooly and Robert H Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior research methods & instrumentation*, 12(4):395–427.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the visual concreteness of words and topics in multimodal datasets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2194–2205, New Orleans, Louisiana. Association for Computational Linguistics.

Marc A. Kastner, Ichiro Ide, Frank Nack, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, and Hiroshi Murase. 2020. Estimating the imageability of words by mining visual characteristics from crawled image data. *Multimedia Tools and Applications*, 79:18167–18199.

Alessandro Lenci, Gianluca E Lebani, and Lucia C Passaro. 2018. The emotions of abstract words: A distributional semantic analysis. *Topics in Cognitive Science*, 10(3):550–572.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Ting Liu, Kit Cho, G. Aaron Broadwell, Samira Shaikh, Tomek Strzalkowski, John Lien, Sarah Taylor, Laurie Feldman, Boris Yamrom, Nick Webb, Umit Boz, Ignacio Cases, and Ching-sheng Lin. 2014. Automatic expansion of the MRC psycholinguistic database imageability ratings. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 2800–2805, Reykjavik, Iceland.

Gonzalo Martínez, Juan Diego Molero, Sandra González, Javier Conde, Marc Brysbaert, and Pedro Reviriego. 2024. Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, 57(1):5.

Daniela Naumann, Diego Frassinelli, and Sabine Schulte im Walde. 2018. Quantitative semantic variation in the contexts of concrete and abstract words. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, pages 76–85, New Orleans, Louisiana. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset.

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

John T. E. Richardson. 1976. Imageability and concreteness. *Bulletin of the psychonomic society*, 7:429–431.

Jocelyn Schock, Michael J Cortese, and Maya M Khanna. 2012. Imageability estimates for 3,000 disyllabic words. *Behavior Research Methods*, 44:374–379.

Sabine Schulte im Walde and Diego Frassinelli. 2022. Distributional measures of semantic abstraction. *Frontiers in Artificial Intelligence*, 4.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Tarun Tater, Sabine Schulte Im Walde, and Diego Frassinelli. 2024. Unveiling the mystery of visual attributes of concrete and abstract concepts: Variability, nearest neighbors, and challenging categories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21581–21597, Miami, Florida, USA. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Michael P Toglia and William F Battig. 1978. *Handbook of semantic word norms.* Lawrence Erlbaum.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Michael Wilson. 1988. MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20:6–10.

Si Wu and David Smith. 2023. Composition and deformance: Measuring imageability with a text-to-image model. In *Proceedings of the 5th Workshop on Narrative Understanding*, pages 106–117, Toronto, Canada. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval.

Qingsheng Zhu, Ji Feng, and Jinlong Huang. 2016. Natural neighbor: A self-adaptive neighborhood method without parameter k. *Pattern Recognition Letters*, 80:30–36.

A Effect of radius

The Neighborhood Stability Measure (NSM) has a single hyperparameter, k, which represents the size of the neighborhood for which we calculate the stability measure. In our main experiments, we used a validation set to tune k by sweeping the interval $[64, 4096]$ in increments of 64. In this

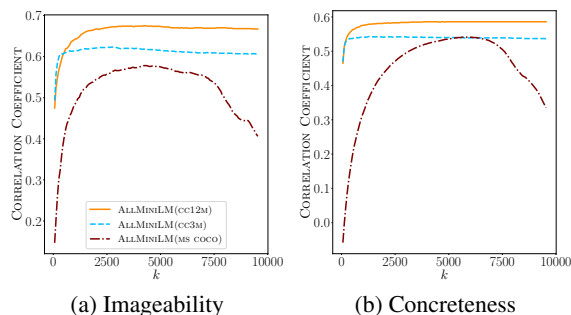(a) Imageability

(b) Concreteness

Figure 5: Effect of the NSM radius, k, on the correlation with imageability and concreteness, for select embedding collections. A value between 2^{10} and 2^{12} yields peak performance for reasonably large collections.

supplementary section, we examine the effect k has on the correlation between NSM and imageability, as well as NSM and concreteness.

Figure 5 illustrates the Spearman's correlation coefficient as a function of k. In other words, for each value of k, we compute the NSM scores for all words in the imageability and concreteness ratings datasets, and compute the correlation of the scores with ground-truth ratings. Note that, while Figure 5 visualizes this result for ALLMINILM embeddings, other embedding collections exhibit a similar trend.

As the plots show, a small value of k leads to poor correlation. The reason behind this is that, NSM takes a value in $\{0, \frac{1}{k}, \frac{2}{k}, \ldots, 1\}$. When k is too small, the score set becomes too coarse. On the other hand, the imageability and concreteness ratings are fine-grained. This discrepancy between the granularity of NSM score set and the ratings leads to poor correlation, as NSM is unable to distinguish between ratings that are close to each other.

As k becomes larger, correlation stabilizes. That trend continues until k is too large relative to the size of the collection. Interestingly, independent of the size of the dataset, $k \in [2^{10}, 2^{12}]$ appears to yield peak correlation. A k in that range can thus serve as a reasonable default value where a validation set is unavailable.