

Semantic Composition in Visually Grounded Language Models

by

Rohan Shrirang Pandey

Undergraduate Thesis submitted to the
Language Technologies Institute
in partial fulfillment of the requirements for
Honors from the School of Computer Science

Supervised by:
Dr. Louis-Philippe Morency
Paul Pu Liang

Abstract

What is sentence meaning and its ideal representation? Much of the expressive power of human language derives from semantic composition, the mind’s ability to represent meaning hierarchically & relationally over constituents. At the same time, much sentential meaning is outside the text and requires grounding in sensory, motor, and experiential modalities to be adequately learned. Although large language models display considerable compositional ability, recent work shows that visually-grounded language models drastically fail to represent compositional structure. In this thesis, we explore whether & how models compose visually grounded semantics, and how we might improve their ability to do so.

Specifically, we introduce 1) WinogroundVQA, a new compositional visual question answering benchmark, 2) Syntactic Neural Module Distillation, a measure of compositional ability in sentence embedding models, 3) Causal Tracing for Image Captioning Models to locate neural representations vital for vision-language composition, 4) Syntactic MeanPool to inject a compositional inductive bias into sentence embeddings, and 5) Cross-modal Attention Congruence Regularization, a self-supervised objective function for vision-language relation alignment. We close by discussing connections of our work to neuroscience, psycholinguistics, formal semantics, and philosophy.

Acknowledgements

I would like to begin by thanking my thesis advisors, LP Morency and Paul Liang, for the continual guidance on the several projects this thesis encompasses. The current state of my research would also not be possible without my previous advisors Graham Neubig, Adrian Brasoveanu, Marilyn Walker, Tim Mullen, and Narges Norouzi. I'd also like to thank my collaborators Rulin Shao, Uri Alon, Frank Xu, Kevin Bowden, as well as everyone I've gone to for advice, Tristan Thrush, Chunyuan Li, Tom McCoy, Emmy Liu, Brad Mahon, Brian MacWhinney, Mike Tarr, Chris Potts, and Paul Smolensky.

I'd like to thank my friends Kamil Kisielewicz, Shiv Rustagi, Aryaman Arora, Miguel Tenant de la Tour, Max Alfano-Smith, Adam Farris, Justus Mattern, Bradley Hsu, Roanak Baviskar, Avik Jain, Steven Feng, and Abhinav Tumu for being critical in my journey as a researcher and builder. Also, thank you to the brothers of $\Sigma\Phi\Theta$ PA- Θ for making my time at CMU enjoyable enough to have the energy to write this thesis. I'm also grateful to all the amazing people and teams I worked with at Microsoft (Man Fong), Wordcab (Aleks Smechov), Vizerto (Sanjay Shitole), Bunch (John Wehr), Intheon (Christian Kothe), and SapientX (David Colleen) over the course of the last 4 years.

To my father I am thankful for the pragmatic, solution-seeking mindset I was raised with and to my mother I am thankful for fostering my curiosity and connections to the Indian intellectual tradition, without which much of my motivation on these questions would be absent. In that vein, I thank my intellectual forefathers in whose writings I found meaning in semantics: Bharadvāja Bārhaspatya, Pāṇini, Kātyāyana, Jaimini, Bhartṛhari, Kumārila Bhaṭṭa, Vācaspati Miśra, Gaṅgeśa Upādhyāya, and Vimalakṛṣṇa Matilal.

Finally, I thank the reviewers of all work included in this thesis for their insightful and valuable comments.

वैश्वानरायं मुतिर्नव्यसी शुचिः
सोमं इव पवते चारुरुग्नये ॥

For That of all men, a new thought clarifies
pleasantly like bright *sóma* for *Agni*.

Table of Contents

1	Introduction	1
1.1	Semantic Composition	1
1.2	Visual Grounding	2
1.3	Observations	2
1.4	Vision-Language Deep Learning	3
1.4.1	Problem	4
1.5	Outline	4
2	Can models compose visually grounded semantics?	6
2.1	Past Work	6
2.2	WinogroundVQA Benchmark	7
2.2.1	Results	7
3	How may models compose visually grounded semantics?	8
3.1	Past Work	8
3.2	Syntactic Neural Module Distillation to Probe Compositionalty	9
3.2.1	Methods	9
3.2.2	Experiments	11
3.2.3	Results	11
3.2.4	Conclusion	14
3.3	Syntax in Attention Flow	14
3.4	Locating Grounded Composition Circuits with Causal Tracing	16

4 Enabling models to compose visually grounded semantics	17
4.1 Past Work	17
4.2 Cross-modal Attention Congruence Regularization	19
4.2.1 Method	20
4.2.2 Results	25
4.2.3 Analysis	26
4.2.4 Conclusion	29
4.3 Syntactic MeanPool for Sentence Embedding	29
5 Conclusion	32
5.1 Multimodal Brain	32
5.2 Vision-Language Psychology	32
5.3 Multimodal Semantics	33
5.3.1 Indian Philosophy	33
References	35
APPENDICES	35

Chapter 1

Introduction

The principle of compositionality and problem of symbol grounding have a long history in Cognitive Science & Analytic Philosophy, tracing back to Gottlob Frege in the late 19th century; an even longer history may be found among the Nyāya & Mīmāṃsā schools c. 7-12th century [53]. For much of the 20th century, logicians [78], formal semanticists [58], and theoretical syntacticians [18] explored how symbolic rule systems could describe the compositionality of human language without trying to describe how the symbols might map onto the multimodal world our language interacts with.

1.1 Semantic Composition

At its core, compositionality is about the regularity of a system. In a fully compositional system, the meaning of an expression is perfectly predictable by its constituents and their relational structure [67]. Compositionality is a useful property of language because it enables high productivity and generalization while requiring less representational capacity [10].

Although strictly compositionality is a measurement of a linguistic system, we often discuss a model’s compositional ability or understanding. By this, we mean a model builds representations that capture the underlying compositional structure in its input. Neither humans nor language models are explicitly trained on syntax trees, but humans implicitly learn syntax which enables them to understand compositional distinctions in semantics.

Compositional ability is vital to forming structured representations, symbolic reasoning, and various high-level system 2 cognitive capacities [11]. Ultimately, semantics is about interpretation—assigning a meaning to an expression abstracted away from the specifics of a language. Interpretation is what allows us to bridge the gap from one language to another language or modality. Traditionally, this meant interpreting utterances like “the car is red” as true if and only if the car is red, locking us into a self-referential system with symbols all the way down.

1.2 Visual Grounding

Although there is certainly utility in approaches that elegantly lift the compositional structure of language to a more formalized level, one cannot deny the dissatisfaction of purely symbolic semantic interpretations. How can we know if the car is actually red when we can't look up from the page? How can we ground our semantic interpretation in something beyond symbols, as humans do through interaction with the world?

The problem of symbol grounding is quite broad, with the earliest Western ideas exploring the relationship between sign & meaning [26] and sense & reference [30]. Simply put, a distinction must be made between the surface level linguistic expression, the meaning it conveys, and the entities in the world that it refers to. While considerable progress has been made over the past century formalizing the relation between the first 2 (syntax & semantics) [59], relatively little progress had been made on grounding semantics to sensory representations of the world until the deep learning era.

Prior to deep learning, it was intractable to compute high quality image representations, relegating all progress in semantics to more symbolic domains. In the deep learning era, the shared semantic framework of vectorial embeddings enables interaction of language meaning representations with image representations. Grounding a model's semantics in sensory modalities will be critical for developing robots that can see, hear, and feel the real world [12].

1.3 Observations

In order to better understand the problem we face—visually grounding the compositional semantics of language—let us begin with a Wittgensteinian (or perhaps sūtranic), first principles approach to observing some general truths about the vision-language duality.

1. Fundamentally, both visual (2D) and audio-linguistic (1D) stimuli are unstructured.
2. The ability to identify constituents (objects or tokens) of a stimulus (scene or sentence) is learned with inductive biases.

2.1 Contiguity

- 2.11 An object usually occupies a contiguous region of space.
- 2.12 A token usually occupies a contiguous span of time.

2.2 Repetition

- 2.21 Seeing a cluster of pixels corresponding to a rabbit in multiple scenes.
- 2.22 Hearing a sequence of syllables corresponding to “bunny” in multiple sentences.

3. A constituent is usually an instance of a category (class or word).

- 3.1 A set of objects from different scenes can visually vary considerably but still instantiate the same class.

3.11 A leaf may appear with different colors, occlusions, orientations depending on context.

3.2 A set of tokens from different sentences can auditorily vary considerably but still instantiate the same word

3.21 "Leaf" may be heard with different phonetics, prosody, or morphology depending on context.

4. Every visual class corresponds to a word.

5. Some words correspond to a visual class.

6. Constituents form relations with each other in a stimulus.

6.1 Objects are organized in visual space.

6.11 person ON bed; blanket ON person; eyelids IN person; top-eyelids TOUCH bottom-eyelids.

6.2 Tokens are ordered in time.

6.21 "The man sleeps".

7. Constituents in a stimulus are usually modified for relations to form.

7.1 A blanket ON a sleeping person will not be neatly folded or completely flat, it will broadly contour their body.

7.2 "sleep" is modified to "sleeps" when related to a single subject.

8. There are restrictions on valid stimuli.

8.1 Two objects in a scene may not occupy the same spatial position.

8.2 Two tokens in a sentence may not fill the same argument of a relation.

9. A scene may be described by different sentences.

10. A sentence may be interpreted as different scenes.

1.4 Vision-Language Deep Learning

Vision-language deep learning research [46] synthesizes advances in computer vision and natural language processing to build models capable of solving multimodal tasks like image-text matching, image2text captioning, and text2image generation. There are a number of methods for learning joint representations of vision and language, but most leverage the transformer architecture [90].

Encoder models generally either use a dual unimodal encoder architecture [70], a single cross-modal encoder [17], or a combination [80]. Image-conditioned text generation models often consist of an image encoder that feeds into a causal transformer decoder [43]. Text-based image generation models are quite varied but often condition an image diffusion model on the output of a text encoder [73].

1.4.1 Problem

While several results in vision-language deep learning are quite impressive, recent work has shown that models consistently run into issues capturing compositional structure. Winoground [88] is a simple vision-language compositionality task that tests a VLM’s ability to match syntactic permutations of a caption to their corresponding images, finding that all recent and state-of-the-art VLMs perform below chance levels on the Winoground evaluation dataset. Contemporaneously, Milewski et al. [57] probe for structural knowledge in VLM’s, finding that they encode significantly less linguistic syntax than LM’s and virtually no visual structure. Tejankar et al. [86] find that VLM’s perform no better than a bag-of-words at an image-text matching task.

In order to improve these abilities, Nikolaus et al. [64] find that VLM’s trained using fine-grained vision-language objectives display better syntactic understanding. Yuksekgonul et al. [103] show that training VLM’s with order-permuted hard negatives improves compositional performance. We seek to contribute to this pool of work that addresses issues of fine-grained, compositional, or syntactic understanding in visually grounded language models.

1.5 Outline

In this thesis, we explore whether & how models compose visually grounded semantics, and how we might improve their ability to do so.

First, we propose WinogroundVQA (Sec. 2.2¹) to test the ability of generative vision-language foundation models to capture compositional distinctions, building on previous work that only tested image-text matching models.

Next, we propose methods for understanding how models perform semantic composition. Syntactic Neural Module Distillation (Sec. 3.2) enables us to test whether a sentence’s syntax tree is a strong causal model of its embedding’s composition. We build a simple visualization tool (Sec. 3.3) to observe if a transformer’s attention activation flow, seen as its causal structure, aligns with a sentence’s syntax, a symbolic model of its composition. We then apply this intuition more rigorously to adapt a mechanistic interpretability approach, causal tracing, to locate neural representations important for composition in image captioning models (Sec. 3.4²).

Having better understood composition in visually grounded models, we now explore approaches for improving their ability. We propose CACR (Sec. 4.2), a self-supervised objective that encourages models to represent unimodal relations in a way that better enables cross-modal relation alignment, improving on the current Winoground state-of-the-art. To inject a compositional inductive bias into unimodal embeddings, we propose the Syntactic MeanPool (Sec. 4.3) which seeks to improve CLIP’s image-text match score on Winoground.

¹Results in progress

²Experiment implementation in progress

In closing, we draw connections to the cognitive sciences (Sec. 5). We explore how some of the behaviors we observe in models may have parallels in the brain since multi-modal semantic representations are localized in the left anterior temporal lobe. We discuss vision-language compositional ability from a psycholinguistic perspective by testing a time-constrained version of Winoground on humans. We finally conjecture a multimodal semantics framework to formalize compositional behavior in vision-language embedding space that draws inspiration from homotopy type theory.

Chapter 2

Can models compose visually grounded semantics?

2.1 Past Work

Several datasets to test vision-language models’ compositional ability have recently cropped up [87, 50, 104], but the question has been studied since early in the deep learning era [85, 6]. We’re particularly interested in Winoground [87] and how we can leverage its general approach to benchmarking visually grounded compositional ability. The Winoground dataset consists of pairs of image-caption pairs, where the captions share the same words but in a slightly different order—for example, “mug in grass” v.s. “grass in mug”. If a vision-language encoder model produces a higher image-text match score between the correct pairs than the incorrect pairs, then it understands the compositional distinctions.

Just as unimodal NLP research has shifted from encoder models like BERT [27] to generative models like GPT-3 [14] over the past few years, vision-language research appears to be shifting now from joint encoder models like UNITER [17] and FLAVA [80] to image-conditioned text generation models like BLIP2 [44] and GPT-4 [65]. Unfortunately, it is not straightforward to evaluate these models on Winoground. One could extract the conditional probability of a caption given an image, but these scores are often unavailable due to large models being hosted on private servers with restricted access.

Furthermore, recent work [28] shows that 19.5% of Winoground examples require complex reasoning over the image and text. Capturing the complexities of a caption’s reasoning in a single embedding and representing an image in such a way that it can be reasoned over using a simple cosine similarity seems difficult and perhaps topologically implausible, making it quite unlikely that two stream models trained using a cross-modal contrastive loss would be able to solve these examples.

Empirically, this is partially backed by FLAVA’s [80] performance on Winoground, which is only 9.00 when measured contrastively but 14.25 when measured using image-text matching score. In this latter setting, the model is given the ability to attend cross-modally between vision and language, potentially enabling a simple form of reasoning in

the forward pass. By switching to a generative setting, we enable the model to perform deeper reasoning, opening the possibility for visually-grounded chain-of-thought prompting [94] on our task.

Another concern raised by Diwan et al. [28] is that 12.5% of Winoground examples contain unusual text which is out of distribution from standard image captioning datasets or even borderline ungrammatical.

2.2 WinogroundVQA Benchmark

To address these issues, we propose WinogroundVQA, the Winoground dataset rephrased as a visual question answering problem. Instead of presenting an image-text pair to an encoder model and calculating the match score, we provide an image along with a question like “Is the mug in the grass?” and score the model based on whether it answers “yes” or “no”. This approach solves a number of the previously mentioned issues:

1. It enables us to test the Winoground ability of visually grounded large language models that don’t produce a match score and don’t expose caption probabilities.
2. It places Winoground in a more cognitively plausible setting where reasoning is handled through language generation rather than in a single embedding.
3. It cleans up out of distribution captions by converting them into questions that are phrased less unusually.

Methodologically, we convert Winoground captions to our VQA form with a simple prompt that we query GPT-3.5 with: “Here is a description of a picture: {caption}. Rewrite the description as a yes-or-no question but change as little of the phrasing as possible.”

2.2.1 Results

Model	Text	Image	Group
BLIP	2.0	2.25	1.5
MiniGPT-4	6.5	6.25	5.75

Table 2.1: Performance of some visually-grounded language models on WinogroundVQA

Chapter 3

How may models compose visually grounded semantics?

3.1 Past Work

The principle of semantic compositionality suggests that the meaning of a sentence should derive from its subconstituents in a regular, structured fashion [60]. In recent years, transformers [90] have become effective at producing sentential meaning representations useful for downstream tasks such as Natural Language Inference, Image-Text Matching, and Document Classification [22, 71]. However, it has famously been conjectured that "You can't cram the meaning of a whole %&!\$# sentence into a single \$&!#* vector" [61]. Since recent models do appear to capture sentence meaning effectively, one wonders how they compose arbitrarily many word meanings together such that their relational structure is captured in a single, fixed-dimensional sentence embedding.

Much work has sought to probe these models for syntax representations and their causal relevance to embedding output. Conneau et al. [23] train linear probes to determine if models encode syntactic features like tree distance and depth [41, 35]. One line seeks out direct mappings between neural representations and tree structures [55, 19, 37, 83, 62]. Other work raises methodological issues with probing [29, 108] such as choice of formalism [42] and semantic entanglement [54]. Ravichander et al. [74] raise the possibility that probing may identify causally un-used features; Tucker et al. [89] partly address this concern to show that some syntactic features are causally relevant. Another line of work explores the geometry of semantic representations [75, 34] and the linearity [9] of syntactic analogies [107].

Rather than directly analyze sentence embedding models, Neural Module Nets [3] seek to improve compositionality by modularizing semantic functions. We see this effort as ultimately similar to probing for structural representations since the former explores whether explicit structure improves performance and the latter explores whether performant models implicitly learn structure. Geiger et al. [31] discovers logical tree causal structures in BERT and Wu et al. [96] then guides model distillation using this structure.

3.2 Syntactic Neural Module Distillation to Probe Compositionality

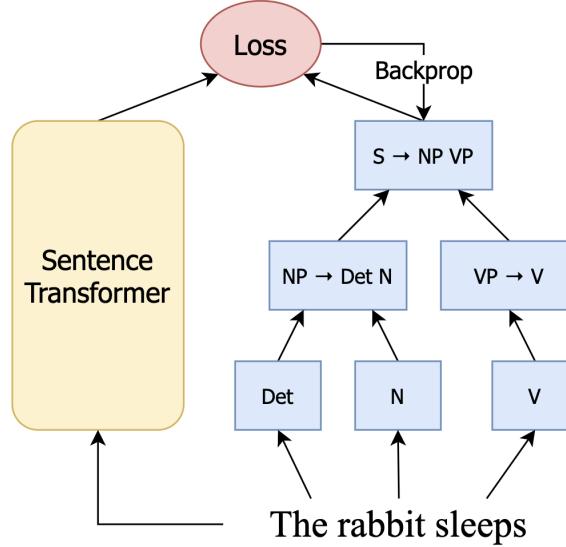


Figure 3.1: Distilling a transformer to a neural module net structured by the sentence’s syntax

Our work builds on these findings by strictly taking syntax as the causal structure of sentential semantics and linearity as the geometry of syntax-guided composition; we conduct experiments to test the distillability of transformer-based sentence embedding models to a Syntactic Neural Module Net (SynNaMoN), an architecture we introduce that implements these two priors. The extent to which a model can be distilled to a SynNaMoN tells us about its internal syntax representations & compositional ability.

3.2.1 Methods

Unlike prior work [4, 20], SynNaMoN modules don’t approximate high-level objectives like ‘Find’ or ‘Count’ but rather correspond to specific syntactic rules like ‘ $S \rightarrow NP VP$ ’ and ‘ $NP \rightarrow DT JJ NN$ ’. Each module receives an input of dimensionality $(1, N * D)$ where N is the number of constituents on the syntax rule’s right-hand side, and D is the dimensionality of the embedding space (768)—in other words, the input embeddings are concatenated. Though computationally more expensive, concatenation enables the module to learn an arbitrary function over the inputs rather than restricting it to a function over their sum or mean; this enables the module to converge on its ideal composition function which is likely not invariant under summation or averaging. Finally, though our implementation of SynNaMoN includes ‘part-of-speech’ modules at the bottom of the parse tree, one could conceivably remove this bottom layer with the hypothesis that the word embeddings already capture part-of-speech information.

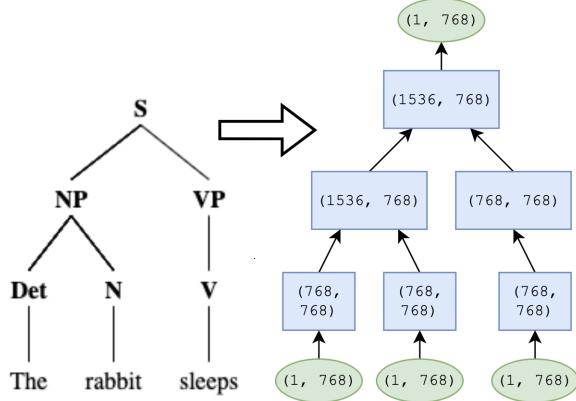


Figure 3.2: Constructing a sentence’s SynNaMoN from its syntax tree; module input and output dimensionalities labeled on the right.

Internal Module Architecture

To explore the geometry of semantic composition under syntax, we implement 3 module architectures: a linear layer (**Linear**), a linear layer + a ReLU activation (**Nonlin**), and a linear layer + ReLU + another linear layer (**Double**). We explore these 3 architectures to see whether syntax is enough of an inductive bias to linearly approximate sentence embeddings, or if adding non-linearities and additional layers considerably improves performance. The extent to which adding parameters improves our approximation of the teacher model beyond the syntactic structure alone could reveal how much isn’t captured by this inductive bias.

Linguistic Formalism

We choose to use the Transformational Grammar presented by Penn Treebank [51], but in principle any Constituency Grammar could be easily used with SynNaMoN, and Dependency Grammars can be adapted with some effort. Since prior work has shown how the choice of linguistic formalism can significantly influence probing results [42], we float the possibility of such an effect being at play in this work as well. If a student SynNaMoN fails to capture much of the teacher embedding model, perhaps it isn’t because of the teacher’s non-compositional causal structure, but rather because the formalism used to structure the SynNaMoN is inadequate. Indeed, recent state-of-the-art neural approaches to syntax parsing have learned grammatical tagsets that often differ starkly from human-produced syntactic theories [40]. We leave these problems to future work which may explore the exciting possibility that certain linguistic formalisms (perhaps even semantic rather than syntactic) are better proxies for a model’s compositional structure than others.

3.2.2 Experiments

Our main experiment runs 5 sentence embedding models (BERT-base [27], MP-Net [82], GTR-T5-base, GTR-T5-large, and GTR-T5-xl [63]) on 3 SynNaMoNs with differing internal architectures (Linear, Nonlin, Double; see Sec. 3.2.1). For BERT-base, we extract input word embeddings for each token and use the CLS token as the sentence embedding as is common practice. For the other 4 models, we encode each token alone to serve as its embedding and use the output as the sentence embedding. When words are encoded as more than 1 token, we compute the mean across the subtokens to serve as its word embedding.

In order to heuristically select a learning rate, 5 training runs were conducted with SynNaMoNs optimizing for BERT-base, and learning rate manually set at increments between 10^{-5} and 10^{-3} . We finally chose a rate of 5×10^{-5} , but recognize from results that optimal learning rate will likely vary by teacher model & SynNaMoN internal architecture. Analysis would best be reported on the optimal scores achieved by a SynNaMoN after hyperparameter tuning, but due to compute restrictions (1 NVIDIA K80 GPU with 12GB of RAM), this was unfeasible.

Additionally, due to the number of modules (originally 900, each with 1M parameters on average), we encountered frequent out-of-memory errors both on CPU & GPU. Since each module corresponds to a syntax rule and is initialized upon encountering the rule in the dataset, we constrained our data to minimize the number of modules needed.

Specifically, we first constrained our trees to those of height 4 & 5 ($n=16492$) in PTB, and then further constrained the trees to those that use a subset of the 300 most common production rules among them. This resulted in 1494 trees, from which we generated a train-validation split of 1250-244. Furthermore, we ensured that all the productions present in trees of the validation split were also included amongst trees in the training split. All this finally resulted in 273 production rules present in our dataset, and the instantiation of 273 modules.

3.2.3 Results

In Tab. 3.1, we present scores for all 5 sentence embedding models across the 3 SynNaMon architectures. We compute the average MSE between sentence embeddings in the complete dataset for each model and divide each model’s MSE loss by this mean distance to normalize results. The normalized scores we present may intuitively be seen as the portion of variance in a model’s sentence embeddings that a SynNaMoN fails to explain. From a probing perspective, the lower a model’s score, the more it can be causally approximated by composition along syntactic lines.

First, notice that GTR-T5-xl outperforms all the other models across all the SynNaMoN architectures. This seems to confirm our expectation that larger models should produce more compositional sentence embeddings. However, GTR-T5-xl only marginally outperforms other sizes of GTR-T5 (except on Nonlin, for which it does far better), suggesting that size actually isn’t a significant factor in compositionality. The lower performance

Sent. Emb. Model	Linear	Nonlin	Double
BERT-base-CLS	.765	4.17	.625
MP-Net-base	.606	.963	.538
GTR-T5-base	.541	.844	.499
GTR-T5-large	.550	.898	.502
GTR-T5-xl	.536	.775	.498

Table 3.1: Best validation MSE loss of sentence embedding models on each SynNaMoN probe, normalized by chance-level MSE between embeddings

of GTR-T5-large further corroborates this, but considering its anomalously lower average embedding MSE, the issue requires more work. The fact that GTR-T5 models all display high compositionality despite variance in size suggests something about their architecture or training approach is important—perhaps the representational bottleneck.

All 3 GTR-T5 models perform better than MP-Net, which in turn outperforms BERT CLS. This first fact is slightly surprising considering that on standard sentence representation tasks [76], MP-Net (63.30) marginally outperforms all GTR-T5 (base: 59.40, large: 62.38, xl: 62.88) models. Evaluation of these sentence embedding models on large-scale, human-interpretable compositionality tasks may reveal that GTR-T5 does indeed produce better compositional representations than MP-Net. Although BERT’s CLS token embedding is widely used for sentence representation, these results show that it fails to capture nearly as much compositional information as more targeted sentence embedding models.

Next, observe that although distilling to a Double SynNaMoN is intuitively easier than to a Linear SynNaMoN due to increased parameterization, there aren’t always major improvements in distillability. It is possible that the geometric expressivity of the Double SynNaMoN will kick in with scaling of training data, but we hypothesize that this Double score will still approach a limit for all sentence embedding models. This is because syntax only describes a subset of sentence meaning, and the strictness of SynNaMoN’s structure prevents this non-compositional component from being learned.

For example, a strictly syntactic compositional interpretation of ”village on the river”, would represent the village as being literally on top of the river since this is the semantic geometry learned for syntactic structures of the form ”NP on NP”. A SynNaMoN that includes non-linearities may better learn the geometry of this non-literal ”on” relation, but a transformer model would best learn to handle non-compositional phrases due to its lack of strict syntactic constraints. Our broader takeaway from comparing Linear & Double scores is that much composition along syntactic lines is linear, and non-linearities in transformers primarily serve a purpose other than syntax-guided composition—perhaps in handling non-compositional phrases.

On a less theoretical note, we observe that our learning curves for Linear SynNaMoN on GTR-T5 (Fig. 3.3) are clearly overfitted due to fixing hyperparameters as mentioned in Sec. 3.2.2. We remediate this issue in Tab. 3.1 by reporting the best scores (minimum across epochs) for each learning curve. Since we want to construct the best possible SynNaMoN for a transformer model (as this most accurately reveals the transformer’s distillable

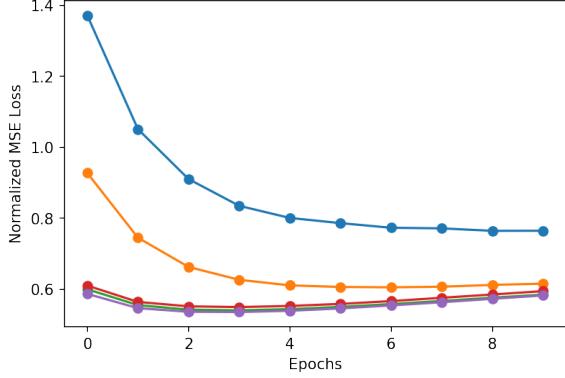


Figure 3.3: Normalized validation learning curves for Linear SynNaMoN on sentence embedding models (blue: BERT, orange: MP-Net, red: GTR-T5-large, green: GTR-T5-base, purple: GTR-T5-xl)

compositional ability), scores could be slightly improved with further hyperparameter tuning.

Analysis

Finally, we explore a single module to determine whether its compositional geometry meets intuitive notions of semantic generalization. Due to methodological difficulties with assessing a single module extracted from our end-to-end training paradigm, we train a Linear module for ‘NP → Det N’ on its own. Determiner-noun composition intuitively lies on a spectrum with adjective-noun composition on the other end and quantifier-noun composition in between. While quantifiers like ‘some’ and ‘all’ seem more like determiners, other quantifiers like ‘several’ and ‘twelve’ appear more comparable to adjectives like ‘swarming’ and ‘grouped’. Intuitively then, we should expect the geometry of quantifier-noun composition to be intermediate to determiners and adjectives.

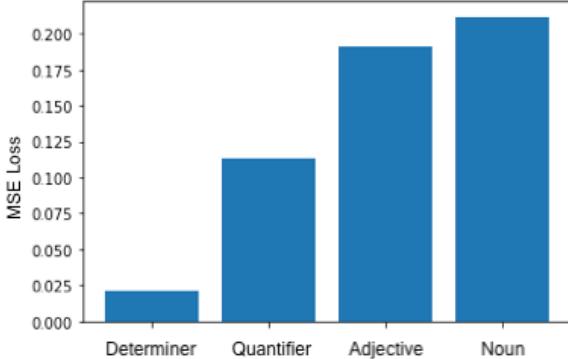


Figure 3.4: Generalization ability of Determiner Phrase module’s linear geometry varies by part-of-speech

And this behavior is precisely what we find in our ‘NP → Det N’ module. Since it’s trained on determiners, it obviously has the lowest MSE for this part-of-speech; we

include noun-noun pairs (e.g. ‘tree cow’) as a control. As seen in Fig. 3.4, the module generalizes to quantifiers intermediately to determiners and adjectives. This demonstrates how SynNaMoN modules may enable interesting analyses of the compositional geometry of syntactic operations in sentence embedding models.

3.2.4 Conclusion

The human ability to apprehend the unitary meaning of a sentence corresponds to a neural model’s ability to construct compositional sentence embeddings. In this work, we introduced Syntactic Neural Module Nets and used it in a distillation approach to assess how well syntax explains the sentential semantics computed by a transformer model. We showed that some models are more compositional by this metric, syntax-guided composition is largely linear, and modules learn composition functions that correspond to our semantic intuition.

Future work could explore this approach’s alignment with other compositionality metrics and the non-compositional semantics left uncaptured by SynNaMoNs. We are also interested in how SynNaMoNs of different linguistic formalisms vary in distillability, as well as other potential use cases of SynNaMoNs beyond probing.

3.3 Syntax in Attention Flow

In the above work, we saw how distillation allows us to test whether a sentence embedding model’s compositional behavior can be described by a syntactic neural module net. However, this does not necessarily prove that the model itself internally implements such syntax-guided compositional structures.

To examine whether this is actually the case, we may visualize the attention flow [1] of a sentence embedding model to explore whether it follows patterns predicted by syntax. From an intuitive perspective, the query-key softmax term of self-attention ($\sigma(QK^\top)V$) modulates the extent to which other words are composed into one word’s representation. Thus, from a syntactic perspective, we should expect that words compose within their local clausal boundaries first before composing with distant words. Furthermore, dependencies should contribute more towards the representation of a clausal head than vice versa.

In this work, we do not quantify either of these hypotheses but simply build a visualization tool to explore how attention flow may align with syntax.

In Fig. 3.5, we can observe the attention flow of two sentences through BERT [27]. In each column, the self-attention activations averaged across all attention heads (usually visualized as a matrix for each layer) are plotted such that the direction of lines from bottom to the top indicate the words that compose into other words’ representations. One may thus follow a word’s attention flow from the top down to observe the other words that contribute to its representation and in what order they compose.

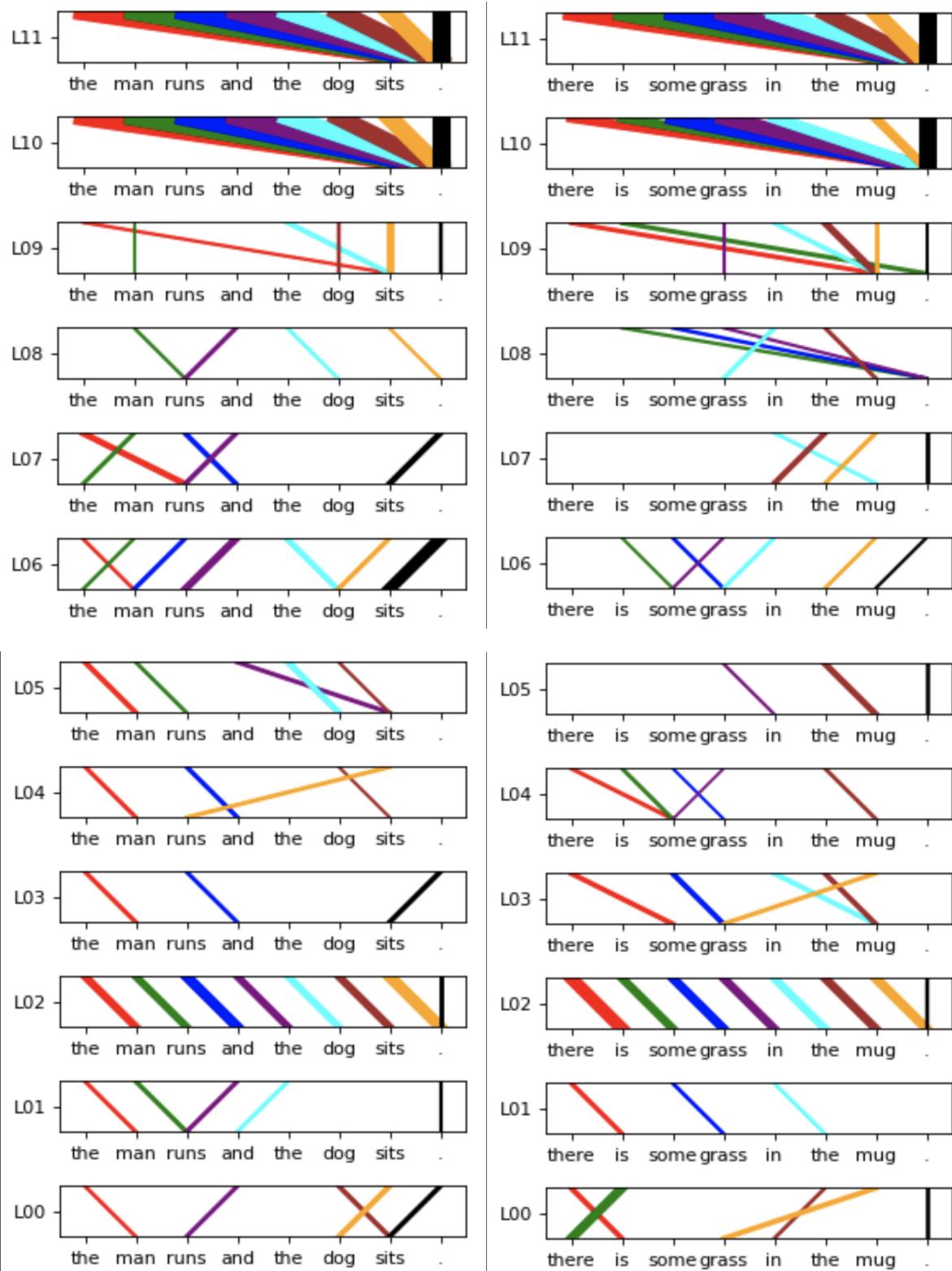


Figure 3.5: Attention Flow Visualization

The color of the line simply indicates the index of the target word representation while the thickness indicates how strong the attention activation is. Note that not all attention activations are plotted—only those that are greater than $k = 2$ standard deviations from the mean attention activation for that layer.

3.4 Locating Grounded Composition Circuits with Causal Tracing

Having gained some intuition on how representations are composed in language models, we now turn our attention to rigorously identifying where in a visually grounded language model these compositional representations form. To accomplish this, we adopt a method from causal abstraction analysis [32] and mechanistic interpretability, causal tracing [56].

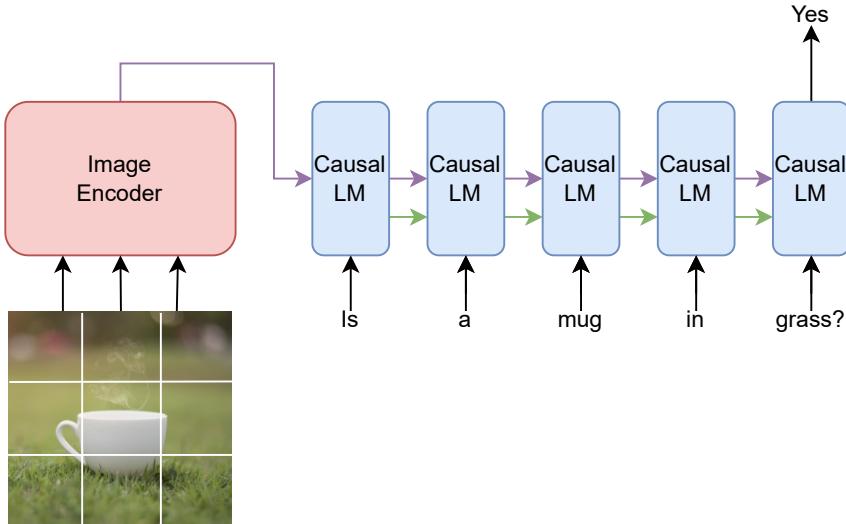


Figure 3.6: Visual Question Answering using an image-conditioned causal language model. Green arrows are causal attention and purple arrows are cross-attention.

Causal Tracing allows us to identify hidden representations in a neural network that strongly contribute to a specific behavior. In our case, we’re interested in the ability of a visually-grounded causal language model to answer questions that depend on the compositional structure of an image.

In Causal Tracing, we perform a normal forward pass as shown in Fig. 3.6, as well as a forward pass where we corrupt part of the input—in our case, the image. Then, we patch each hidden representation from the normal forward pass into the corrupted forward pass to see which hidden representations we patch result in the most significant improvement in performance (answering “yes” or “no”). To verify that we’re identifying neural states relevant for compositional representation and not just for answering “yes” or “no”, we can perform a similar operation on Winoground pairs.

Chapter 4

Enabling models to compose visually grounded semantics

4.1 Past Work

Compositionality is the ability to combine meanings of constituents according to structured rules. Recent work shows that Vision-Language Models (VLMs) fail to construct compositional representations and generally ignore syntactic & structural information [88, 57, 46]. Winoground [88] is a vision-language compositionality task that tests a VLM’s ability to match syntactic permutations of text with their visual interpretations, for example correctly matching “grass in mug” and “mug in grass” to their corresponding images. Winoground finds that all recent state-of-the-art VLMs perform below chance levels on this compositionality task. Contemporaneously, Milewski et al. [57] probe for structural knowledge in VLMs, finding that they encode significantly less linguistic syntax than Language Models (LMs) and virtually no visual structure. Recently, Yuksekgonul et al. [103] built a large dataset confirming that VLMs treat images as a ‘bag of objects’ and don’t adequately represent visuo-linguistic relations.

Since models must determine whether the compositional structure of an image matches that of the caption, it’s important for the model to learn to cross-modally align intra-modal relations. That is, if the relation from ‘mug’ to ‘grass’ is ‘in-ness’, the model should recognize when the equivalent physical relation holds between a mug and grass in the image, and representationally align these relations such that an image-text matching head may more easily determine whether the relations are cross-modally equivalent. In simpler terms, the compositional structure of input for each modality should be represented such that they can be cross-modally matched even for difficult examples like Winoground.

Unfortunately, there has been less highly influential work on **relation alignment** between vision & language, and Thrush et al. [88] did not benchmark any such models. In this work, we begin exploration of these relation alignment approaches by tentatively grouping them into 3 categories:

1. Structural Data: training a model on data that explicitly captures relational structure

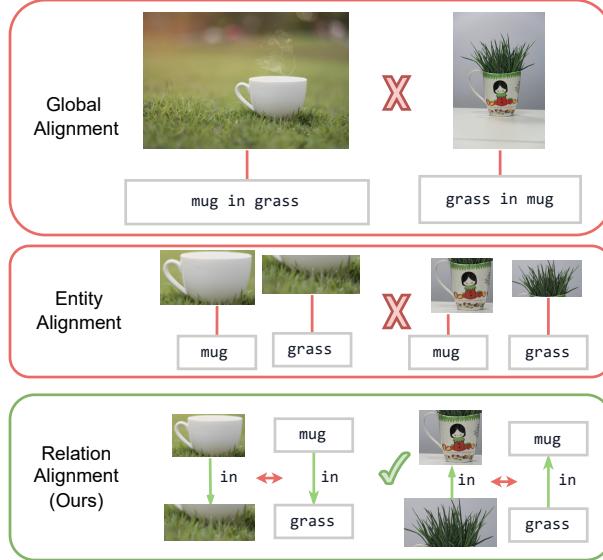


Figure 4.1: Global Alignment (GA) only aligns the entire image with the corresponding caption. Entity Alignment (EA) extracts entities from the image and caption for finer-grained alignment. Relation Alignment (RA) cross-modally aligns the intra-modal relations between entities in both the image and the text. We show RA is vital to improve compositional performance.

[95, 106, 102, 25, 91, 38]

2. Structural Model: infusing an inductive bias into the architecture of the model that enables more compositional representations [5, 33, 36, 105, 93, 39, 92]
3. Structural Training: modifying the objective function or imposing a parameter constraint to encourage relation alignment [77, 99, 100, 97]

While some of these works introduce ideas from multiple of these categories, we group them by their core contribution. For example, ROSITA proposes a graphical data pre-training approach, and a self-supervised objective to accompany it; we consider it a Structural Data approach since the training objective ultimately is just a necessity for the data being provided.

Unfortunately, many of these works do not provide publicly available code or pre-trained checkpoints, so we were unable to complete an exhaustive analysis of the compositional performance of these relation alignment approaches. Due to the added complexity of Structural Model approaches, we leave exploration of their compositional abilities to future work.

Since Structural Data approaches require complex annotations and Structural Model approaches are often incompatible with large transformers, we identify Structural Training as a promising avenue for providing compositional inductive biases to VLMs due to their architecture-agnostic compatibility and computational scalability.

4.2 Cross-modal Attention Congruence Regularization

We chose one exemplar for both Structural Data (ROSITA) and Structural Training (IAIS) that made their pre-trained image-text matching checkpoints available; we generated their scores on Winoground, which have not previously been calculated. In Tab. 4.1, we present these two relation alignment models’ Winoground scores alongside a few entity alignment and global alignment models.

Model	Text	Image	Group
MTurk Human	89.50	88.50	85.50
IAIS (RA-ST)	42.50	19.75	16.00
OSCAR+ (EA)	37.75	17.75	14.50
ROSITA (RA-SD)	35.25	15.25	12.25
UNITER (EA)	38.00	14.00	10.50
CLIP (GA)	30.75	10.50	8.00
LXMERT (GA)	19.25	7.00	4.00

Table 4.1: Comparison of Winoground scores for models using Global Alignment (GA), Entity Alignment (EA), Relation Alignment with Structural Data (RA-SD), and Relation Alignment with Structural Training (RA-ST). We find that IAIS, a recent relation alignment approach that uses attention regularization for structural training achieves universal performance improvements.

Notice that global alignment approaches tend to perform the lowest on Winoground, even when scaled considerably. Entity alignment approaches perform intermediately and OSCAR+ specifically held the state-of-the-art prior to our benchmarking of these relation alignment models. Of the two relation alignment approaches we benchmark, IAIS beats out OSCAR+ and achieves a new state-of-the-art on Winoground. But ROSITA, despite providing structural data to encourage cross-modal relation alignment, underperforms OSCAR+. We attribute this partly to the improved visual features OSCAR+ has access to as a result of VinVL, but further comparison of IAIS and ROSITA is explored in our recent work.

In this work, we propose a Structural Training approach for relation alignment that uses the cross-modal attention matrix as a change of basis¹ to the opposite modality, which we then compare to the original modality to calculate a divergence loss, effectively measuring cross-modal congruence between intra-modal attentions.

We show how our approach, Cross-modal Attention Congruence Regularization (CACR), generalizes previous Structural Training work on cross-modal attention regularization (IAIS [77]) by taking into account all possible entity alignments and computationally simplifying relation alignment. The CACR regularization term can easily be dropped into most transformer-based Vision-Language model objectives with no added data and minimal computational overhead, to encourage relation alignment during training. Finally, we show

¹not defined in a strict linear algebraic sense

that CACR_{base} improves on IAIS_{base}—where IAIS_{large} holds the current state-of-the-art on Winoground.

4.2.1 Method

How can we infuse the vision-language model’s training objective with an implicit structural prior that encourages cross-modal alignment of relations?

To attempt a solution to this question, we begin by noting that attention activations encode some degree of relational information. Attention values in transformers may be seen as an informational gating mechanism that implicitly encode how representations are composed [1]. For example, past work in language has shown how syntax trees may be extracted [52] from attention across layers and used to guide attention [8, 45] for improved compositionality. In this section, we extend this intuition to the multimodal domain by proposing to use the cross-modal attentions, which as a change-of-basis matrix encode a transformation from one modality’s compositional structure to the opposite modality’s, to encourage cross-modal relation alignment.

Relation Alignment Using Attention

In specific, we focus on the self-attention matrix S computed in a transformer by

$$S = QK^\top = (XW^Q)(XW^K)^\top \quad (4.1)$$

Then, some row i in S corresponds to a distribution over columns j_0, \dots, j_n where $S_{i,j}$ tells us how much of the previous layer’s entity representation j we want to infuse into the current layer’s entity representation i , intuitively their compositional relation. Since X is a series of visual and linguistic tokens, we can segment S into four submatrices for intra- and cross-modal relations [15]. Denote the intra-modal attention submatrices in the last multimodal encoder layer as S_{VV} (vision to vision) and S_{LL} (language to language); the cross-modal attention matrices as S_{VL} (vision to language) and S_{LV} (language to vision).

$$S = \begin{pmatrix} S_{LL} & S_{LV} \\ S_{VL} & S_{VV} \end{pmatrix} \quad (4.2)$$

If an image and caption have the same underlying compositional structure, the entities that cross-modally correspond to each other should bear similar intra-modal compositional structure. That is, a word w should attend to other words (in S_{LL}) in a similar way that its visual object counterpart o attends to other objects (in S_{VV}). Furthermore, we can use the cross-modal matrices (S_{LV} and S_{VL}) to identify entities that cross-modally correspond as they will generally attend to each other [2]. Unfortunately, since representations are heavily contextualized by the final layer, clear bijective correspondences between words and objects may not always be identified using an argmax over the cross-modal attention matrix as Ren et al. [77] attempts. Deeper analysis of when their model, IAIS, fails to identify cross-modal bijective correspondences is provided in Sec. 4.2.3.

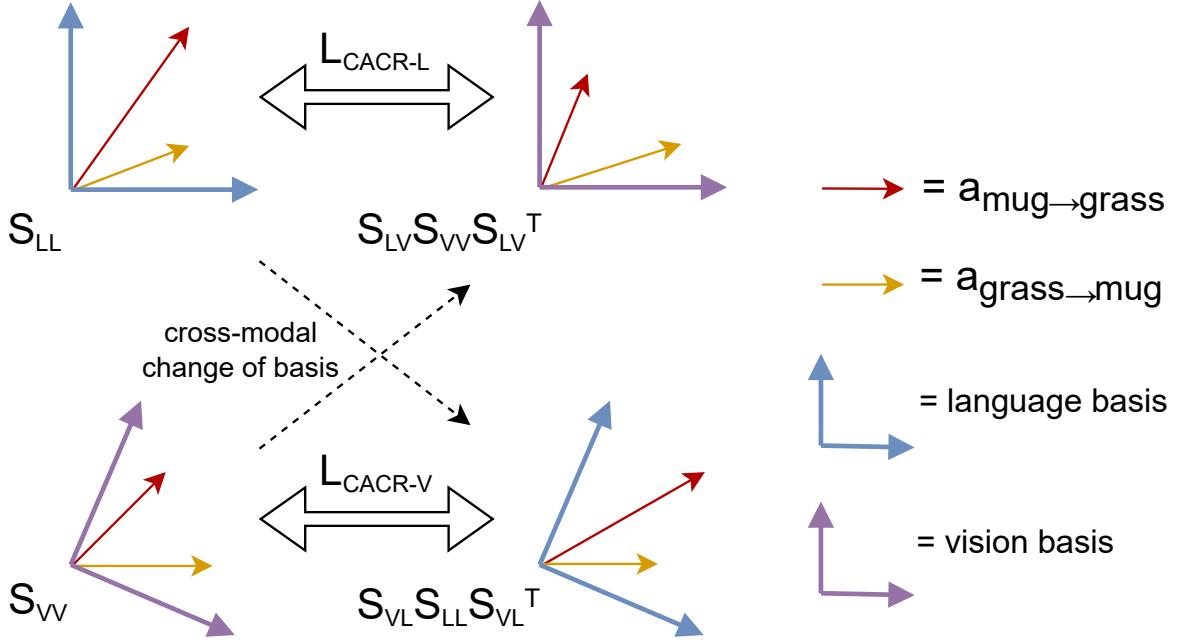


Figure 4.2: Top: language attention (S_{LL}) is aligned with the visual attention projected into the language basis ($S_{LV}S_{VV}S_{LV}^T$) to calculate \mathcal{L}_{CACR-L} ; specific attention values (yellow, red) capturing intra-modal relations are cross-modally aligned as a result. Bottom: as above, but in the vision basis.

Attention Congruence

We opt to use the cross-modal matrices (S_{LV} and S_{VL}) as a whole to ‘change basis’ to the opposite modality, with which we can then calculate ‘congruence’ with the original modality. However, we use ‘change of basis’ and ‘congruence’ loosely since the cross-modal matrices are not guaranteed to be square and thus do not satisfy strict linear algebraic definitions. We formulate S_{VV} in the language basis as $S_{LV}S_{VV}S_{LV}^T$, which we then encourage to be similar to S_{LL} .

Under the hood, this says that for each $a_{i \rightarrow j} \in S_{LL}$, we can use row vectors $S_{LV,i}$ and $S_{LV,j}^T$ to calculate a weighted sum $a_{i \rightarrow j}^*$ over S_{VV} . If we were to do this for all i, j , we would construct a matrix of the same shape as S_{LL} where each entry is $a_{i \rightarrow j}^*$, i.e. an approximation of the visual correspondent of the relation $a_{i \rightarrow j}$ taking into account all the possible cross-modal alignments of i and j . Since this computation intuitively makes a lot of sense and may more easily be compared to previous approaches, we choose to illustrate it in Fig. 4.3. However, since this computation is relatively expensive, we instead use the $S_{LV}S_{VV}S_{LV}^T$ formulation which produces the same matrix of $a_{i \rightarrow j}^*$ values but with considerably fewer operations. This also enables us to view the operation as a ‘change-of-basis’ to the opposite modality and the CACR loss as encouraging a sense of cross-modal ‘congruence’.

Specifically, we align the original S_{LL} with the language-basis S_{VV} matrix using $\mathcal{L}_{\text{CACR-L}}$:

$$\mathcal{L}_{\text{CACR-L}} = \text{m-KL}(\sigma(S_{LV}S_{VV}S_{LV}^\top), \sigma(S_{LL})). \quad (4.3)$$

We apply a softmax to normalize both matrices since $S_{LV}S_{VV}S_{LV}^\top$ will generally be larger in scale due to summation. Additionally, m-KL(\cdot) [77] is a symmetric matrix-based Kullback-Leibler Divergence (m-KL) which measures the distance between two matrices S and S' :

$$\text{m-KL}(S, S') = \sum_i^N \text{KL}(S_i || S'_i) + \text{KL}(S'_i || S_i), \quad (4.4)$$

where $(\cdot)_i$ stands for the i^{th} row-vector in the matrix.

Similarly, we have $\mathcal{L}_{\text{CACR-V}}$:

$$\mathcal{L}_{\text{CACR-V}} = \text{m-KL}(\sigma(S_{VL}S_{LL}S_{VL}^\top), \sigma(S_{VV})), \quad (4.5)$$

Combining $\mathcal{L}_{\text{CACR-V}}$ and $\mathcal{L}_{\text{CACR-L}}$, we present our $\mathcal{L}_{\text{CACR}}$ objective, an attention activation regularizer for cross-modal relation alignment:

$$\mathcal{L}_{\text{CACR}} = \mathcal{L}_{\text{CACR-V}} + \mathcal{L}_{\text{CACR-L}}. \quad (4.6)$$

When the vision inputs and the language inputs have the same sequence length and S_{VL}, S_{LV} are invertible, then S_{VV} and $S_{VL}S_{LL}S_{VL}^\top$ (as well as S_{LL} and $S_{LV}S_{VV}S_{LV}^\top$) can become strictly congruent. In this case, $S_{VL}S_{LL}S_{VL}^\top$ can be interpreted as the language view of S_{VV} . Aligning $S_{VL}S_{LL}S_{VL}^\top$ and S_{VV} leads to cross-modal relation alignment. It is similar for $S_{LV}S_{VV}S_{LV}^\top$ and S_{LL} . In the general case where the vision inputs and the language inputs may have different sequence lengths, the two forms are not linear algebraically congruent but the relevant intuition still holds.

Hard and Soft Cross-modal Equivalence

In this section, we show that CACR can be interpreted as leveraging cross-modal soft equivalences, where IAIS [77] uses hard bijective equivalences. In their approach, each element in the intra-modal attention matrix is aligned with a single counterpart in the opposite modality. This is built upon a strict assumption that there exists a one-to-one mapping (provided by an argmax over the cross-modal attention) from S_{LL} to S_{VV} and vice versa, which is unsatisfied in practical cases. CACR may be seen as a soft cross-modal equivalence method which instead uses the whole S_{LV} (or S_{VL}) to implicitly build an ‘equivalence weighting’ which is then used to compute a weighted mean over S_{VV} (or S_{LL}). We illustrate and compare hard cross-modal equivalence and our soft cross-modal equivalence in Figure 4.3, taking the language-side alignment as an example.

We note that IAIS could be seen as a special case of soft cross-modal equivalence by forcing the cross-modal attention map to be a one-hot matrix, i.e., taking the argmax of the attention matrix as the index of the cross-modal counterpart. We show in Section 4.2.3

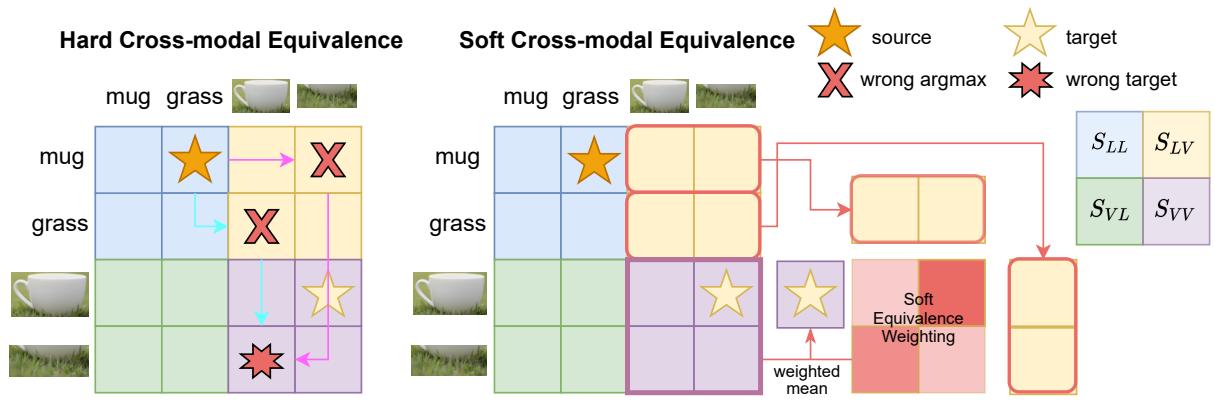


Figure 4.3: Comparison of the hard cross-modal equivalence used in IAIS (left) and the soft cross-modal equivalence we propose in CACR (right), with an example from Winoground to illustrate how the target visual equivalent (yellow star) of a source linguistic relation (orange star) is calculated. Hard cross-modal equivalence attempts to build a one-to-one mapping between language and vision by applying argmax (red crosses) to the S_{LV} row vectors. Our soft cross-modal equivalence instead uses the whole S_{LV} row vectors to calculate a weighted (red) mean over S_{VV} . The scalar that is produced corresponds to the attention from ‘mug’ to ‘grass’ but in a visual basis. We note that IAIS can be seen as a special case of soft cross-modal equivalence by forcing the attention matrix (red) to be a one-hot matrix where the max value is set to 1 and all others to 0. CACR linear algebraically simplifies soft cross-modal equivalence for computational efficiency.

that IAIS can have inferior performance when a clear bijective cross-modal correspondence isn't available.

In Alg. 1, we show the pseudo-code of the soft cross-modal equivalence method for calculating the vision-side loss. S_{CACR-V} can be computed similarly. Computing the hard and soft cross-modal equivalence is computationally complex and difficult to be parallelized due to indexing operations. For practical applications, we sought to simplify this soft cross-modal equivalence algorithm to a mathematical equivalent that would improve computational tractability. From here, we arrive at CACR, which is a closed-form formulation of soft cross-modal equivalence which utilizes only differentiable matrix multiplications. Therefore, our CACR is more **computationally efficient** and **easier to be parallelized** than soft cross-modal equivalence.

Algorithm 1 Soft Cross-modal Equivalence (V)

Require: $S_{LL} \in N \times N, S_{VL} \in N \times M, S_{VV} \in M \times M$

```

1:  $\mathcal{L} \leftarrow 0$ 
2: for  $i, j \in S_{VV}$  do
3:    $W \leftarrow S_{VL}[i] \cdot S_{VL}^\top[j]$                                  $\triangleright$  soft weighting
4:    $a_{i \rightarrow j}^* \leftarrow \overline{W} \circ S_{LL}$                            $\triangleright$  element-wise weighted mean
5:    $\mathcal{L} = \mathcal{L} + \text{m-KL}(a_{i \rightarrow j}^*, S_{VV}[i, j])$ 
6: end for
7: return  $\mathcal{L}$ 

```

Proof of Equivalence Between CACR and Soft Cross-modal Equivalence

Computing the hard (IAIS) and soft cross-modal equivalence is computationally complex and difficult to parallelize due to indexing operations. However, CACR loss is mathematically equivalent to soft cross-modal equivalence but can be computed efficiently. We take $CACR_V$ for illustration of this equivalence, but $CACR_L$ can be proved in the same way.

Beginning with the visual-basis form of S_{LL} in CACR, the attention at index $[i, j]$ in $S_{VL}S_{LL}S_{VL}^\top$ is

$$\begin{aligned}
& (S_{VL}S_{LL}S_{VL}^\top)[i, j] \\
&= \sum_p^{N_L} \sum_k^{N_L} a_{v_i \rightarrow l_k} a_{l_p \rightarrow l_k} a_{v_j \rightarrow l_p} \\
&= \sum_p^{N_L} \sum_k^{N_L} \underbrace{S_{VL}[i, k]S_{VL}[j, p]}_{\text{soft weighting}} S_{LL}[p, k]
\end{aligned} \tag{4.7}$$

where $a_{v_i \rightarrow l_j}$ stands for the attention from the i -th visual token to the j -th linguistic token, N_L is the total number of language tokens and N_V is the total number of the visual tokens. Comparing Eq.4.7 and Alg.1, we observe that the summation we arrive at above is equivalent to the content of the for-loop (line 3-5). Thus, although of seemingly different linear algebraic form, CACR generalizes IAIS by way of its equivalence to the Soft Cross-modal Equivalence formulation presented above.

4.2.2 Results

How does CACR compare to other vision-language models in its compositional ability?

Model	Text	Image	Group
MTurk Human	89.50	88.50	85.50
CACR _{base}	39.25	17.75	14.25
UNITER _{large}	43.50	14.75	13.75
IAIS _{base}	37.50	16.75	13.00
UNITER _{base}	32.75	11.75	8.50

Table 4.2: CACR outperforms its pre-trained baseline (UNITER) and an alternative attention regularization approach (IAIS) across all Winoground scores.

In Tab. 4.2, we present our approach’s scores alongside a few other models. Since we use CACR to fine-tune UNITER, we include scores for the two baseline UNITER sizes. We also include scores for IAIS_{base} which is also built on UNITER.

The fact that CACR_{base} outperforms IAIS_{base} suggests that, with adequate computational resources, CACR_{large} could similarly outperform IAIS_{large}, potentially achieving a new state-of-the-art on Winoground. Furthermore, its performance compared to UNITER_{large} is impressive considering that CACR_{base} is approximately half its size in parameters.

Model	Image	Image	Text	Text
	R@1	R@10	R@1	R@10
IAIS _{base}	73.54	96.32	86.10	99.10
UNITER _{base}	72.52	96.08	85.90	98.80
CACR _{base}	70.88	95.68	83.50	98.80

Table 4.3: CACR performance on Flickr30k has marginal reductions from UNITER, suggesting performance could be improved even further with a hyperparameter search.

Finally, we report Flickr30k retrieval scores in Tab. 4.3 to verify that we are not somehow overfitting to Winoground. Though CACR takes some minor losses to its retrieval scores, this may be attributed to imperfect hyperparameters, suggesting that CACR’s performance on Winoground could be even higher with adequate hyperparameter tuning. It’s also important to remember here that we’re only training on Flickr30k, so this isn’t a case of our model overfitting to Winoground and ‘forgetting’ its true image-text matching ability. Rather, it shows that the hyperparameters that we adapted from IAIS need to be modified to more perfectly train CACR on Flickr30k, which would then carry over to compositional improvements on Winoground.

We fine-tuned CACR on Flickr30k [101] for 5000 epochs using PyTorch [68] with a train-validation-test split of 90-5-5. The training batch size is 4 and 31 negative samples are provided for every individual positive sample in a standard image-text matching training setup. We use a learning rate of 5×10^{-5} , the *AdamW* optimizer [47], and introduce \mathcal{L}_{CACR} with an exponential warmup schedule. Training was completed on a node with 4 NVIDIA GTX 1080 Ti’s, each with 11 GB of memory.

4.2.3 Analysis

Why does CACR’s soft cross-modal equivalence approach outperform hard cross-modal equivalence?

Qualitative

Hard cross-modal equivalence, implemented by IAIS, assumes that cross-modal submatrices can be used to find a singular equivalent of an entity in the opposite modality. Specifically, if $i^* = \text{argmax}(S_{LV}[i])$ then $S_{LL}[i]$ should correspond to $S_{VV}[i^*]$. In simple terms, IAIS says the following: if word A attends most to object A and word B attends most to object B, then word A should attend to word B in a similar way that object A attends to object B. Underlying IAIS is the hard assumption that argmaxing over the cross-modal attention submatrix is an effective means of identifying the opposite modality equivalent of an entity. However, we show in this section that this is often not the case.

Given the argmaxes for rows in the S_{LV} submatrix, we can identify the bounding box that each token maximally attends to, which IAIS assumes is its visual equivalent. In Fig. 4.4a, we visualize an example where ‘clouds’ maximally attends (green) to the ground, which would prevent IAIS from identifying the correct cross-modal equivalence. ‘Turbines’ (Fig. 4.4b), on the other hand, maximally attends to a bounding box that better matches our intuition. It is qualitatively clear from the several examples displayed that the argmax assumption often fails to identify the correct cross-modal equivalence. Since words may attend to several visual tokens for different reasons, we shouldn’t assume that the cross-modal argmax provides us with a clear bijective correspondence.

Instead, the cross-modal matrices should be seen as providing useful high-level information about what visual entities are relevant to a word, and vice versa. We can certainly gain useful information about cross-modal correspondences using it, but it isn’t as simple as using an argmax, due to words having multiple referents and entity representations being intermixed. Instead, our soft cross-modal equivalence approach takes all the possible cross-modal alignments into account with a weighted sum.

To illustrate how the soft approach accounts for critical cross-modal alignment information, we present a few Winoground examples with UNITER’s cross-modal attention activations in Fig. 4.4 and 4.5. We use UNITER since this is the baseline model from which attentional information is bootstrapped to calculate cross-modal alignments. For example, in Fig. 4.5c, using the representation for the bounding box covering the mug’s handle may not adequately capture the visual referent of ‘mug’ and therefore disrupt our ability to calculate the visual-basis relation between ‘mug’ and ‘grass’ if restricted by an argmax.

Quantitative

In the absence of annotations, we attempted a quantitative measurement of whether overlap in argmaxes (several words attending to one bounding box or vice versa) as quantified by

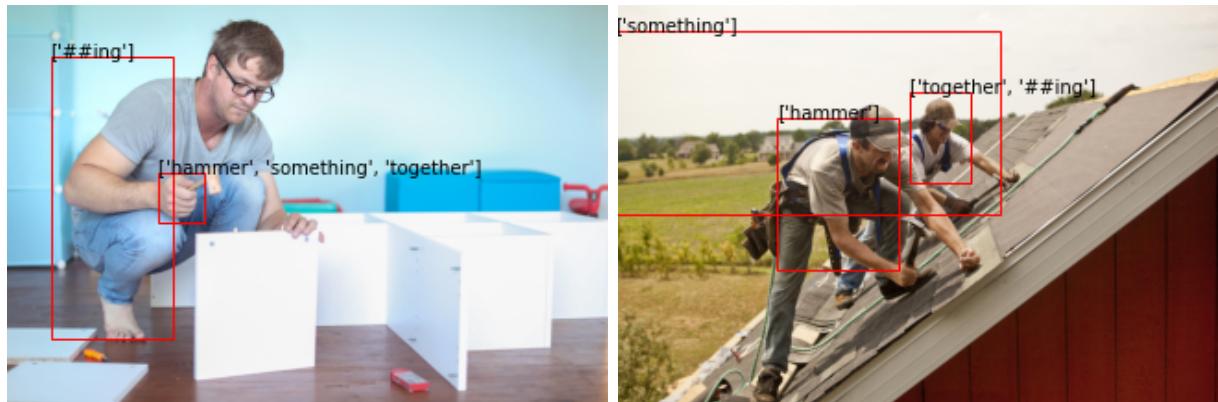
(a) *clouds*(b) *turbines*(c) *"hammering something together"*(d) *"together hammering something"*

Figure 4.4: Top: UNITER S_{LV} attention for caption “*a few clouds and many wind turbines*”, with the bounding box maximally attended to by the token in green; other highly attended boxes in red. Bottom: UNITER S_{LV} attention with bounding boxes labeled with the tokens that maximally attend to them. Note that argmaxes often fail to precisely identify cross-modal equivalence.

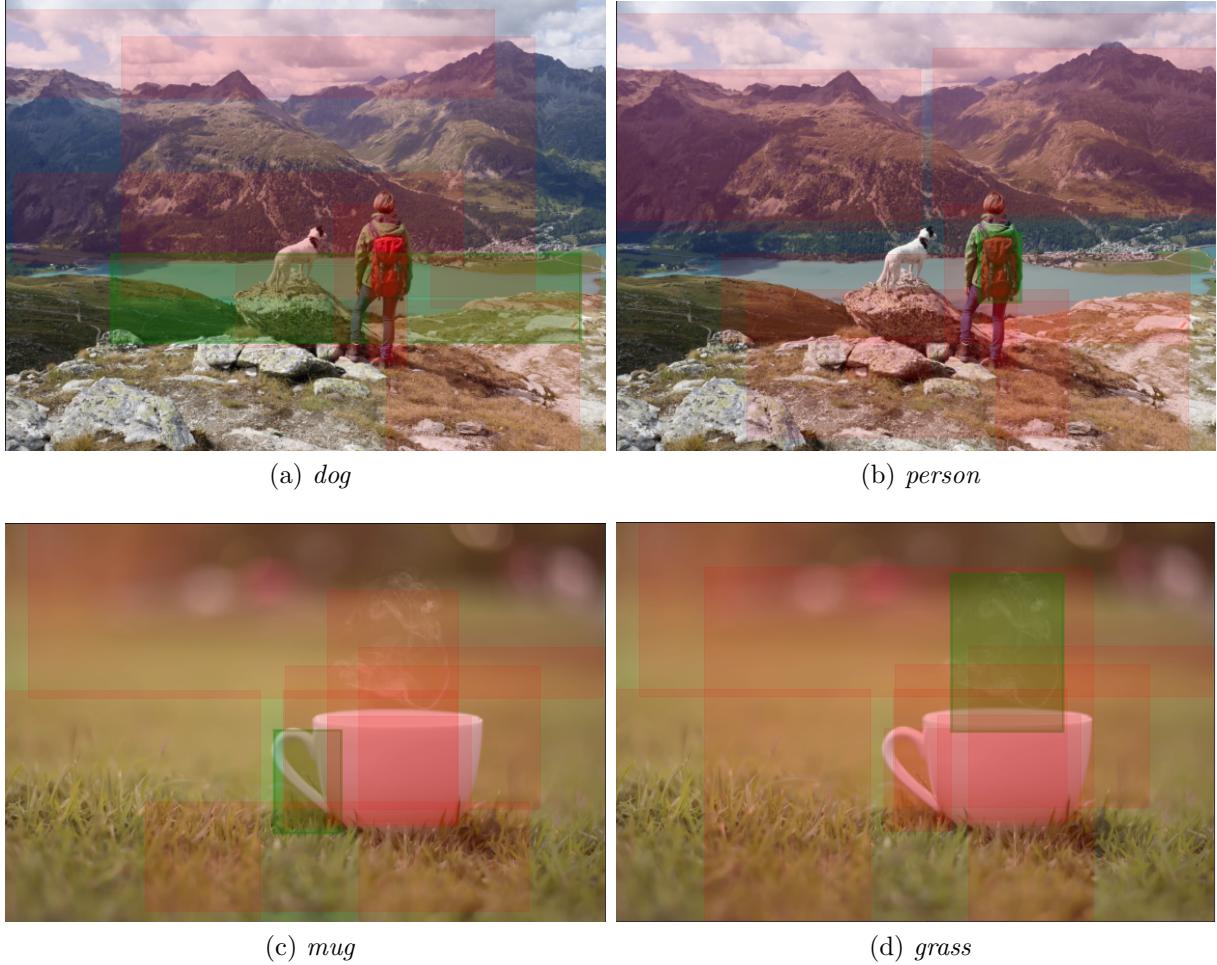


Figure 4.5: UNITER S_{VL} attention for captions “*a dog on a rock next to a person*” and “*there is a mug in some grass*”. Shown are boxes that attend highly to the displayed token, with the maximally attending bounding box in green; others in red. Observe that although the argmax often does pick up on a relevant bounding box, it is prone to missing critical visual information, e.g. focusing on only the backpack in (b).

the Shannon Entropy of argmax indices inversely correlates with soft Winoground score. Intuitively, if an example has more like a one-to-one mapping between text and image, the entropy of its cross-modal argmaxes should be higher as each token will attend to a different box, which would suggest that the model is better aligning entities. However, we found no significant correlation with Winoground score, which we attribute to the fact that high entropy on its own doesn’t mean *correct* entity alignment.

Rather, high entropy in argmax indices could still be produced by a bad representation if ‘mug’ attends to the grass & ‘grass’ attends to the mug; conversely, low entropy could be produced by a good representation for an example like ‘fire truck’ where two tokens refer to a single object. Quantitative exploration of cross-modal attention is difficult without annotations and we leave this task to future work to explore in a multimodal compositionality context.

As a general takeaway, while the cross-modal argmax assumption of IAIS does hold in some cases and may be more meaningful during the course of IAIS training, it is clearly quite a strict assumption that could suffer if an entity attends to several cross-modal entities or there are no corresponding cross-modal entities. Furthermore, since IAIS is only active in the final self-attention layer, all the token representations are intermixed and therefore don’t necessarily have a one-to-one correspondence with our intuitive notions of what they should be—the word ‘turbine’ may not solely represent the traditional meaning of that word but perhaps the entire scene that includes the turbines, clouds, and ground.

We hypothesize that by removing the hard argmax assumption, our approach better accounts for varying cross-modal entity equivalences and thus enables stronger relation alignment. By also calculating alignment between all pairs of source and target modality entities, CACR should considerably improve sample efficiency, which is important considering that the final layer S matrix of the converged IAIS model is largely flat. Therefore it’s important to backpropagate as much alignment knowledge over the course of training as possible, which CACR’s soft equivalence weighting implicitly enables.

4.2.4 Conclusion

In this work, we identified that a key factor holding back models from vision-language representational compositionality is cross-modal relation alignment. We categorized recent compositional inductive bias approaches into 3 categories: Structural Model, Structural Data, and Structural Training, showing that a previous Structural Training model (IAIS) achieves state-of-the-art performance on Winoground. We then identified a potential key weakness in IAIS, its hard argmax assumption, and developed a soft cross-modal equivalence approach to address it. Having linear algebraically simplified this approach, we arrived at CACR, an auxiliary loss that encourages cross-modal congruence of intra-modal attention. CACR improves on IAIS’ performance on Winoground, and even outperforms a UNITER model nearly twice as large.

As computational scaling becomes more widespread, it’s necessary to develop compositional inductive biases that do not require complex annotated data or exotic model architectures. Our work illustrates how taking advantage of the transformer’s own attentional structure can improve the quality of fine-grained vision-language representations, opening the avenue for large scale approaches to visually-grounded compositionality.

4.3 Syntactic MeanPool for Sentence Embedding

Recent work has shown that state-of-the-art vision-language models often fail to represent compositional or relational structure. In simple terms, they are liable to conflate representations of phrases such as “the mug is in the grass” and “the grass is in the mug”, resulting in failure to cross-modally distinguish them. Standard practice is to produce embeddings through mean-pooling token embeddings from the final layer. However, in order to capture

the compositional structure inherent in language, we instead propose hierarchically averaging token vectors given a syntax parse of the text, a novel Structural Model approach according to the taxonomy of Sec. 4.

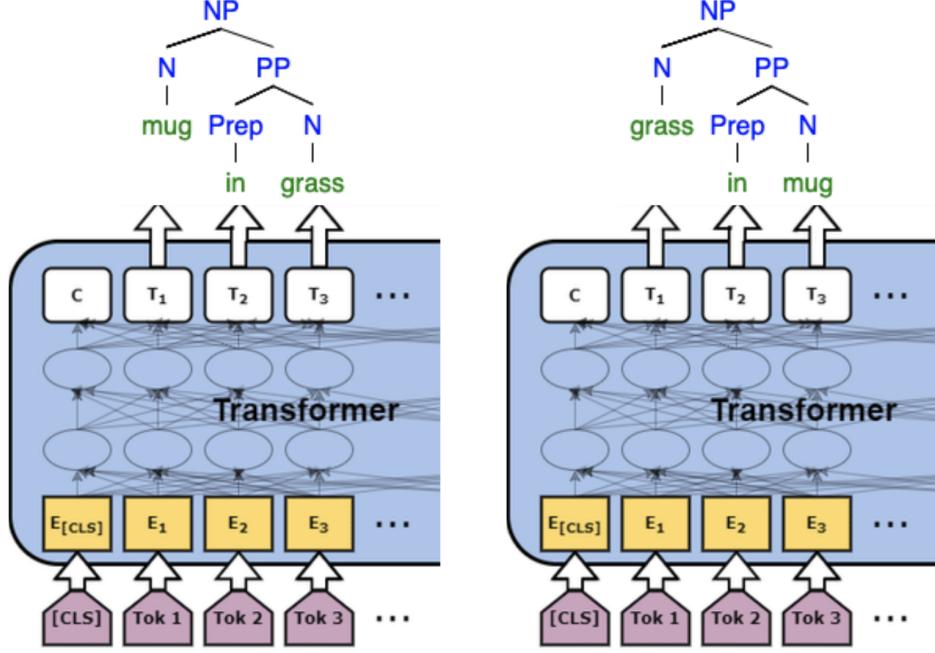


Figure 4.6: Syntactic MeanPool leverages the non-associativity of the mean operation to inject a compositional inductive bias into the hierarchical pooling of a sentence embedding.

We empirically show that our Syntax-guided MeanPool increases the distance between compositionally distinct pairs of sentences compared to MeanPool alone, across multiple embedding models in Tab. 4.4.

Performance	SynMeanPool	MeanPool	CLS Pool
Euc Dist between Winoground captions (BERT)	3.21	2.14	2.57
Spearman Correlation with STS (BERT)	.302	.476	.203
Euc Dist between Winoground captions (CLIP)	.382	.287	.224
Winoground Group Score (CLIP)	6.00	7.75	8.25

Table 4.4: Three sentence embedding pooling strategies, with their influence on Euclidean Distance, Semantic Textual Similarity [16], and Winoground Score.

Despite increasing Euclidean distance between Winoground captions, intuitively making them more easy to distinguish, why does SynMeanPool not result in increased Winoground score? We suggest that this is likely due to the fact that CLIP [70] is actually trained for, as evidenced by its Winoground performance being the

highest. The obvious next step is to fine-tune CLIP using SynMeanPool to see whether we can successfully inject it with this compositional prior, but we have yet to run this experiment due to implementation issues regarding the PyTorch [68] backprop computation graph.

Chapter 5

Conclusion

5.1 Multimodal Brain

Multimodal “Hub & Spoke” structures have been identified in the human brain that enable embodied semantic cognition in the anterior & inferior temporal lobe [7, 72]. Through the use of continuous vectorial representations, DL similarly provides a paradigm in which multiple modalities can be aligned to learn joint embeddings of vision and language in the form of VLMs [48]. While VLMs can clearly map simple words to images, it’s natural to ask whether they can ground highly compositional utterances as humans do [81].

5.2 Vision-Language Psychology

More broadly, the ability to impose structure upon the world where discrete objects can be related hierarchically is important for human intelligence and likely critical for developing artificial general intelligence. The role of language in imposing these world structures is potentially causal but certainly connected [49], a controversial debate in the cognitive sciences [24].

This line of work may also comment on debates in linguistics over the innateness of syntax (universal grammar) [98]. If structural knowledge can be learned in an unsupervised or self-supervised manner without an innate ”syntax module”, we contradict the Chomskyan position. However, if we find that structural augmentation methods that require some innate ability to produce hierarchical structures from input are more performant, this could imply that syntactic ability is indeed innate. The architecture of such a syntax module for structural augmentation may then inform or corroborate neuroscience work exploring the cortical basis of universal grammar and its role in multimodal/visual cognition.

In a related work, we present a time constraint based approach to explore how the mind aligns relations between imagery and text. Vision-language relation alignment refers to the ability to determine that an image of a *lightbulb surrounded by plants* is indeed that and not instead ‘a *lightbulb surrounding plants*’. We expose an image to the subject for some

amount of time t and then asked them to choose which of two captions—that only differ in relation—matched the image. We find that performing this task takes significantly longer than simple object recognition, requiring 500ms to achieve 75% and 1000ms to achieve 90% accuracy. We explore what this might suggest about cognitive architecture, with the lack of a narrow threshold suggesting no explicit module for vision-language relation alignment.

5.3 Multimodal Semantics

Finally, we may seek to draw a more tenuous philosophical parallel between vectorial semantics and contemporary work on type theoretic foundations of mathematics. Past work has sought to treat grammatical types as vectorspaces such that syntactic compositions of types reflect their compositional meaning in the equivalent vectorspace, using algebraic and categorical formalizations [21]. Concretely, if we have "rabbits hop", our simplified parse tree is $[S [NP [\text{rabbits}]] [VP [\text{hop}]]]$. The token "rabbits" is located in the space corresponding to type NP, while "hop" is in the VP space. The syntax rule that composes them ($S \rightarrow NP VP$) is a type reduction which can be treated as a morphism in the category of vectorspaces, enabling their embeddings to compose according to the sentence's syntax.

Logic	Type Theory	Homotopy	Semantics	Multimodal
Proposition	Type	Space	Denotation	Text
Proof	Term	Point	Referent	Image

Using related mathematical machinery, we suggest a parallel between grounded semantic vectorspaces and homotopy types. Homotopy type theory [69] builds on the Curry-Howard isomorphism by treating propositions (types) as spaces and their proofs (terms of a type) as points in their space. We give HoTT a linguistic interpretation, where spaces correspond to a word's semantic denotation and points in the space are intuitively its grounded referents. Thus the meaning of "rabbit" is captured by the topological structure of the embedding subspace consisting of images of rabbits. Composing these spaces is enabled by the same type theoretic operations that underlie logical composition in HoTT. Denotation spaces may hierarchically nest in broader type universes corresponding to semantic fields or syntactic types. Whether any of this is empirically true or derivable from VLM embedding spaces remains yet to be shown, but such a formalism may elegantly capture grounded meaning in a way historically unattempted in formal semantics.

5.3.1 Indian Philosophy

My interest in multimodal semantic compositionality is not simply a pragmatic concern with what I believe is critical to develop Artificial General Intelligence. Rather, it's a direct consequence of my ethno-cultural identity. Philologists have noted that the importance of Pythagoras & Euclid's mathematics in the Western intellectual tradition is much the same

as Pāṇini’s linguistics in India [84]; the latter developed the first ever generative grammar using a concise system of symbols, functions, and meta-rules.

The Sanskrit tradition directly contributed to the rise of historical and structural linguistics in the West, with pioneers like Saussure & Chomsky directly citing the work of Bhartṛhari & Pāṇini thousands of years prior [66]. In fact, India witnessed extensive debate on issues of semantic compositionality [79] & perceptual grounding [13] among the Nyāya, Mīmāṃsā, and Buddhist schools c. 6th-12th centuries AD. In this way, I see my work on the word-world interface as simply continuing the work of my ancestors using the best tools available to me—computers (and a whole lotta GPUs).

References

- [1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, 2020.
- [2] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21406–21415, 2022.
- [3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1545–1554, 2016.
- [5] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016.
- [6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- [7] Sharon M Antonucci, P  lagie M Beeson, David M Labiner, and Steven Z Rapcsak. Lexical retrieval and semantic knowledge in patients with left inferior temporal lobe lesions. *Aphasiology*, 22(3):281–304, 2008.
- [8] Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. Syntax-bert: Improving pre-trained transformers with syntax trees. *arXiv preprint arXiv:2103.04350*, 2021.
- [9] Petra Baran  kov   and Ond  ej Bojar. In search for linear relations in sentence embedding spaces. *arXiv preprint arXiv:1910.03375*, 2019.

- [10] Marco Baroni. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791):20190307, 2020.
- [11] Yoshua Bengio. From system 1 deep learning to system 2 deep learning. In *Neural Information Processing Systems*, 2019.
- [12] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, 2020.
- [13] Johannes Bronkhorst. A note on nirvikalpaka and savikalpaka perception. *Philosophy East and West*, 61(2):373–379, 2011.
- [14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [15] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multi-modal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994, 2021. doi: 10.1162/tacl_a_00408. URL <https://aclanthology.org/2021.tacl-1.58>.
- [16] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, 2017.
- [17] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.
- [18] Noam Chomsky. Syntactic structures. In *Syntactic Structures*. De Gruyter Mouton, 1957.
- [19] Grzegorz Chrupała and Afra Alishahi. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, 2019.
- [20] Volkan Cirik, Taylor Berg-Kirkpatrick, and Louis-Philippe Morency. Using syntax to ground referring expressions in natural images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [21] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen J Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36(1):345–384, 2010.

- [22] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*, 2017.
- [23] Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $\$ \& ! \# ^ *$ vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, 2018.
- [24] Michael C Corballis. The recursive mind. In *The Recursive Mind*. Princeton University Press, 2014.
- [25] Yuhao Cui, Zhou Yu, Chunqi Wang, Zhongzhou Zhao, Ji Zhang, Meng Wang, and Jun Yu. Rosita: Enhancing vision-and-language semantic alignments via cross-and intra-modal knowledge integration. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 797–806, 2021.
- [26] Ferdinand De Saussure. *Course in general linguistics*. Columbia University Press, 1916.
- [27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- [28] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. *arXiv preprint arXiv:2211.00768*, 2022.
- [29] Steffen Eger, Andreas Rücklé, and Iryna Gurevych. Pitfalls in the evaluation of sentence embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 55–60, 2019.
- [30] Gottlob Frege. Über sinn und bedeutung. *Zeitschrift für Philosophie Und Philosophische Kritik*, 100(1):25–50, 1892.
- [31] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34: 9574–9586, 2021.
- [32] Atticus Geiger, Chris Potts, and Thomas Icard. Causal abstraction for faithful model interpretation. *arXiv preprint arXiv:2301.04709*, 2023.
- [33] Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. Aligning linguistic words and visual semantic units for image captioning. In *Proceedings of the 27th ACM International Conference on Multimedia, MM '19*, page

- 765–773, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368896. doi: 10.1145/3343031.3350943. URL <https://doi.org/10.1145/3343031.3350943>.
- [34] Evan Hernandez and Jacob Andreas. The low-dimensional linear geometry of contextualized word representations. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 82–93, 2021.
 - [35] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.
 - [36] Yining Hong, Qing Li, Song-Chun Zhu, and Siyuan Huang. Vlgrammar: Grounded grammar induction of vision and language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1665–1674, October 2021.
 - [37] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1356. URL <https://aclanthology.org/P19-1356>.
 - [38] Zaid Khan, Vijay Kumar BG, Xiang Yu, Samuel Schulter, Manmohan Chandraker, and Yun Fu. Single-stream multi-level alignment for vision-language pretraining. *arXiv preprint arXiv:2203.14395*, 2022.
 - [39] Taehyeong Kim, Hyeonseop Song, and Byoung-Tak Zhang. Cross-modal alignment learning of vision-language conceptual systems. *arXiv preprint arXiv:2208.01744*, 2022.
 - [40] Nikita Kitaev, Thomas Lu, and Dan Klein. Learned incremental representations for parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3086–3095, 2022.
 - [41] Katarzyna Krasnowska-Kieraś and Alina Wróblewska. Empirical linguistic study of sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5729–5739, 2019.
 - [42] Ilia Kuznetsov and Iryna Gurevych. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–182, 2020.
 - [43] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

- [44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [45] Zhongli Li, Qingyu Zhou, Chao Li, Ke Xu, and Yunbo Cao. Improving bert with syntax-aware local attention. *arXiv preprint arXiv:2012.15150*, 2020.
- [46] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022.
- [47] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [48] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [49] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 2021.
- [50] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. Crepe: Can vision-language foundation models reason compositionally? *arXiv preprint arXiv:2212.07796*, 2022.
- [51] Mitch Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.
- [52] David Mareček and Rudolf Rosa. From balustrades to pierre vinken: Looking for syntax in transformer self-attentions. *arXiv preprint arXiv:1906.01958*, 2019.
- [53] Bimal Krishna Matilal. Reference and existence in nyaya and buddhist logic. *Journal of Indian Philosophy*, 1(1):83–110, 1970.
- [54] Rowan Hall Maudslay and Ryan Cotterell. Do syntactic probes probe syntax? experiments with jabberwocky probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131, 2021.
- [55] R Thomas McCoy, Tal Linzen, Ewan Dunbar, and Paul Smolensky. Rnns implicitly implement tensor product representations. *arXiv preprint arXiv:1812.08718*, 2018.
- [56] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

- [57] Victor Milewski, Miryam de Lhoneux, and Marie Francine Moens. Finding structural knowledge in multimodal-bert. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5658–5671, 2022.
- [58] Richard Montague. English as a formal language. In *Logic and philosophy for linguists*, pages 94–121. De Gruyter Mouton, 1970.
- [59] Richard Montague. Universal grammar. *Theoria*, 36(3):373–398, 1970. doi: 10.1111/j.1755-2567.1970.tb00434.x.
- [60] Richard Montague. English as a formal language. 1970.
- [61] Raymond J Mooney. Semantic parsing: Past, present, and future. In *Presentation slides from the ACL Workshop on Semantic Parsing*, 2014.
- [62] Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D Manning. Characterizing intrinsic compositionality in transformers with tree projections. In *International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=sAO0eI878Ns>.
- [63] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*, 2021.
- [64] Mitja Nikolaus, Emmanuelle Salin, Stephane Ayache, Abdellah Fourtassi, and Benoit Favre. Do vision-and-language transformers learn grounded predicate-noun dependencies? *arXiv preprint arXiv:2210.12079*, 2022.
- [65] OpenAI. Gpt-4 technical report, 2023.
- [66] Rohan Pandey. Echos of classical indian linguistics in natural language processing. Microsoft Artificial Intelligence Platform Talks, 2022.
- [67] Barbara H Partee. *Compositionality in formal semantics: Selected papers*. John Wiley & Sons, 2008.
- [68] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
- [69] The Univalent Foundations Program. Homotopy type theory: univalent foundations of mathematics. *arXiv preprint arXiv:1308.0729*, 2013.
- [70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

- [71] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [72] Matthew A Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T Rogers. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42–55, 2017.
- [73] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.
- [74] Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, 2021.
- [75] Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32, 2019.
- [76] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [77] Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. Learning relation alignment for calibrated cross-modal retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 514–524, 2021.
- [78] Bertrand Russell. *An inquiry into meaning and truth*. Routledge, 1940.
- [79] Shishir Rajan Saxena. *Linguistic and Phenomenological Theories of Verbal Cognition in Mīmāṃsā: A Study of the Arguments in Śālikanātha’s Vākyārthamāṭrkā-I and the Response in Sucarita’s Kāśikātīkā*. PhD thesis, University of Cambridge, 2019.
- [80] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.
- [81] Paul Smolensky, R Thomas McCoy, Roland Fernandez, Matthew Goldrick, and Jianfeng Gao. Neurocompositional computing: From the central paradox of cognition to a new generation of ai systems. *arXiv preprint arXiv:2205.01128*, 2022.
- [82] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

- [83] Paul Soulos, R Thomas McCoy, Tal Linzen, and Paul Smolensky. Discovering the compositional structure of vector representations with role learning networks. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 238–254, 2020.
- [84] JF Staal. Euclid and panini. *Philosophy East and West*, 15(2):99–116, 1965.
- [85] Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2034. URL <https://aclanthology.org/P17-2034>.
- [86] Ajinkya Tejankar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv preprint arXiv:2112.13884*, 2021.
- [87] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.
- [88] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.
- [89] Mycal Tucker, Peng Qian, and Roger Levy. What if this modified that? syntactic interventions with counterfactual embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.76. URL <https://aclanthology.org/2021.findings-acl.76>.
- [90] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [91] Bo Wan, Wenjuan Han, Zilong Zheng, and Tinne Tuytelaars. Unsupervised vision-language grammar induction with shared structure modeling. In *International Conference on Learning Representations*, 2021.
- [92] Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering. *arXiv preprint arXiv:2205.11501*, 2022.
- [93] Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. Sgeitl: Scene graph enhanced image-text learning for visual commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5914–5922, 2022.

- [94] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.
- [95] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [96] Zhengxuan Wu, Atticus Geiger, Josh Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah D Goodman. Causal distillation for language models. *arXiv preprint arXiv:2112.02505*, 2021.
- [97] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. *Advances in Neural Information Processing Systems*, 34: 4514–4528, 2021.
- [98] Charles D Yang. Universal grammar, statistics or both? *Trends in cognitive sciences*, 8(10):451–456, 2004.
- [99] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. Auto-parsing network for image captioning and visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2197–2207, 2021.
- [100] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9847–9857, June 2021.
- [101] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.
- [102] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021.
- [103] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bag-of-words models, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.
- [104] Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A benchmark for compositional visual reasoning. *arXiv preprint arXiv:2206.05379*, 2022.

- [105] Bryan Zhang. Improve MT for search with selected translation memory using search signals. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 123–131, Orlando, USA, September 2022. Association for Machine Translation in the Americas. URL <https://aclanthology.org/2022.amta-upg.9>.
- [106] Junchao Zhang and Yuxin Peng. Hierarchical vision-language alignment for video captioning. In *International Conference on Multimedia Modeling*, pages 42–54. Springer, 2019.
- [107] Xunjie Zhu and Gerard de Melo. Sentence analogies: Linguistic regularities in sentence embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3389–3400, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.300. URL <https://aclanthology.org/2020.coling-main.300>.
- [108] Zining Zhu and Frank Rudzicz. An information theoretic view on selecting linguistic probes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9251–9262, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.744. URL <https://aclanthology.org/2020.emnlp-main.744>.