Abstract vs Concrete

BMVC 2019 Submission # ??

Abstract

Properly learning to model information between language and vision is of paramount importance in multimodal deep learning. Though multimodal fusion methods and taxonomies grow increasingly more complex, they are largely inspired-by and adapted-from existing techniques e.g. novel application of attention. We draw inspiration from neurological insights in 'Dual-Coding Theorey' to model multimodal representations inline with how humans distinctly process and store 'abstract' and 'concrete' concepts.
We present a ...
We find that...

1 Introduction

Current neural network architectures are inherently inspired by the extraordinary capacity of the brain. Though much of this capacity is still unknown, advances in psychology and neuroscience have given us a strong understanding of the nature of the brain [16, 33, 46, 49, 50, 59]. Though modern artificial neural networks (ANNs) remain comparatively primitive, recent neurologically inspired designs show promising results [4, 32, 37, 45, 53]. However, insights for specifically multimodal processing remain largely unexplored. Strictly speaking, the multimodal family of bilinear pooling (BLP) models [9, 23, 24, 36, 38, 65, 66] evolved from earlier unimodal, vision-only bilinear models [42, 58] that were themselves reminiscent of the 'two-stream' model of vision [25, 43][1]. In reality, this neurological inspiration has been abandoned in the shift from vision to multimodal BLP as an equivalent 'two-factor/stream' model of vision and language has not been established or discussed. Instead, the motivation for BLP techniques is that 'higher-order interactions between text and image facilitate representations of fine-grained cross modal information'. This motivation is rather speculative and unsatisfactory, what makes a BLP-enforced 'higher-order representation' any better than a non-linear projection of feature concatenation? In parallel to promising work in recent surveys and taxonomies focus on explaining, motivating and categorising the nature of multimodal representations [6, 26, 67] (e.g. joint/co-ordinated representations), we explore text-image multimodal fusion inspired by 'Dual Coding Theory' (DCT). DCT [47] broadly considers the interactions between the verbal and non-verbal systems in the brain (recently surveyed here [48]) by way of 'logogens' and 'imagens' respectively, i.e. units of verbal and non-verbal recognition. Imagens may be multimodal, i.e. haptic, visual, smell, taste, motory etc (more formally, information is encoded in multiple ways). We see these concepts paralleled in multimodal deep learning, with textual features as logogens and visual (and sometimes audio) features as visual (or auditory) imagens. A key insight from DCT explores the different

[1] Can be thought of more explicitly as the 'shared terminal' hypothesis detailed in [43]

ways 'abstract' 'concrete' concepts (non-imageable and imageble) are stored and accessed by the brain, and the differences in cognitive processing (i.e. free-call) associated with either concept type. [12, 21] find evidence implying that abstract and concrete words initially activate similar brain regions and then later separate ones, with concrete concepts activating regions associated with visualisation (this makes intuitive sense). Most interestingly, [20] find evidence that, irrespective of the type of information, abstract and concrete concepts are stored in structurally different ways. Abstract concepts are represented in associative frameworks, near other concepts associated with it (but not necessarily similar in meaning) e.g. 'jury' and 'courtroom'. Concrete concepts appear to be more categorically organised i.e. stored in more rigid, semantically related networks. It is thought that all words are polysemous to some extent as their precise meaning changes in different contexts [27], inspiring speculation that a strict 'associative/categorical dichotomy' is overly simplistic [20]. Concepts of middling concreteness (e.g. nurse, chemistry) are thought to have both associative and categorical connections.
Motivated by these substantial insights, in this paper we ...
we find that ...
Our implementation is available on github [2].

2 Experimental Proposals

Note that the introduction is significantly more detailed than it will be in the final version to fully flesh out the story for you guys. In this section, we'll outline and visualise the experimental ideas we're currently considering.
DCT papers I have read in detail: [12, 20, 21, 31, 48]
Multimodal deep learning survey papers: I've read to gain inspiration across all multi-modal deep learning: [26, 67]. To a lesser extent, i must finish this [6].

2.1 Abstract and Concrete Word Lists For Ground Truth

There is a wealth of datasets giving the concreteness, imageability and other words norms that are separated by syntactic units (i.e. verbs, nouns, adjectives, adverbs etc...). See Section 4 for an (almost) exhaustive overview of datasets handling abstract or concrete concepts. We have our lists!

2.2 NEW Experiment 6: Bidirectional Mapping from Concrete to Abstract and Back Through Metaphorical Variety

As intuition would dictate, image representations of abstract concepts have been shown to be far more diverse i.e. the prototypicality of a concept [8]. However, there is ongoing evidence that abstract concepts are still grounded in the perceptual system (2003 Barsalou et al, and [30, 35]). Highly 'dispersed' abstract concepts can be imaged in many ways.
Our IDEA: Using the 'visual variety' precedent set by "Estimating the visual variety of concepts by referring to Web popularity", when training on an abstract concept that is thought to be dispersed or even metaphorical ("Formal Distinctiveness of High- and Low-Imageability Nouns: Analyses and Theoretical Implications"), we can actively train on the abstract metaphorical image representations.

[2]https://github.com/Jumperkables/a_vs_c

2.3 NEW Experiment 7: Reducing Abstract/Concrete Bias

This idea is relatively simple, emulate the model-agnostic RUBi (reducing unimodal bias) scheme with a concrete-vs-abstract alternative training scheme.

2.4 NEW Experiment 8: Abstract-Concrete Changing Priors Rearrangement

Another relatively simple idea. Following the VQA-CP (changing priors) scheme discussed before (i.e. make sure your training and validation sets have different statistical priors, so shortcut exploitations are punished at test time). We can reorient existing multimodal datasets such that the highly concrete/imageability biases that appear are separated. We can reverse this logic for metaphorical or abstract biases too.

THE FOLLOWING ARE THE OLDER EXPERIMENTS THAT HAVE BEEN UPDATED WITH REFERENCE TO DATASETS RECENTLY FOUND

2.5 Experiment 0: Modified Multiple-Instance-Learning

We count from 0 as is proper. [22] introduced me to multiple-instance-learning, which learns discriminative visual signatures for each word (surveyed here [15]).

The main idea: Bags of words are initialised. A bag is positive if any of its objects are present, and negative if none are present.

Our Use: This may be used to emulate associative/semantic interconnections. Bags of associated words or semantically similar words can be used generated for abstract/concrete words in particular. We could compare and contrast using *abstract/concrete aware bags* vs *normal, unguided bags* to ascertain if abstract-concrete aware processing is useful.

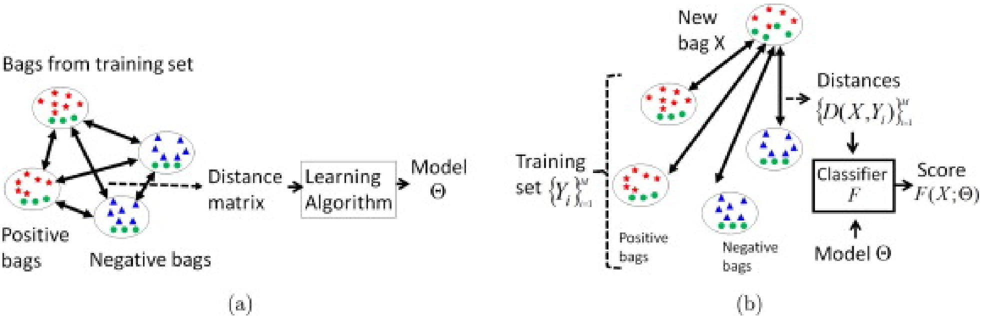

Figure 1: Multiple instance Learning. We can adapt this scheme to have abstract vs non-abstract concepts, or perhaps abstract vs concrete concepts.

2.6 Experiment 1: Unified VSE Relational Networks

Main paper is here [61]. Humans are able to establish accurate alignments between vision and language. We note that concrete concepts are heavily aligned with vision. Though visual concepts are aligned in language at different levels, i.e. objects, relations, sentences. Where

	Language	**Vision**
Objects	Noun-phrases	Visual objects
Attributes	Prenominal phrases	Visual attributes
Scene	Sentence	Image

Table 1: Breakdown that could be expanded upon.

each concept exists at different levels in language, a starting breakdown approximating this could be:

A noted problem in unified VSE is that its difficult to distinguish exactly what visual feature a linguistic unit is referring to (pictures of keyboards will often accompany monitors). Contrastive training (i.e. if language includes a clock, then one picture will have a clock, and one will not) to resolve referential ambiguities.

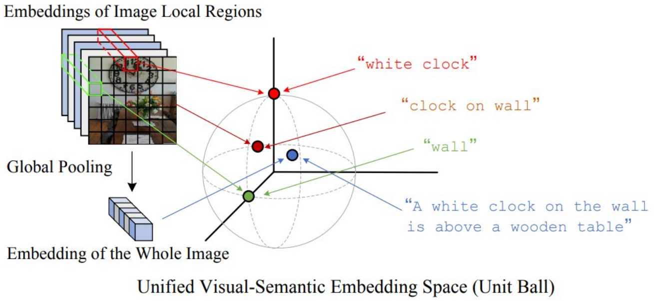

Figure 2: Unified VSE example.

Our Idea: Where unified VSE generates a joint embedding space of text and vision at the different levels and context thats they exist in, so too can we use unified VSE to capture the concrete and abstract elements of words at different levels, see again Figure **??**. Learn a more contextual joint embedding space, where we understand the contextual difference between 'building' the abstract concept, and 'building' the concrete example of the scene characters are in. This can be thought of similarly to the clock example in Figure 2.
Further explicit idea: Generate a separate middle, abstract and concrete embedding with respect to this scheme for each concept.

2.7 Experiment 2: Semantic and Associative Search and Concept Resolution

Consider the sentence "justice is done", this could be applied in context to many scenes. When parsing this sentence, if we find a very abstract word i.e. 'justice', then we can search the surrounding associative space of 'justice' for concrete words that could appear in a scene. We can perform this search by moving around association matricies for concepts provided by USF Free Association Norms dataset. We can follow CSLB cosine similarities for concrete concepts in turn.

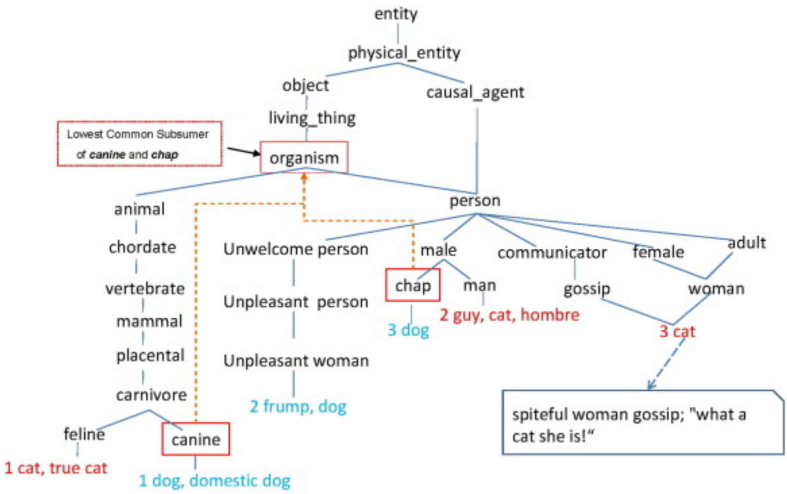

Figure 3: Example trees we could search or parse.

2.8 Experiment 3: Pointer-Generator

We can consider a pointer-generator attention mechanism reminiscent of multi-task learning models that Hudson has worked on [54].

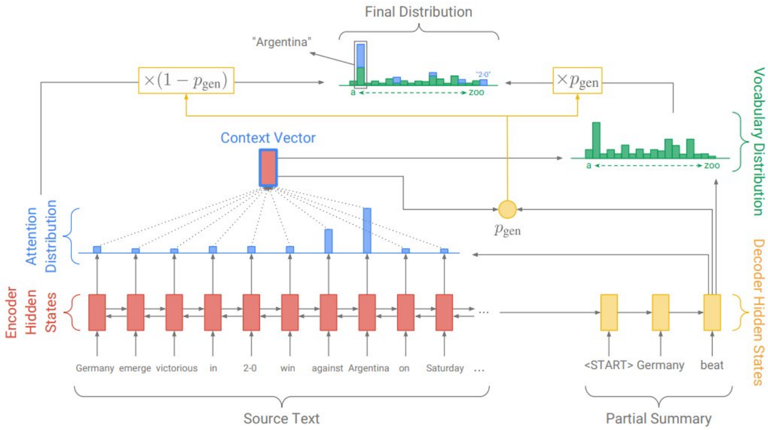

Figure 4: Pointer Generator mechanism learning end-to-end where to draw information.

We could create a network designed to learn, end-to-end, to draw more heavily from concrete embeddings and networks or from associative abstract networks where needed. A joint multimodal embedding EViLBERT (sense embedding) structure can be applied to this too.

2.9 Experiment 4: Neural Module Networks (NMN)

Neural Module Networks (NMN) [5] are in essence a neural network that trains different separate substructures for different purposes, and decides in forward propragation if a certain subnetwork is relevant to current processing.

- They basically apply attention across a bunch of neural network modules trained jointly end to end and would specialise in different things.
- Parse the input question using the Stanford Parser getting universal dependency representations
- Filter set of dependencies between 5w word and copula
- “Is there a circle next to the square?” -> is(circle, next-to(square))
- All leaves become find modules, all internal nodes become transform or combine depending on their arity and all root nodes become describe or measure depending on their domain.
- (Dodgey) They use a simple LSTM question encoder and think its good for 2 reasons:
  - A vaguer question understanding removes ambiguity between answers, i.e. is vs are
  - Allows them to capture ‘semantic regularities with missing or low quality image data’, i.e. guessing a bear is brown is reasonable but not green, (their explanation sucks)
- Some modules are updated more than others, so adaptive per-weight learning rates are best
- They introduce the shapes dataset.
- Performs especially well on questions answered by an object or an attribute.

This bears some relationship to ‘Neural Symbolic VQA’ [62].

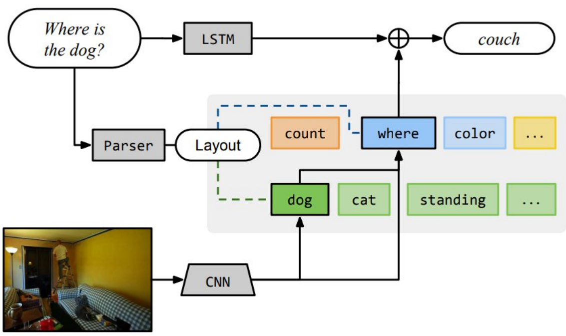

Figure 5: NMN overview. Note the network decides from language to use different actual network regions that have learned different functions.

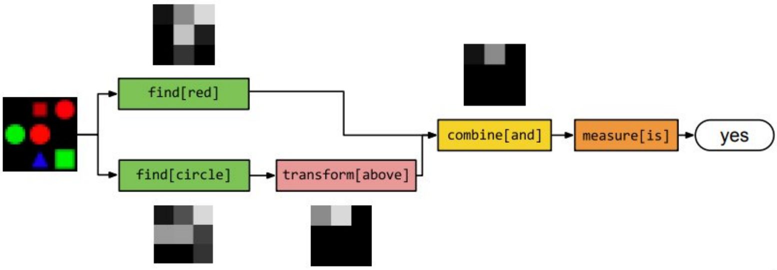

Figure 6: An explicit example of an NMN forward pass.

Our Idea: We can adapt NMNs to use explicitly abstract or concrete transformations on explicitly abstract or concrete inputs. Furthermore, we should create abstract noun/verb, and concrete noun/verb submodules, which we can train using the appropriate labels from the many datasets from Section 4

2.10 Experiment 5: Bottom-up and Top-Down attention

Bottom-up and Top-Down attention mechanism [4]. This paper is mainly inspired from [19] (and another paper cited in their main study) looking at attention in humans. This is a neurologically inspired piece of work that follows how the brain chooses to focus, and rapidly adjust its priorities when faced with new stimuli:

- Top-down control: I.e. cognitive brain to task, when our 'attentional set' is guiding our attention
- Bottom-up: When salient features (sensory stimulus of sorts) grab our attention, e.g. an alarm going off
- So bottom-up acts like a circuit breaker to the current attentional load. Switching focus to new salient images. This will be more pronounced in videos.
- So in this paper their attention mechanisms driven by non-visual or task-specific context as top-down (simple one pass attention model, more could be applied), and purely visual feed-forward attention mechanisms as bottom-up (Faster-RCNN).

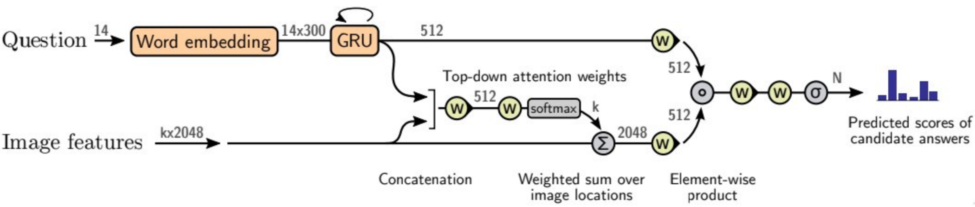

Figure 7: Bottom-up Top-down Attention using image information to adjust and refocus the text narrative in a 'circuit-breaker' manner.

Our Idea: We can repurpose concrete/abstract concepts to push the attentional set towards either concrete/abstract network structures appropriately.

3 Datasets: Concreteness/Association/Verbs

These datasets are about properties of concepts (words)

3.1 CSLB Norms

- 638 Concrete Concepts
- Semantic Properties for each (from human annotators)
- 'Production frequency' vector of a concept (Production Frequency. The total number of participants who gave a response that was mapped to the given feature label. All concept/feature pairs with PF > 1 are given)
- For each concept/feature pair:
  - Living / Nonliving
  - visual perceptual - other perceptual - functional - taxonomic - encyclopaedic
  - Concept name from McRae et al (2005)
  - Final normalised feature label
  - Production Frequency
  - A semi-colon-delimited list of the linguistic variation that was mapped to the given normalized feature label. Note that automatically re-written variations are not given. In particular, the following syntactic patterns are collapsed into a single feature:
    * does, can, may, might, third person singular of verb
    * Example: does eat, can eat, may eat, might eat, eats -> does eat
    * Singular and plurals are collapsed. Example: is found in gardens; is found in a garden -> is found in gardens
    * Variation in the use of articles is also not shown. Example: is found in a kitchen; is found in the kitchen; is found in kitchens -> is found in kitchens
    * Only linguistic variation in the raw responses relevant to each normalized feature is given.
  - Cosine similarity of features, including and excluding taxonomic features

3.2 USF Free Association Norms

- USF Free association standard norms (linguistic norms, i.e. concreteness, imageability etc...)
- word embeddings built from USF https://github.com/jocarema/Wan2Vec
- full information here: http://w3.usf.edu/FreeAssociation/
- 'Free association' measure of given words, i.e. given a word, whats the first one that comes to mind
- 5,019 words

- Work done by over 6000 participants
- nouns (76%), but adjectives (13%) and verbs (7%), and other parts of speech are also represented. 16% are identified as homographs (spelled the same but not pronounced the same)
- Actual content:
  - Appendix A: All of the normed words(cues) listed alphabetically, their responses(target and related information
  - Appendix B: All of the responses (targets) listed alphabetically, the normed words (cues) that produce them in free association and related information
  - N x N associative matrices showing connections among the associates of each normed word
  - Appendix D: All normed words (cues) and their idiosyncratic responses
  - Appendix E: Accessibility index: Responses ranked by how many normed words produce them as associates
  - Appendix F: Norms for rhyme, beginning stem cues, ending stem cues, beginning fragment cues, and ending fragment cues

3.3 "2014 Learning Abstract Concepts..."[29]

- They use USF concepts and rate their concreteness
- I cannot find this dataset but could request it
- Each USF concept used has also been ranked on a Likert scale 1-7 by a bunch of human annotators to get its concreteness
- Spearmen correlation between association scores and cosine similarity of vec reps
- They draw noun/verb relationships as they arent done from before, getting 4 lists of noun-noun, verb-noun etc..

3.4 Vinson Dataset: Semantic feature production norms for a large set of objects and events

- 456 words (169 nouns referring to objects,71 nouns referring to events,and 216 verbs referring to events)
- Types, (action verb, action noun, object ..)
- Semantic label (e.g. contact, change-state, noise, tool, action(body-action))

3.5 McRae Dataset

- 541 concepts (living vs non-living)
- concept names and frequencies (of collecting subjects)
- similaritiers between concepts
- Brain region labels and where on the display participants saw

3.6 SimVerb Dataset

- https://github.com/benathi/word2gm/tree/master/evaluation_data/simverb/data
- A dataset of verb similarity
- 3500 verb pairs
- score 0-6 (likert) projected to 0-10 to match other datasets
- Lexical relation types: "SYNONYMS", "ANTONYMS", "HYPER/HYPONYMS", "COHYPONYMS", "NONE"

3.7 A collecition of multimodal grounding for verbs

- Collection: https://public.ukp.informatik.tu-darmstadt.de/coling18-multimodalSurvey/
- Github: https://github.com/UKPLab/coling2018-multimodalSurvey

3.8 "Quantifying the visual concreteness of words and topics in multimodal datasets"

- Implementation of concreteness extraction (can be used for any dataset): https://github.com/jml
- Concreteness scores of topics and words in: COCO, Flickr, Wikipeida and British library sets

3.9 MT40k: Brysbaert et al 2014

- Concreteness ratings for 40,000 generallt known engish word lemmas (rating 1-5)
- 37,058 English words and 2,896 two-word expressions (such as zebra crossing and zoom in)
- Scores here: http://crr.ugent.be/papers/Concreteness_ratings_Brysbaert_et_al_BRM.txt
  - Concreteness scores (and normalised)
  - Marked for adjectives in dom_pos or subtlex

3.10 PYM: Paivio 1968

- 925 nouns
- ratings for concreteness, imagery, meaningfulness

3.11 CP: Clark and Paivio 2004

- An extension and alternative rating list of the above PYM
- 2,311 words

3.12 Toronto Word Pool

- Imagery and concreteness for 1080 words from thorndike-lorge word count
- Nouns, verbs, adjectives, adverbs, prepositions
- Small overlap with PYM

3.13 Newcombe

- 200 abstract and concrete particular words picked from TWP and PYM
- "Predicting Word concreteness and imagery" are currently the best bundle of concreteness and imagery datasets around
- They cite a bunch more smaller concreteness sources

3.14 "Metaphorical Sense Identification through Concrete and Abstract Context"

- An algorithm to classifying words as literal or metaphorical.
- evaluate this algorithm with a set of adjective noun phrases (e.g., in dark comedy, the adjective dark is used metaphorically; in dark hair, it is used literally) and with the TroFi (Trope Finder) Example Base of literal and nonliteral usage for fifty verbs

3.15 BabelPic Dataset

- An image/synset association dataset that focuses on NON-CONCRETE CONCEPTS
- Multimodal 'sense embeddings' generated encoding text and image

3.16 MRC Psycholinguistic Database

- An EXTREMELY EXTENSIVE AND USEFUL TOOL https://websites.psychology.uwa.edu
- concreteness, imageability, part of speech and much much more
- All this for words, word chunks and a lot more too. YOU SHOULD LOOK AT THE TOOL FOR THIS
- 150,837 words, 26 psycholinguistic and linguistic attributes

3.17 Cotese et al 2004

- Imageability ratings for 3000 monosyllabic words
- Psycholinguistic markers from experiments, i.e. reaction time for words

3.18 “Formal Distinctiveness of High- and Low-Imageability Nouns: Analyses and Theoretical Implications”

- High- and low-imageability nouns differed by length, etymology, prosody, affixation, phonological neighborhood density, and rates of consonant clustering
- Cannot find this dataset right now, must look again

3.19 “Estimating the visual variety of concepts by referring to Web popularity”

- For a given concept, this dataset has multiple images aimed at spanning the different kinds of images associated with that concept, i.e. the ‘visual variety’
- I cannot find their dataset

3.20 Battig Dataset

- Montague Categorized Word Norms (Battig) - This dataset, from Battig Montague (1968) comprises a ranked list of 5231 words listed in 56 taxonomic categories by people who were asked to list as many exemplars of a given category
- https://github.com/friendly/WordPools

4 Datasets: Non-linguistic

These are the non-linguistic related (usually image based) datasets that have been used in deep learning abstract-vs-concrete work. They may be useful to us.

4.1 ESPGame

- 100,000 images
- Each annotated with a list of lexical concepts that appear in the image (singleword concepts etc..)
- Used by [29] by appending concept tokens from the images to bags for frequencies

4.2 Google Syntactic N-Grams Corpus

- Syntactic Ngrams (counted from dependency tree fragments)
- 10 billion distinct items ‘covering a wide range of syntactic configurations’
- Syntactic Ngrams
  - Content-words and Functional-markers
  - Conjunctions and Prepositions
  - Multiword Expressions

- Nodes, arcs, biarcs, triarcs, quadarcs
- Ngrams of verbs and their immediate arguements
- Nouns and their immediate arguements
- Both of the above for the most popular words

4.3 TACOS Corpus: Grounding Actions in Videos

- 127 videos, each with 20 different text descriptions
- https://www.aclweb.org/anthology/Q13-1003.pdf
- Datset for grounding sentences describing videos
- Sentences and videos are aligned
- Paraphrases describing similar scenes
- Alignments for ‘low-level activites’, i.e. groundings for verbs

4.4 imSitu Dataset

CLIPPING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	VET
SOURCE	SHEEP	SOURCE	DOG
TOOL	SHEARS	TOOL	CLIPPER
ITEM	WOOL	ITEM	CLAW
PLACE	FIELD	PLACE	ROOM

JUMPING

ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

SPRAYING

ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

Figure 1. Six images that depict situations where actors, objects, substances, and locations play roles in an activity. Below each image is a *realized frame* that summarizes the situation: the left columns (blue) list activity-specific roles (derived from FrameNet, a broad coverage verb lexicon) while the right columns (green) list values (from ImageNet) for each role. Three different activities are shown, highlighting that visual properties can vary widely between role values (e.g., clipping a sheep’s wool looks very different from clipping a dog’s nails).

- Dataset of verbs and images of actions happening ‘in-situ’
- 500 verbs, 125,000 images

YOU CAN STOP READING HERE, ALL GOOD, ABANDON HOPE ALL YE WHO CONTINUE

Recent Reading

5 Kastner's Thesis

- Concepts
  - "Mental Image": Visual experience where the content does not directly relate to any afferent stimulus but is derived from working memory. (See [96][98])
  - Metrics to decide whether a concept should be concrete or abstract
  - Psycholinguistics is split into language production, comphrenesion and acquisition
  -
- Methods/results
- Motivation/Claims
  - Crawling across web and social media: Core assumption that the average mental image regarding words across society is reflected in the images available through web and social media
- Datasets/Tools
  - Imagability/concreteness scales [30,58,59,60]
  - LIWC [61], Empath [63] that connect words and language to motivation, thoughts, emotions and other sentiment-based numiercal ratings
  -
- Ideas i got
  - Other quantisations that aren't just visual variety/distance. Ambiguity etc... What is the relationship between ambiguity and abstractness?
  - Ablate across visual variety
  - The difference between first and second language learners. Perhaps fine tuning should follow this trend? I.e. train like a first language and fine-tune like a second one
  - Concreteness and imagability are different. Astrolabe is highly concrete but not imagable [97]. Consider [78]
  - Do abstract words really have broader physical characteristics just because thyre less defined
  - Figure out the true distribution of words via tasks. The core assumption of Kastner is insufficient. What if car is not an average truly but a subset. Narrow this down by regression across a task and let the model figure out for itself which average image it is (they did this)
  - The above point, Castner has though of this. Image tagging. Consider the different between relative and absolute measurements between concepts

- ESTIMATING CONCRETENESS AND ABSTRACTNESS PREDICTIVELY. COULD ALSO REGRESS WITH IMAGEABILITY AND CONCRETENESSSCORES BOTH

95 perception and cognition and how they talk to eachother, some others inbetween, [96] found both visual perception and mental imagery share the same neural structure

11 discuss imageability of verbs on grammar usage for different contexts. "It is considered to be used for syntactic as well as semantic processes in the human mind". There is a relationship on imagability of words to age of aquisition and reading comprehension [51,55]. MY IDEA. START WITH IMAGEABLE. FINETUNE WITH CONCRETENESS.

- Read papers from section 2.3.2
- Use the rebalanced version of imagenet and wordnet Kastner creates
- He does a lot of feature engineering, which is fine in motivation, but away from the transformer trend
- Not convinved by his results for abstrct vs concrete
- READ [8]

6 ASVD as suitable?

7 Deep Learning Abstract vs Concrete

The 10 papers i gathered:

- Learning Abstract Concept Embeddings from Multimodal Data (Since you probably cant see what i mean) [29]:
  - Motivation/Claims
    * They claim abstract representations may prove highly applicable for multitask/multi-domain transfer learning
    * Hill et al 2014 ridge regssion proposed so certain abstract concepts can be enhanced by multimodal models by combining ‘perceptual and linguistic input’. They improve on this
    * Inspired by the process of human language learning
    * Moderates training input to include more perceptual info about commonly occurring concepts and less about rarer ones
    * An updated process for integrating linguistic and perceptual info based on backpropagation
    * Propagates extra linguistic in inputs for concrete nouns to improve performance
    * Text-only abstract, perceptual gets contextual and ‘perceptual psuedo-sentence’ updates
    * Based on assumption that frequency in domain-general corpora correlates with likelihood of experiencing a concept in the real world (a few citations for this)

    * Guiding questions
      - 0) which model architectures perform best at combining information pertinent to multiple modalities
      - 1)which model architectures propagate info best
      - 2) Is it preferable to give all perceptual, or filter in some way?
      - 3) how much percept vs linguistic is best for various concept types
    * Cited, abstract concepts are generally more subjective and less reliable than concrete concepts
    * Like a language learner in that ‘once you encounter the concrete word, you start trying to formulate it in your head for a little and then continue’. I LIKE THIS
  - Datasets Tools (ALL EXTREMELY VALUABLE)
    * ESPGame dataset of annotated images
      - For each concept in ESP image, construct a bag-of-features, and share them between overlaps
    * CSLB concept property norms
      - Similar to previous McRae et al 2005 property norms
    - They convert properties to ‘lexical form’, is$_g reen->green.`bydoingthis,theytreatt$
    * USF free association gold standard
      - Each USF concept used has also been ranked on a Likert scale 1-7 by a bunch of human annotators to get its concreteness
      - Spearmen correlation between association scores and cosine similarity of vec reps
      - They draw noun/verb relationships as they arent done from before, getting 4 lists of noun-noun, verb-noun etc..
  - Methods
    * Model learns from target-word/context-word pairs. Selected in k context windows around each word
    * Maximise the log probabilities across all of these examples
    * Perceptual information is introduced whenever designated concrete concpets are encountered (VERY LIMITED). (They claim this has the effect of introducing more commonly experience concrete concepts and less from rarer ones) (fair)
    * Associative array of (typically) concrete words to perceptual features
    * Model starts training on language, when it finds a concept from concrete domain, it begins learning from sentence of alternating concept, sampled$_b(w), concept$

* They use alpha to scale the relative ratio of linguistic to perceptual features
* Free association scores as empirical measure of cognitive conceptual proximity

  - Results
    * They find more efficient multimodal combination than other models, giving more ability to model the USF free association gold standard (for concrete nouns)

        * Propagation of extra linguistic input 'can extend the advantage of multimodal approach to many more concepts than simple concrete nouns'
        * However, the benefit of adding erceptual inputs appear to decrease as the target concept becomes more abstract
        * For most abstract concept, language only is still SotA
        * They find a set of concepts that benefit or do not with perceptual inputs etc... (input wise analysis)
        * The embeddings they get have 'higher correlation to USR data regardless of perceptual input source'
        * They claim that the correlations seeming somewhat low are a consequence of how hard it is to model USF data
        * Concrete verbs are 69% better
        * Abstract verbs a little less
        * None of the results did good for abstract nouns. Implying that the info is so far removed that youre best leaving it to text only stuff right now. (wow)
        * They think their moderately worse performance comes from the more enforced inter modal dependence.
        * Their results "reflect a clear manifestation of abstract/concrete distinction, concrete verbs and nouns can be effectively represented from perceptual information sources"
        * "Clearly counterproductive" for abstract nouns, BUT NOT ABSTRACT VERBS
        * "Model learns higher quality representations of abstract verbs if perceptual input is restricted to concrete nouns than if none at all, or both conc noun and abstract verb". They claim this supports the idea of gradual scale of concreteness. Cos the abstract concepts representations are improved by information of concrete. This is overall a moderately flawed arguement IMO.
        * They say that alpha = 1 implies that concrete concepts should be given approximately equal weight from language and perception
        * Type findings
            · Type I) Concepts that can be effectively represented directly in the perceptual modality, generally concrete nouns or verbs benefit from their combination technique and give better muiltimodal features
            · Type II) Concepts including abstract that can be effectively represented but improve from joint learning. Type I learning helps these type II in their model
            · Type III) Abstract concepts (including nouns) that are best handled by language only. Multimodal stuff not helpful for them
    – Insights
        * Andrews et al 2009 is apparently motivating original human word learning
        * Visual grounding and semantic representation is well cited in this paper
        * Concept categorisation, Silberer and Lapata 2014, is a task (so is predicting compositionality)
        * They cite DCT

        - They cite 72% of noun and verb tokens are rated by human judges as more abstract that the noun ‘war’
        - Their scheme means that once a concrete concept is found, they update its embedding first with language, then with perception
        - They use CCA
        -
    - Ideas i got
        - What about regularly used concepts. **I SHOULD CROSS CHECK THE PROPORTION OF NOUNS IN USAGE VS THE STATIC OVERALL LIST TO DETERMINE IF CONCRETE CONCEPTS ARE REALLY USED IN DAY TO DAY THINGS AS MUCH AS THEY THINK**
        - Video-BERT adaption fo their training scheme would be a nice contiuations of their work
        - Edit this array: Associative array of (typically) concrete words to perceptual features
        - Generate associative structures like this: For each concept in ESP image, construct a bag-of-features, and share them between overlaps
        - Free association scores should drive our associative network
        - I like their language learner idea, once you hit the concrete, continue thinking about it for a while. Should adapt this directly

- 2014 Multimodal Models for Concrete and Abstract Concept Meaning [30]
    - Motviation/Insight
        - Motivated by Barsalou et al 2003 (word meanings are grounded in perceptual system) are recent works
        - But these only work for concrete nouns, and concrete nouns are a subset of all linguistic units
        - Paper questions
            - Which modality gives rise to different concepts
            - Can perceptual inputs be propagated from concrete to abstract words
            - Best way to combine these different info sources?
        - Concreteness is closely related to more functional lexical distributions (cited)
        - Grammatical role is a strong predictor of semantics (gildea and jurafsky 2002)
    - Methods
        - They check distributional features between nouns/verb etc because it is well known that word meaning can be inferred from nearby words in a corpora
        - Since words can function as more than one PoS, this variation could hold meaning information. Nouns shiver and walk are processes not entities. They count frequency of occurence with the PoS categories adj, adverb, noun, verb
        - Since Grammatical role is a strong predictor of semantics (gildea and jurafsky 2002), theycount freq of nouns in range of syntactics contexts and of verbs in one of the 6 most common subcategory-frame classes defined in Van de Cruys et al (2012). Their table 1

    * Make abs and conc sets
        - nouns and verbs from USF word pairs based on majority PoS
        - Abs conc ordering is drawn by ordering words according to concreteness and sampling from 1st and 4th quartiles ....
    * They create a proximity function to cosine similarity
    * Ridge regression for concrete nouns to perceptual
    * Their weight ngram idea contains many useful insights
- Datasets Tools
    * Google Syntactic N-Grams Corpus (has linguistic features)
        - Dependency tree fragments for 10bn words in English Google Book Corpus, used for distributional abstract/concrete analysis
    * ESPGame too
    * They provide USF scores from their website
        - This data reflects the cognitive proximity of concepts
    * McRae Dataset
        - perceptual information , properties of 500 concrete noun concepts
- Results
    * concreteness determins both which linguistic features are most informative and the impact of perceptual information
    * Ridge regression to propagate perceptual info from concrete nouns to more abstract concepts (better than some previous), and it works?
    * 'weighted gram matrix combination' combining reps fro distinct modalities that outperform alternatives when 'both are sufficiently rich'
        - It beats CCA
    * Linguistic features overall effectively reflect meanings of all concept types
    * Features encoding syntactic patterns are only valuable for abstract concepts
    * Perceptual input useful for concrete concept reps
    * results indicate that 3 feature classes convery distinct info
    * McRae data most valuable for concrete nouns and verbs, abstract nouns liked combination of ESP-game and McRae
    * They claim this result underlines the link between concreteness and cognition in the literature
    * One drawback of multiplicative fusion is the joining of source, i.e. what i hate about BLP
    * preceptual stuff can be successfully propagated from concrete to abstract concepts
- Ideas i got
    * Remember abstract words are more common
    * The connection between lexical function and concreteness suggests that an awarenss of concreteness could improve models that already use PoS distinctions
    * Follow their approach for constructing abstract and concrete sets

    * Spearmans p seen to be appropriate for free associations, use for us. Refer heavily to table 3
    * Harris 1954, the value of lexical co-occurence statstics in conveying word meaning is expressed in the well known distributional hypothesis
    * The importance of such features ..... -> arguement in favour of abs vs conc
    * They have found when multimodal models should or should not aim to distinguish them
    * Supports weakly the 2003 Barsalou et al idea that abstract concepts are still grounded in the perceptual subsystem

- 2017 Exploring multi-modal text+image models to distinguish between abstract and concrete nouns [39]
  - motivation/insights
    * Citations for grounding theory
  - methods
    * binary classifier regression to distinguish between abstract and concrete
    * Noun work only really here
    *
  - datasets/tools
    * Brysbaert et al 2014 [13] collecion of concreteness ratings for 40,000 english words
  - results
    * Seems very weak finding. They find thatboth text and image 'seem to provide reliable, non-complementary information to represent both abstract and concrete words'
    * They interestingly find more concrete than abstract nouns, at odds with Hill's earlier paper?
    * When trying to binary classify abs vs concrete, text features are shown to be slightly better
    * Combined are slightly better
    * I think specifically what theyre saying is that when trying to predict concreteness or abstractness, the words where the difference between the model predicition and human gold standards are very low contain a mix of concrete and abstract words, for both image heavy reliance and text heavy reliance. They say this implies that text vs images give no particular advantage to A or C. There are a few things wrong with this, what about general trends??
    *
  - ideas for me
    * There are lots of works distinguishing between abstract and concrete concepts
- 2018 Multimodal grounding for language processing [8]
  - motivation/insights

    * Discuss the benefits of multimodal grounding for language processing tasks and the difficulties with respect to cognitive models of human processing
    * cognitive theories for grounding distributional semantics (baroni 2016)
    * Inspired by chomsky 1986 mental models of language that dont directly incorporate perceptual information?
    * They want to see a more compositional structure of multimodal representations, beyond nouns and adjectives
    *
- methods
    * They focus on multimodal grounding of verbs which 'play a crucial role in the compositional power of language'
    * they propose classifying multimodal task wrt information flow between modalities
    * Concept representation, Projection, Grounding concepts
    * 4.3 discussion on various psychological findings about learning
- datasets/tools
    * Regneri et al 2013 [52] build a corpus that grounds descriptions of actions in videos. Better TVQA
    * imSitu datast of images depicting verbs and annotations which link the verb arguements to visual referents
    * SimVerb dataset
    * These guys get a dataset that illuminates the embodiment of verbs
- results (most of this is a review)
    * In multimodal processing, grounding is usually limted to concrete conepts leading to a reduction of referntial ambiguity
    * Bottom of section 4.1 they have inconclusive findings for if visual info actually helps concrete concepts etc..
    * their results on quality of verb representation indicate that one should directly obtain visual representation for verbs instead of projecting meaning
    * Where concrete words are captured more adequately by multimodal representations
- ideas for me
    * useful guiding citations
    * cites recent works that indicate processing a word activates areas in the brain that correspond to the associated sensory modality of its sdemantic categories, kick = motor cortex, cup = visual
    * This is another multimodal taxonomy paper
    * Pay attention to: Bottom of section 4.1 they have inconclusive findings for if visual info actually helps concrete concepts etc..
    * Midway section 4.2, theres a vey interesting discussion on howabstract concepts may move to the metaphoric side of things. 'time is a stream'
    * Combining complementary informatio 5.1. is an extremely useful section and should be looked at again

    * It is argued that (Bruni et al 2014) highly relevant visual properties are often not represented by linguistic models because theyre too obvious to be explicitly mentioned in text
    * Collel et al 2017 consider verbs (alongisde hill, potentially add this to the paper)
    * BIG IDEA. WE WANT SKETCHES OF METAPHORS FOR ABSTRACT PHRASES BIG BIG BIG BIG BIG IDEA. Along with a textual descrption of why these images represent that using CONCRETE WORDS TO REDESCRIBE THEM. A DOUBLE GROUNDING EFFECT. This is explored in the measure of prototypicality of a concept as measured by image dispersion scores (how representative an image is of a category)

- Quantifying the Visual Concreteness of words and topics in multimodal dataset [28]
  - motivation/insights
    * algorithm for computing concreteness of words and topics in MM datasets
    * intuitively, a visdually concrete concept is one associated with locally similar sets of images
    * ‘Allowing concreteness to be dataset-specific is an important innovation because concreteness is contextual’
    * ‘readily scalable method’
    * they do still expect some correlation of concreteness and frequency (gorman 1961 citatation)
  - methods
    * Algorithm to compute concreteness of words and topics in MM datasets
    * ‘predict the capcity of ML algorithms to learn text/visual relationships’
    * For a fixed
    * each image is associated with discrete words and tags
    * measure how clustered a word is in image space: measure how often images are associated with a given word
      · MNI, mutually neighboring images score
      · normalise for infrequent words
    * this is also extended to contiuous topics
    * resnet imagenet
    * their null hypothesis is that concreteness is just measuring frequency. this is not fair because they scale down frequency?
    * calculated the correlation between Flickr and COCO to assess concreteness across datasets
    * they calculate joint embedding space, using a specific task because joint embeddings is often a ’first step’
    * they map image features to text features throug linear transform, with equation 4 minimised
    * negative sampling
    * they define retreivability, you might expect dog to be more retrievable than beautiful

        · retrievability is higher if instances associated with a concept are more easily revtrievd as measured by thei metric
  - datasets/tools
    * their datasets are imagey and texty stuff
    * They use USF, spearman correlation again
    *
  - results
    * concrete concepts easier to learn
    * Lots of algoritshm they look at have the similar failure cases
    * 'The precise positive relationship between concreteness and performance varies between datasets'
    * concreteness varies intuitively through topics between datasets
    * they observe moderate-to-strong correlation between infrequency and concreteness, (this is similar to previous findings)
    * direction of text-to-image and vice vera mappings matter
    * strong correlation between retrievability and concreteness, which gives strong evidence
    * there is little correlation between frequency and retrievability, implying concreteness isnt measuring frequency
    * evidence that concrete concepts are easier for classification tasks
  - ideas for me
    * follow this paper's intuitition to create abstract and concrete semantic spaces as planned from before
    * concrete-vs-abstract aware class performance of used datasets a-priori (curiculum learning)
    *

- 2019 Predicting word concreteness and imagery [17]
  - motivation/insights
    * bad assumption off the bat "we assume that concrete nouns occur in other contexts than .... words than nouns with a low imagery" (ok not exclusive, just trends and im fine)
    * noted by rabinovich et al 2018, certain suffixes can be important in determining concreteness
  - methods
    * regression model on 7 datasets, concreteness and imagery values can be predicted with high accuracy
    * they control for suffixes cos like '-ness' changes nouns to adjectives etc.., its very cool
    * they use an SVM
    * they excluded words from intersections of datasets
  - datasets/tools

    * MT40K? Brysbaert et al 2014
    * PYMc (paivio 1968, concrete)
    * PYMi (imagery)
    * CPa clark and pavio (2004)
    * CPe same
    * TWPifriendly et al 1982 (TWP) (HAS LOTS OF GOODIES)
    * TWPc same
    * Newcombe (newcombe et al 2012)
    * they create ‘training corpus’
    * section on ‘further sources for concreteness’
    * ukWaC, ferraresi et al 2008
  - results
    * very good at predicting 1-5 scores for concrete/abstract
    * fastText better than googlenews
    * adding suffix and POS increases performance slightly for the better fastText
    * correlations are much higher
    * spearman coefficient is best feature combination
    * since they train on concreteness values (? but they also have 3 datasets of imageability), their prediictably lower imagable scores are a thing
    * they didnt find patterns for differences in the predictions
    * they use their findings to conclude that concrete words and abstract words appears in different contexts. hmmm
  - ideas for me
    * 2018 overview of datasets, approaches etc very useful
    * Brybaert et al 2014 found that subjects largely rate haptic and visual experiences, even when explicitly asked to take into account experiences involving any sense
    * it was shown that 5 is the max number of categories humans can work with reliably (citation?)
    * noted by rabinovich et al 2018, certain suffixes can be important in determining concreteness
    * they suspect they will need to combine concreteness with other relevant measures

- Visually Grounded Neural Syntax Acquisition [55]
  - motivation/insights
    * learn syntactic structures and representations with explicit supervision
    * better syntactic structure leads to better representation of constituents which them lead to better alignement between vision and language
    * at test time no images paired with text are needed
    * they define concreteness for spans instead of words
  - methods

    * constituency parse tree of text, recursively compose representations for constituents and match them with images
    * define concreteness of constituents by matching their scores with images
    * for their text representations, sequential units are made, a score is made , 2 of the n are selected and combined, leaving n-1 for the next time step
    * during the matching with images they ignore the tree structure of the parse and index them as a list of constituents
    * cosine vector alignment of text and visual into a joint space
    * they optimise their constiuency maker with a hinge triplet loss
    * they also discourage abstract things from relating, read more about it
  - datasets/tools
    * multi 30k dataset
    * Turney et al 2011
    * Hessel et al 2018
    *
  - results
    * their concreteness definition scales well with linguistic definitions
    * their approach is more stable to random instatiation than older ones
    * best f1 scores against gold parse trees
    * model is substantially better with noun phrases and prepositional phrases
    * more efficient usage of text than previous approaches
    * easily extended to multiple language
    *
  - ideas for me
    * many citations for concreteness data (unimodal and multimodal)
    * they are inspired by semantic bootstrapping (children squire syntax by first understanding the meaning of words and phrases and linking them with the syntax of words (Pinker 1984))
    * definitelky consider their parser holy christ i think its cute as f
    * they think more complex GRU encoders tend to focus too much on cats in a caption about cats for exmaple
    * They deal with head-inductive bias, making you treat the ordering of adjectives and nouns properly
    * build structural spaces

- 2020 BabelPic, a Multimodal Dataset for Non-Concrete Concepts [14]
  - motivation/insights
    * previous image datasets focus on concrete concepts
    * build starting from concepts related to events and emotions cos thats whats popular with the ML bois
  - methods

    * Propose a non-concrete concept dataset that is handpicked by cleaning image-synset association from LKB
    * Vision-language pretraining models
        · Reduce the classification task to VQA 'yes/no' essentially
        · VLP model pretrained on conceptual captions dataset
        · Faster-RCNN 100 detections per image
    * They create the gold dataset by selecting a set of NC synsets from WordNet 'on the basis of their paradigmatic nature and relations in the knowledge base'. Gather corresponding images in BabelNet, then manually validate the synset-images mapping
        · They dont allow an image to be mapped to more than one concept and vice versa
        · there are more specifics that should be looked at again if you consider this
    * Negative instances in their dataset
        · **Sibling:** there exists a synset s2 s.t s and s1 are connected to s2 by a hypernymy relation
        · **Polysemy:** both s and s1 contain the same lemma
        · **Unrelated:** no relation connecting in babelnet
    * VERY GOOD, THEY FORCE VALIDATION AND TEST SETS TO CONTAIN INSTANCES REFERRING TO SYNSETS THAT ARENT PRESENT IN TRAINING SET. GOOD BOIS
    * Evaluate performance on different types of negative instances
  – datasets/tools
    * BabelPic, created from BabelNet Lexical Knowledge Base
        · built by manually validating associations available in BabelNet
    * MultiSense dataset
    * VerSe dataset (gella et al 2016)
    *
  – results
    * pretrained language-vision systems can be used to further expand the resource by exploiting natural language knowledge from the LKB
    * High performance on zero-shot classification
    * F-VLP (on VQA 2.0) is most stable for the task
    * both are robust to zero shot learning
        · F-VLP in particular can verify associations between unseen synsets and images with 77.67% precision
    * For analysis on negatives instances:
        · unrelated synset-image pair are correctly classified well
        · sibling is more difficult
        · humans have a tough time between dissapointment and boredom too tbf
        · Polysemy hardest

  - ideas for me
    * Use vision-language pretraining model (Zhou et al 2020)
    * Use a reducing concrete-bias idea (Like rubi) (The tencent ML-Images dataset focuses on concrete, we could us this to reverse train etc..)
    * changiong priors approach/rearrangement to any dataset between val/training
    *

- 2020 Predicting the concreteness of German words [18]
  - motivation/insights
    *
  - methods
    * These guys create a conjoined dataset of German things
    *
  - datasets/tools
    * MRC dataset psycholinguistic database (concreteness values)
    * Kendalls tau
    * pearson correlation
    * Web Word Norms (WWN) (German?)
    * Leipzig affective word norms(German!)
    * Berlin word norms (German!)
  - results
    * No difference between their regression and older model
  - ideas for me
    * They cite things for metaphor detection
    * 2020 updated related works including (They cite very useful things in general)
      · Adopting concreteness value from similar, related and neighbouring words synonyms and hypernyms of wordnet sentence level concrete assignment
      · Identifying a dimension in word embeddings that correspond to concreteness Using SVD, Hollis and Westbury found a dimension of word embeddings that corresponded to concreteness
      · traiing regressing models on features of words Paetzold and specia 2016, regression model to predict 4 word norms, including concreteness

- Estimating the imageability of words by mining visual characteristics from crawled image data [35]
  - motivation/insights
    * previous works used visual variety trying to align things with their representation in real life
  - methods
    * Mean Absolute Error (MAE) is very useful, they argue to capture trends of prediciton (high vs low imageability)

- datasets/tools
    - 10 [36] imageability datasets
    - ∗ Sentiment evaluation
        - · LIWC, EMPATH
    - ∗ Tehir previous dataset [24]
- results
    - ∗ their eigenvalue matricies can get very low
    - ∗ low level features better for abstract words, high level better for concrete, hmmmmmmmmm
    - ∗ a very loaded results section
    - ∗ the error for abstract words is significantly bigger than concrete (as expected)
- ideas for me
    - ∗ They have a good tool to help get data
    - ∗ think of this, i.e. abstract things are likley imaged in lots of different representations? (flawed)
        - · Expected concrete things to be highly imageable across all images and abstract to be low
    - ∗ Zhang et al [52] look at parallel equivalent, non-equivalent, non parallel stuff in slogans etc...
    - ∗ Give the datasets for 'need' or 'challenge' to people without the label, see what they call it, see how that relates and generate negative samples with them
    - ∗ reread all of the tables

8 Cognitive Load

Cognitive and Perceptual Load, Probably different things.
There was a debate between early and late processing
-Early=focused attention ignores early processing of perceptual information
-late=all of the early info are processed, but theyre ignored later on in the later postperceptual processes

- (1994) Lavies OG PLT [40]
    - They say you can resolve the debated 'concerning the locus of attentional selection' by specifying the conditions under which early selection is possible
    - Supports that 'clear physical distinction between relevant and irrelevant information is not sufficient to prevent irrelevant processing'
    - Locus of selection means 'where selection happens' i.e. early or late
    - Selective set experiments along with a few other named ones are used to gauge the reality of attention and perception
    - The clear physical distinctiveness of relevant information has proved to be insufficient for early selective processing. They say that although its important for processing priority, it cant itself prevent processing of irrelevant information

  - Resources as an internal input essential for processingbut limited and shared across tasks
  - They call focused attention as trying to ignore some stimuli or aspects of it defined as irrelevant
  - They propose that alongside previous upper limit constraint on resources, there is also a lower limit one
  - They focus on perceptual distinctiveness by location
  -

- (2004) Load theory of selective attentional cognitive control [41]
  - Try to reconcile early and late arguement with 2 different mechanism proposals
  - They 'demonstrate' that high perceptual load reduces distractor interference, working memory load of dual-task co-ordination load increases distractor interference
  - They say these findings suggest 2 selective attention mechanisms:
    * One prceptual selector to serving to reduce distractor perception in situations of high perceptual load that exhaust perceptual capacity in processing relevant stimuli
      · They say its passive because its just ignored becuase there isnt enough resources
    * A cognitive controller that reduces interference from perceived distractors as long as cognitive control functions are available to maintain current priorities (low cognitive load)
      · They say the second one is more active cos there are resources availible in low load
      · They say that, contrary to predicted effect of perceptual load, high load on these controlling higher cognitive functions means they dont have capacity to actively regulate distractors anymore, and thus distractors increase
      · There were some experiments showing that with distractors that are incongruent (vs congruent and no response), congnitive processing slowed, implying that even though its ignored its still being noticed somewhat. sometimes even where the distractors were 'clearly separated from the target'
    * Therefore, different types of loads should effect these two things differently and thus they can be dissociated. Fair enough. Idk yet

- (2013) Conceptual and methodological concerns in the theory of perceptual load [10]
  - They aregue perceptual load is circularly defined
  - Its fuzzy definition overlap with cognitive load and sensory load
  - Its argued PLT is contained with working memory ablations

9 Extra Notes

ToDo:
-How to collect own multimodal dataset
-DCT papers
-Finish reading BERT
-Related works for visual modelling
-Get a list of abstract and cncrete words
-Multiple instance learning for associative/semantic interconnections
-MY NOVEL CONTRIBUTION I NEED ONE
-One could consider the closest work to ours as alignment of verbal and non-verbal concepts done properly?? Maybe not actually
-Abstract and concrete being on a continuum could making using 2 embeddings a good natural fit

10 Multimodal Tasks, Works and Notes

- Standout notes for some Tasks
  - Captioning
  - Visual Grounding
    - There is knowledge base VQA which could help me use relational things
  - Image-text matching
  - Phrase localisation
  - Text-to-image synthesis
    - Stacked GANs improve on GAN-INT-CLS
    - Hierarchically nested discriminator GANs
    - Obj-GAN focuses on improving generation with and object-wise discriminator
  - Visual Reasoning
    - The task of in general learning things about visual stuff, could be a subsection of VQA for example
    - Neural Module Network [5]: Specify a framework of modular, composable, jointly-trained neural networks. Think about different questions, some may be simply getting an object location, i.e. ‘where is the truck?’, some may require multiple glimpses etc..
      - They basically apply attention across a bunch of neural network modules trained jointly end to end and would specialise in different things.
      - Parse the input question using the Stanford Parser getting universal dependency representations
      - Filter set of dependencies between 5w word and copula
      - “Is there a circle next to the square?” -> is(circle, next-to(square))
      - All leaves become find modules, all internal nodes become transform or combine depending on their arity and all root nodes become describe or measure depending on their domain.

      - (dodgey)They use a simple LSTM question encoder and think its good for 2 reasons: A vaguer question understanding removes ambiguity between answers, i.e. is vs are   Allows them to capture 'semantic regularities with missing or low quality image data', i.e. guessing a bear is brown is reasonable but not green, (their explanation sucks)
      - Some modules are updated more than others, so adaptive per-weight learning rates are best
      - They introduce the shapes dataset
      - Performs especially well onquestions answered by an object or an attribute
    - Neural Symbolic VQA [62]
      - Structural Scene representation from image
      - Program trace from question
      - Then execute program on scene to get answer
      - 'Fully disentangles vision and language understanding from reasoning'

- Multimodal Transformers
  - 2019 Multiview [64]
    - multimodal transformer for image captioning.
    - Focuses on intra modality relations alongside inter modality.
    - By self attentiona and coattention, stacking attention blocks.
  - 2019 Unaligned [60]
    -
  - 2019 LXMERT [57]
    - Learn vision-and-language connections
    - Object relationship encoder
      -
    - Language Encoder
      -
    - Cross-modality encoder
      -
  - Video BERT [56]
    -
- Multimodal Surveys
  - Deep multimodal representation learning: A survey [26]
  - Multimodal Intelligence: Rep learning, Information fusion and applications [67]
    - Page 3: "by assuming the corresponding representations to have similar neighbourhood structures across modalities, the representationn of a concept with zero training sample in one modal can be found based on its representations grounded in other modalities which have training datae". - Interesting, but this a strong assumption that doesnt necesarily mesh with DCT. e.g. closest word vectors can be retrieved as labels by projecting image into text space, But this maybe be only for concrete things.

    *

- Important Papers
  - From captions to visual concepts and back [22].
    * Introduced me to multiple instance learning, which learns discriminative visual signatures for each word. Surveyed here [15]. I.e. a bag is positive if any of its objects are present, and negative if all are present. This may be used to emulate associative/semantic interconnections.
    * Multiple instance learning can be useful for me.
    * Word detection, Generate word scores, Generate the caption from them.
  - Deep fragment embeddings for bidirectiona image-sentence mappings [34].
    * Embed fragments of images and sentences into joint embedding space
  - Unified Visual-Semantic Embeddings: [61]
    * Unifies embeddings of concepts at different levels, objects, attributes, relations, full scenes.
    * This is defnitely the closest piece of work to ours that ive seen
    * They use a bidirectional margin-based ranking loss.
    * They use an off shelf semantic parser
    * "For simplicity we only consider single-word nouns as adjectives and single-word adjectives for object attributes"
    * So for each word in vocab they make a semantic embedding and 'modifier semantic embedding', and combine them differently for nouns and adjective-noun pairs.
    * They align vision and language in the unified space using contrastive learning on different semantic levels i.e.

- Plan
  - Getting a list of abstract or concrete words:
    * Extract object labels from detectors
    * Look again for a centralised list
    * Learn some stronger linguistic things (i think theyre called semantic components), i.e. noun-phrases relate to whatever (see unified VSE paper for good details).
    * Content and functional morphemes may help breakdowns and parsing.
    * Check BERT's vocbulary and cross refernce with popular lists from internet??
    *
  - Experiments:
    * A modified MIL learning scheme have bags of semantically or associatively similar things between abstract and concrete and ablate performance. GET CODE FROM THIS.

        * Unified VSE, the closest thing to what i have, experiments on that. GET CODE FROM THIS. Edit the relational tree structure they have and experiment in that vein.
        * For each word, similar to Unified VSE, initialise an abstract embeddings and a concrete embedding and learn those and omfg that would be so amazing OMG OMG OMG AN IDEA IS COMING TOGETHER. Cite the 2005 paper and supporting that finds these are on a spectrum.
        * MY idea: "justice is done". Justice is abstract, i can search the surrounding semantic space for justice for a close concrete word. Then assess if the space of semantically close things fits. I COULD MODEL THIS AS A SEARCH SPACE.
        * Think of a combination scheme that would force concrete embeddings down for abstract words and vice versa. Is concatenation enough??
        * Neural Module Networks could be repurposed with abstract/concrete things in mind. Perhaps a describe module would fit perfectly for us.
        * Use a Stanford Dictionary Parser to get out entities and their objects.
    - Things to note:
        * VSE Stuff
            · If i do Unified VSE style stuff, straightforward use of the sentence embedding is vunerable to adversarial attack, a smaller set of semantic components appears in captions. This is 'alleviated' by 'enforcing coverage of the semantic components appearing in the sentences' i.e. combine the sentence representation with an explicit bag-of-components embedding that aggregates all components of a sentence to stop it ignoring anything automatically.
            · Should probably consider negative sampling, Unified VSE use nouns, attribute nouns, relational triplets and sentences.
            · To consider image fragments they generate a relevance map of 7x7 image regions.
            · They do object, attribute and relational attacks.
        * Attention
            · Bottom-up and top-down attention mechanism [4]. Following this paper [19] (there is another check the paper). Another neurologically inspired piece of work. Top-down control: I.e. cognitive brain to task, when our 'attentional set' is guiding our attention Bottom-up: When salient features (sensory stimulus of sorts) grab our attention, e.g. an alarm going off So bottom-up acts like a circuit breaker to the current attentional load. Switching focus to new salient images. This will be more pronounced in videos.

Dual coding theory (DCT) [47] broadly considers the interactions between the verbal and non-verbal systems in the brain (recently surveyed here [48]). DCT considers verbal and non-verbal interactions by way of 'logogens' and 'imagens' respectively, i.e. units of verbal and non-verbal recognition. Imagens may be multimodal, i.e. haptic, visual, smell, taste, motory etc. We should appreciate the distinction between medium and modality: image is both medium and modality and videos are an image based modality. Similarly, text is the medium through which the natural language modality is expressed. We can see parallels in multimodal deep learning and dual coding theory, with textual features as logogens and visual (and sometimes audio) features as visual (or auditory) imagens. There are many insights from DCT that could guide and drive multimodal deep learning: **I)** Logogens and imagens are discrete units of recognition and are often related to tangible concepts (e.g. 'pictogens' [44]). This may imply that multimodal models should additionally focus on deriving more tangible features i.e. discrete convolution maps previously used in vision-only bilinear models [42] as opposed to ImageNet-style feature vector more commonly used in recent BLP models and attention modules could be used to better visualise these learned relations. **II)** Multimodal cognitive behaviours in people can be improved by providing cues. For example, referential processing (naming an object or identifying an object from a word) has been found to additively affect free recall (recite a list of items), with the memory contribution of non-verbal codes (pictures) being twice that of verbal codes [**?**]. [7] find that free recall of 'concrete phrases' (can be visualised) or their constituent words is roughly twice that of 'abstract' phrases. However, this difference increased six-fold for concrete phrases when cued with one of the phrase words, yet using cues for abstract phrases did not help at all. This was named the 'conceptual peg' effect in DCT, and is interpreted as memory images being re-activated by 'a high imagery retrieval cue'. This may imply that future networks could improve in quality by focusing on learning referential relations between 'concrete' words and images and treat 'abstract' words and concepts differently. **III)** [11] explore the differences in student's understanding when text information is presented alongside other modalities. They argue that when meaning is moved from one medium to another semiotic relations are redefined. This paradigm could be emulated to control how networks learn concepts in relation to certain modal information. **IV)** Imagens (and potentially logogens) may be a function of many modalities, i.e. one may recognise something as a function of haptic and auditory experiences alongside visual ones. We believe this implies that non-verbal modalities (vision/sound etc..) should be in some way grouped or aggregated, and that while DCT remains widely accepted, multimodal research should consider 'verbal vs non-verbal' interactions as a whole instead of focusing too intently on 'case-by-case' interactions, i.e. text-vs-image and text-vs-sound. Recently proposed computational models of DCT have had many drawbacks [48], we believe that neural networks are a natural fit for modelling neural correlates explored in DCT and should be considered as a future modelling option. Of these insights, **we are most interested in exploring Abstract-vs-Concrete aware fusion techniques.** Our current idea is to contextually identify concepts in text that are abstract or concrete, and dynamically backpropagate contributions from a joint feature space if and only if the concept is deemed 'concrete'. We believe this narrowed focus will stabilise learning by *focusing on appropriate and realistic interactions* and filter out unhelpful noise.

11 Related Works

In this section we summarise the most related works in deep learning.

11.1 Neurologically Inspired Multimodal Deep Learning

- Two-stream model of vision.(adapt from 21 month review and put it in here).
- Capsules

- Bottom-up and top-down attention mechanism [4]. Following this paper [19] (there is another check the paper). Another neurologically inspired piece of work.
  - Top-down control: I.e. cognitive brain to task, when our 'attentional set' is guiding our attention
  - Bottom-up: When salient features (sensory stimulus of sorts) grab our attention, e.g. an alarm going off
  - So bottom-up acts like a circuit breaker to the current attentional load. Switching focus to new salient images. This will be more pronounced in videos.
  - So in this paper their attention mechanisms driven by non-visual or task-specific context as top-down (simple one pass attention model, more could be applied), and purely visual feed-forward attention mechanisms as bottom-up (Faster-RCNN).
- Paper 150 from 2020 visual modelling survey [32]
- Supervised learning based on temporal coding in spiking neural networks
- [37]. Rate vs temporal deep nn

11.2 Deep Learning

Similar tasks include visual grounding.

11.2.1 Visual Grounding

Visual grounding seeks to identify or 'ground' an object from text in an image. Visual grounding literature includes and is not limited to:

- Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding
- Countering Language Drift via Visual Grounding
- Interpretable Visual Question Answering by Visual Grounding From Attention Supervision Mining
- Learning to Assemble Neural Module Tree Networks for Visual Grounding
- Finding "It": Weakly-Supervised Reference-Aware Visual Grounding in Instructional Videos

11.2.2 Part-of-speech Tagging

Part-of-speech tagging describes the process of accurately identifying the syntactic role of each word in a given text or speech. For our purposes, we may for example assume that verbs and nouns are more likely to relate to 'concrete' concepts than abstract ones. Thus we may consdier part-of-speech paradigms in thsi work.

12 Dual Coding Theory

(Add in the original and survey references for DCT here.) The reading list for me to currently consider for abstract-vs-concrete currently includes:

- (1999) Dual coding, context, availability and concreteness
  - They compare the 2 theories and at this point in time say that neither theory can fully explain the concreteness effect and they do it by way of semantic processing
  - Context availability has a big fuck up not controlling for the effects of imagery on their self resported context availability ratings.
  - The follow up experiment for context availability demonstrate that sufficient support from context elimibates the concreteness advntae, but im not sure thats powerful enough of a claim.
  - if you support abstract words in context hey become as powerfull
  - However dual coding theory doesnt necesarily claim away all this and minor alterations to DCT could account for context availability
  -
- (2005) Distinct brain systems for processing concrete and abstract
  - Abstract
    * Abstract words activated left inferior frontal regions linked with **phonological and verbal working processes**. Almost only left hemisphere for abstract
    * Concrete use bilateral associations
    * Overlapping but distinct brain regions
  - Intro:
    * Some theories Meaning of words partly involves retrieval of other closely associated words (Deese 1965, Noble 1952)
    * Other theories, knowledge about word meaning is largely separate from language involving sensory and motor images through perceptual experience (James 1890, Wernicke 1874)
    * Concrete words are better everything (see introduction)
    * Split brain and brain lesioned patients have a good time with concrete. I'm not sure its true?
    * DCT is still disputed, functional imaging experiments failed to provide evidence for right hemisphere involvement in concrete word processing
    * Concrete word advantages are not always seen in all tasks
    * Contrasting DCT there is 'context availability' model. Posits one system for accessing meaning of abstract and concrete words?
    * **Their Experiments use** Functional MRI during lexical decision making task (Binder, McKiernan et al 2003), brain responses to words are compared with closely matching nonwords while participants classify them as words or nonwords. A further comparison was made between responses to concrete and abstract nouns.

    * Bilateral network of association and posterior multimodal cortices activated during processing of concrete *concepts*. Strongly left lateralised network activated for abstract *nouns*. Firm evidence for DCT
- **Results**
    * Stimuli present in a random order, variable interstimulus interval.
    * Planned constrast show: Reliable advantage of concrete over abstract, and concrete over nonword, and abstract over nonword, no differences between concrete and nonword items or between abstract and nonword items. (idk these 2 sentences)
    * fMRI blood oxygenation level dependent conducted by modelling the indivdual words and nonword trials as discrete events => they could be discrete events entirely
    * Concrete AND abstract words activated a number of common areas in left hemisphere relative to nonwords. Including left angular gyrus, middle and inferior temporal gyri, the dorsal prefrontal cortex, (these comprise a network of cortical regions closely associated with semantic processing).
    * All except for left middle and inferior temporal gyri, activation in these regions appeared more extensive for concrete-nonword contrast.
    * Areas activated only in concrete-nonword contrast include the left posterior cingulate grus and precuneus, left hippocampus and para-hippocampus, right angular gyrus, right supoerior frontal gyrus and right superior cingulate gyrus.
    * Direct comparison between concrete and abstract words stronger activation for concrete bilaterally in angular gyrus, posterio cingulate gyrus, precuneus and left dorsal prefrontal cortex. And for abstract more in inferior front gyrus, premotor cortex and dorsal temporal pole.
    * Bilateral activations occured in the inferior frontal gyrus, premotor cortex, anterior insula, adjacent frontal operculum, snterior cingulate gyrus, supermarginal gyrus, intraparrietal sulcus, anterior thalamus and the midbrain. Ok theres more but i cba
- Discussion
    * "discrimination of words from very word-like nonwords requires access to word-specific knowledge"
    * Previous studies found imageable concrete words more rapidly recognised than abstract words suggesting:
        · Concrete words arouse qualitatively distinct semantic codes not accessed by abstract words (Pavio 1971)
        · Or activate the same kind of semantic codes more efficiently (Schwanenflugel 1991)
    * "The facilatory effect of semantic access on performance can be unmderstood by viewing lexical decision as a signel detection task in whih participants combine orthographic, phonological and semantic information in whatever way most efficiently and accurately classifies stimuli as words or nonwords"
    * (Signal) Detection theory: is a means to measure the ability to differentiate between information-bearing patterns and random patterns that disctract from information. [wikipedia]

    * They claim that because concrete and abstract word sets used there are matched on orthographics and phonological characteristics, then better performance on concrete words compared with abstract words confirms that access to word meaning occured during the physiological recordings.
    * There are lots of studies contracting semantic and nonsemantic tasks
    * Language baed semantic system in the left hemisphere and a nonverbal one in the right
    * Acivations in both brain regions implies comon mechanism such as contextual access
    * There is a best candidate for a semantic brai region
    * To accept an abstract item as a word require holding its phonological form in working memory while retrieving associated words to the item
    *

- (2005) Abstract and Concrete concepts have structurally different representation frameworks
  - Abstract
    * Demonstrate that semantically associate abstract words reliably interfere with one another significantly more than semantically synonymous abstract words
    * Abstract concepts are represneted in associative neural network
    * Their patient had signifcantly greater difficulty identifying high frequency than low frequency abstract words. They say this is first evidence for inverse word frequency effect
  - Intro
    * So this paaper seems to contrast DCT and (i think this one is also cited) context availability theories. i.e. They consider a quantative distinction between abstract and concrete concepts. But these guys argue that the fundimental distinction is rooted in qualitatively different principles of organisation
    * **They compare semantic similarity and semantic association as competing principles of organisation for abstract and concrete word semantics**
  - Experiment 1
    * abstract vs concrete may be on a continuum
    * They are testing the influence of abstractness and word frequency on the patients word identification skills
    * Increased impairment at word identification with greater abstractness of target word, and aninverse frequency effect observedwhich is striking apparently
  - Experiment 2:
    * Temporal factor affects on abstract word comprehension
    * Speed made it worse
  - Experiment 3:
    * Semantic similarity on abstract adjective and verb comprehension
    * Academics with their ridiculously obtuse way of speaking

    * Unlike concrete, rafractoriness does not build up more quickly among semantically related abstract words than semantically unrelated abstract words.
  - Experiment 4:
    * Semantic similarity's influence on abstract and concrete word comphrension
    * So basically, semantic refractory access dysphasia has this thing where semantic distance effects are attributed to the abnormal deleterious effects which activating a concept has on other concepts that share neural space. The fact that experiment 3 shows no difference implies that abstract words with similar meaning dont share neural space.
    * Very few adjectives are highly concrete
    * Yeah they find that refractoriness doesnt build up more quickly in abstract words which have similar meaning vs those that dont. But concrete revealed significant effects
  - Experiment 5:
    * Influence of semantic association on abstract and concrete word comprehension
    * Association = meaning is not synonymous, but are bound together in real world or 'sentential' contexts, salute, army, respect...
    * Refractoriness builds up much more easily among associated than synonymous words.
    * This is not noted in concrete words.
  - Discussion:
    * Abstract and concrete each exhibit double dissociation between similar and not similar semantic-contextual-association and semantic-similarity respectively
    * Experiencing through five senses play a key role in aquiring concrete concepts
    * Abstract may be required from language no sensory input needed
    * Make sure to re-read discussion its great
    * Since they think that abstract vs concrete is on a spectrum, they suggest that(Anderson and Nagy 1991 from 2005 AnC have struct..., including 2005 paper also) associative/categorical dichotomy is relative rather than absolute. i.e. middling items have both in more equal proportions than concepts at either end.
    * Great quote: "Essentially, our findings suggest that attempting to model conceptual knowledge within a unitary system based on a single set of network principles is over simplistic"
    * Some people argue that all words are polysemous because of context
    *

- (2007) Spatio-temporal cortical dynamics underlying abstract and concrete word reading
  - Abstract

    * Findings suggest words are initially understood using a left-lateralised verbal-lingsuitic system and that for concrete words are supplemented after a short delay by right parietal and medial occiptal imagistic network.
  - Intro:
    * DCT vs context-availibility, context availibility explains reaction time advantage of concrete words by stronger links to contextual information in semantic memory.
    * The negativity was extensively studied as the **n400** component. They take quite a decisive stance on N400, saying it is evoked by pronounceable non-words occuring in isolation or sentences. Inversely proportional to the ease with which the stimulus may be integrated into the current cortex.
    * N700 modulated for concrete word processing only
    * Brain responses to abstract and concrete word presentations demostrated similar posterior-to-anterior sequences of cortical recruitment.
  - Discussion:
    * Lots in here, see highlights
    * Left frontotemporal areas associated with N400 responded more to concrete than abstract
    * Rigth anterior temporal areas associated with N400m demonstrace increased response to abstract work, peaking at slightly longer latencies.
    * Confirms previous works suggesting differences due to strnger imagery vs nonimagistic representations
    * n400 triggered by 'potentially meaningful' stimuli, thought to embody processing within an associative semantic network encompassing the integration of a current event with an ongoing context
    * suggest decreased left N400m to avstract words represents a more efficient or extensive representation for these words within frtontotemporal networks
    * May be inferred that relatively decreased right n400m to concrete represents more efficient or extensive representation within right hemisphere networks.
    * Right occipitoparietal imagery-related processing may contribute informatio to the frontotemporal N400m, leading to faster termination of the n400m to concrete words (and ultimately faster reaction times)
    * Their results overall imply a joint processing and then separtion of sorts (rephrase this)

- (2014) Concreteness effects in semantic processing in ERP for spanish words
- (NEED PDF PAYGATING SCUM) Concrete spatial language, see what i mean?

13 Experimental Scope

- Dual coding Theory vs Context Availability theory. Using video BERT, can we gauge how contextual associated knowledge works with abstract and concrete concepts.
- Can indentify present concrete or abstract words using visual concepts or labels from object detectors

- P300 and P400 components could be used to control or isolate incongruencies
- Concret words have categories, abstract words have synonyms, inspired by 2005 abstract and concrete concepts have structurally different representation frameworks
- Concrete vs abstract word categorisation and prediction
- Semantic distance-aware datasets (genome?) and models
- There is no generalised set of abstract vs concrete. Il should extract this myself across all datasets

14 Datasets to Consider

We would like to explore a more diverse range of multimodal datasets. Potentially drawing from VQA, video-Qa, visual dialog and elsewhere.

14.1 AVSD

This is the AVSD challenge citation [3], this is the AVSD paper [2].

14.2 VQA-CP

The VQA-CP v1/2 (changing priors) datasets [1] are adapted from the VQA v1/2 datasets respectively. Link the original datasets here.

14.3 Flikr30k Entities

Original Flikr30k dataset [63]. The Flikr30k entities dataset [51]. **https://github.com/BryanPlum**

15 Criticisms and Discussion

Why would a model via attention not automatically learn to prioritise abstract vs concrete things?:
Good point, well motivation from this comes from the idea that separate cortical systems are used for processing abstract and non-abstract concepts. **This is an ambitious leap from neurological inspiration to the realities of deep learning:**
Yes, yes it is. I'll expand on this later.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[2] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. Audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[3] Huda Alamri, Chiori Hori, Tim K Marks, Dhruv Batra, and Devi Parikh. Audio visual scene-aware dialog (AVSD) track for natural language generation in DSTC7. In *AAAI workshop on the 7th edition of Dialog System Technology Challenge (DSTC7)*, January 2019.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998, 2017. URL http://arxiv.org/abs/1707.07998.

[5] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2016.

[6] Tadas Baltruaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:423–443, 2019.

[7] Ian Maynard Begg. Recall of meaningful phrases. 1972.

[8] Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. Multimodal grounding for language processing. *ArXiv*, abs/1806.06371, 2018.

[9] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8102–8109, 2019.

[10] Hanna Benoni and Yehoshua Tsal. Conceptual and methodological concerns in the theory of perceptual load. *Frontiers in Psychology*, 4, 2013.

[11] Jeff Bezemer and Gunther R. Kress. Writing in multimodal texts a social semiotic account of designs for learning. 2008.

[12] Jeffrey R. Binder, Chris F. Westbury, Kristen A. McKiernan, Edward T. Possing, and David A. Medler. Distinct brain systems for processing concrete and abstract concepts. *Journal of Cognitive Neuroscience*, 17:905–917, 2005.

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

[14] Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. Fatality killed the cat or: Babelpic, a multimodal dataset for non-concrete concepts. In *ACL*, 2020.

[15] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit.*, 77:329–353, 2018.

[16] John B. Carroll. Human cognitive abilities: Survey and analysis of correlational and factor-analytic research on cognitive abilities: Overview of outcomes. 1993.

[17] Jean Charbonnier and Christian Wartena. Predicting word concreteness and imagery. In *IWCS*, 2019.

[18] Jean Charbonnier and Christian Wartena. Predicting the concreteness of german words. In *SwissText/KONVENS*, 2020.

[19] Maurizio Corbetta and Gordon L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3:201–215, 2002.

[20] Sebastian J. Crutch and Elizabeth K. Warrington. Abstract and concrete concepts have structurally different representational frameworks. *Brain : a journal of neurology*, 128 Pt 3:615–27, 2005.

[21] Rupali P. Dhond, Thomas Witzel, Anders M. Dale, and Eric Halgren. Spatiotemporal cortical dynamics underlying abstract and concrete word reading. *Human brain mapping*, 28 4:355–62, 2007.

[22] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1473–1482, 2015.

[23] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *CoRR*, abs/1606.01847, 2016. URL `http://arxiv.org/abs/1606.01847`.

[24] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. *CoRR*, abs/1511.06062, 2015. URL `http://arxiv.org/abs/1511.06062`.

[25] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[26] W. Guo, J. Wang, and S. Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

[27] John T. Guthrie. Comprehension and teaching : research reviews. 1981.

[28] Jack Hessel, David Mimno, and Lillian Lee. Quantifying the visual concreteness of words and topics in multimodal datasets. *ArXiv*, abs/1804.06786, 2018.

[29] Felix Hill and Anna Korhonen. Learning abstract concept embeddings from multi-modal data: Since you probably can't see what i mean. In *EMNLP*, 2014.

[30] Felix Hill, Roi Reichart, and Anna Korhonen. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296, 2014.

[31] Phillip J. Holcomb, John Kounios, John Arnold Edward Anderson, and William Carey West. Dual-coding, context-availability, and concreteness effects in sentence comprehension: an electrophysiological investigation. *Journal of experimental psychology. Learning, memory, and cognition*, 25 3:721–42, 1999.

[32] Beibei Jin, Yu Hu, Qiankun Tang, Jingyu Niu, Zhiping Shi, Yinhe Han, and Xiaowei Li. Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4554–4563, 2020.

[33] Kenneth O Johnson. Neural coding. *Neuron*, 26(3):563–566, 2000.

[34] Andrej Karpathy, Armand Joulin, and Fei Fei Li. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, 2014.

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

[36] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, JungWoo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *CoRR*, abs/1610.04325, 2016. URL http://arxiv.org/abs/1610.04325.

[37] Mikhail Kiselev. Rate coding vs. temporal coding – is optimum between? 07 2016. doi: 10.1109/IJCNN.2016.7727355.

[38] Shu Kong and Charless C. Fowlkes. Low-rank bilinear pooling for fine-grained classification. *CoRR*, abs/1611.05109, 2016. URL http://arxiv.org/abs/1611.05109.

[39] Maximilian Köper, Sabine Schulte im Walde, and Diego Frassinelli. Exploring multi-modal text+image models to distinguish between abstract and concrete nouns. 2017.

[40] Nilli Lavie and Yehoshua Tsal. Perceptual load as a major determinant of the locus of selection in visual attention. *Perception Psychophysics*, 56:183–197, 1994.

[41] Nilli Lavie, Aleksandra Hirst, Jan W. de Fockert, and Essi Viding. Load theory of selective attention and cognitive control. *Journal of experimental psychology. General*, 133 3:339–54, 2004.

[42] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnns for fine-grained visual recognition. *arXiv: Computer Vision and Pattern Recognition*, 2015.

[43] Anthony David Milner. How do the two visual streams interact with each other? *Experimental Brain Research*, 235:1297 – 1308, 2017.

[44] J. Morton. Facilitation in word recognition: Experiments causing change in the logogen model. 1979.

[45] Hesham Mostafa. Supervised learning based on temporal coding in spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 29:3227–3235, 2018.

[46] James H. O'keefe and J Dostrovsky. The hippocampus as a spatial map. preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 34 1:171–5, 1971.

[47] Allan Paivio. *Imagery and verbal processes*. Psychology Press, 2013.

[48] Allan Paivio. Intelligence, dual coding theory, and the brain. 2014.

[49] Allan Paivio and Wallace E. Lambert. Dual coding and bilingual memory. 1981.

[50] Jaak Panksepp. Affective neuroscience: The foundations of human and animal emotions. 1998.

[51] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017.

[52] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

[53] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *NeurIPS*, 2017.

[54] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. *ArXiv*, abs/1704.04368, 2017.

[55] Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. Visually grounded neural syntax acquisition. *ArXiv*, abs/1906.02890, 2019.

[56] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019.

[57] Hao Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP*, 2019.

[58] Joshua B. Tenenbaum and William T. Freeman. Separating style and content with bilinear models. *Neural Comput.*, 12(6):1247–1283, June 2000. ISSN 0899-7667. doi: 10.1162/089976600300015349. URL http://dx.doi.org/10.1162/089976600300015349.

[59] Simon Thorpe. Spike arrival times: A highly efficient coding scheme for neural networks. *Parallel Processing in Neural Systems and Computers*, 01 1990.

[60] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019:6558–6569, 2019.

[61] 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. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611, 2019.

[62] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *ArXiv*, abs/1810.02338, 2018.

[63] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

[64] Jun Yu, Jing Li, Zhou Yu, and Qingming Huang. Multimodal transformer with multi-view visual representation for image captioning. *ArXiv*, abs/1905.07841, 2019.

[65] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *CoRR*, abs/1708.01471, 2017. URL http://arxiv.org/abs/1708.01471.

[66] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29:5947–5959, 2018.

[67] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications, 2019.