ESPGame		CSLB	
Image 1	Image 2	Crocodile	Screwdriver
red	wreck	has 4 legs (7)	has handle (28)
chihuaua	cyan	has tail (18)	has head (5)
eyes	man	has jaw (7)	is long (9)
little	crash	has scales (8)	is plastic (18)
ear	accident	has teeth (20)	is metal (28)
nose	street	is green (10)	
small		is large (10)	

Table 1: Concepts identified in images in the ESP Game (left) and features produced for concepts by human annotators in the CSLB dataset (with feature strength, max=30).

Concept 1	Concept 2	Assoc.
abdomen (6.83)	stomach (6.04)	0.566
throw (4.05)	ball (6.08)	0.234
hope (1.18)	glory (3.53)	0.192
egg (5.79)	milk (6.66)	0.012

Table 2: Example concept pairs (with mean concreteness rating) and free-association scores from the USF dataset.

Wikipedia text, split into sentences and with punctuation removed.

2.2 Evaluation

We evaluate the quality of representations by how well they reflect *free association* scores, an empirical measure of cognitive conceptual proximity. The University of South Florida Norms (USF) (Nelson et al., 2004) contain free association scores for over 40,000 concept pairs, and have been widely used in NLP to evaluate semantic representations (Andrews et al., 2009; Feng and Lapata, 2010; Silberer and Lapata, 2012; Roller and Schulte im Walde, 2013). Each concept that we extract from the USF database has also been rated for conceptual concreteness on a Likert scale of 1-7 by at least 10 human annotators. Following previous studies (Huang et al., 2012; Silberer and Lapata, 2012), we measure the (Spearman ρ) correlation between association scores and the cosine similarity of vector representations.

We create separate abstract and concrete concept lists by ranking the USF concepts according to concreteness and sampling at random from the first and fourth quartiles respectively. We also introduce a complementary noun/verb dichotomy, on the intuition that information propagation may occur differently from noun to noun or from noun to verb (because of their distinct structural relationships in sentences). POS-tags are not assigned as part of the USF data, so we draw the noun/verb distinction based on the majority POS-tag of USF concepts in the lemmatized British National Corpus (Leech et al., 1994). The abstract/concrete and noun/verb dichotomies yield four distinct concept lists. For consistency, the concrete noun list is filtered so that each concrete noun concept w has a perceptual representation $\mathbf{b}$(w) in both $\mathbf{P_{ESP}}$ and $\mathbf{P_{CSLB}}$. For the four resulting concept lists C (concrete/abstract, noun/verb), a corresponding set of evaluation pairs $\{(w_1, w_2) \in USF : w_1, w_2 \in C\}$ is extracted (see Table 3 for details).

Concept Type	List	Pairs	Examples
concrete nouns	541	1418	*yacht, cup*
abstract nouns	100	295	*fear, respect*
all nouns	666	1815	*fear, cup*
concrete verbs	50	66	*kiss, launch*
abstract verbs	50	127	*differ, obey*
all verbs	100	221	*kiss, obey*

Table 3: Details the subsets of USF data used in our evaluations, downloadable from our website.

3 Results and Discussion

Our experiments were designed to answer four questions, outlined in the following subsections: (1) Which model architectures perform best at *combining* information pertinent to multiple modalities when such information exists explicitly (as common for concrete concepts)? (2) Which model architectures best propagate perceptual information to concepts for which it does not exist explicitly (as is common for abstract concepts)? (3) Is it preferable to include all of the perceptual input that can be obtained from a given source, or to filter this input stream in some way? (4) How much perceptual vs. linguistic input is optimal for learning various concept types?

3.1 Combining information sources

To evaluate our approach as a method of information combination we compared its performance on the concrete noun evaluation set against three alternative methods. The first alternative is simple concatenation of these perceptual vectors with linguistic vectors embeddings learned