A Glove and Picturebook Gating

In Table 11 we show the top 10 most frequent words sorted by mean gate activation value across 4 datasets. On all datasets, the strongest scoring words for Glove are abstract (and don't provide meaningful image search results) while the strongest scoring words for Picturebook are almost always visual. Occasionally we also find that Picturebook prefers misspelled words, perhaps this is due to auto-correcting the spelling when performing image search. We note these are a representative sample of typical results across models but different runs can produce different high-scoring words.

A.1 Nearest Neighbours

We also compute nearest neighbours across languages in Table 12. We take an English word and find the closest French and German words. Some of the words capture visual similarity but not necessarily semantic similarity, for example, 'zebra' in English is an animal with black and white stripes, while the nearest neighbours in German are 'zebra', 'tiger' and 'gähnender' (note, our image search results for 'tiger' did not include white tigers). Our image search results for 'gähnender' resulted in various animals yawning. The embedding space also relates 'tourist' in English to 'dorothy' both in French and German. Presumably this is due to our CNN capturing semantic relatedness of Dorothy as a traveller in the Wizard of Oz. Interestingly, the English word 'not' has nearest neighbours of 'interdit', '10' and '1' in French. Our image search returns '1' and '10' with several red digits and circular backgrounds, which is somewhat visually similar to stop signs. The closest German words 'stop', 'kein', 'not' mostly return red signs.

B Hyperparameters

B.1 Natural Language Inference

For BoW models, we represent sentences as the sum of their word embeddings and use a 1024-dimensional 'translation' layer with ReLU activation [5] prior to summing the word embeddings. For BiLSTM-Max, the encoder size is treated as a hyperparameter and searched over, using the translated layer as inputs. As in previous work, the joint embedding layer combines the premise and hypothesis embeddings along with their componentwise product and absolute difference. A feedforward neural net with 2 hidden layers with ReLU activations are added prior to the softmax layer in all models. We search over a) the hidden state size for BiLSTM-Max models, b) the size of the hidden layers c) the dropout rates d) learning rate decay schedule and e) the dimensionality of the fusion layer for Glove+Picturebook models. All models were trained with a minibatch size of 32 and Adam with an initial learning rate of 0.0003. BiLSTM-Max models uses gradient clipping with a norm of 10 and recurrent dropout without memory loss (Semeniuta et al., 2016). On SNLI, the learning rate was decreased exponentially by a factor of 2 every 40K iterations. On MultiNLI, the learning rate was decreased exponentially by a factor of 2 every 30K iterations. Early stopping is employed with the best development set model used to report test set results on all tasks.

B.2 Semantic Relatedness

For all models and datasets, we represent sentences as the sum of their word embeddings. A 'translation' layer with ReLU activation is applied to all models prior to summing the word embeddings. The joint embedding layer combines the two sentence embeddings along with their componentwise product and absolute difference. A feedforward neural net with 2 hidden layers with ReLU activations are added prior to the softmax layer. For each dataset and model, we perform a search over a) the encoder size b) the size of the hidden layers c) learning rate decay schedule and d) the dropout rates. All models use a minibatch size of 16. Models were trained with Adam for 2500 iterations with an initial learning rate of 0.001. The learning rate was decreased exponentially by a factor of 2 every 250 iterations. The best settings found on the development set were then trained 5 times, which we report the average test set result.

B.3 Sentiment and Topic Classification

For all models and datasets, we represent sentences as the sum of their word embeddings. A 1024-dimensional 'translation' layer with ReLU activation is applied to all models prior to summing the word embeddings. A feedforward neural net with 2 hidden layers with ReLU activations are added prior to the softmax layer. For each dataset and model, we perform a search over a) the size of the hidden layers b) learning rate decay schedule

[5] https://github.com/Smerity/keras_snli

SNLI		MultiNLI		AG-News		Yelp	
glove	picturebook	glove	picturebook	glove	picturebook	glove	picturebook
comes	landscaping	that	boots	that	Palestinians	doing	port
instead	excavation	as	drawer	Since	painkiller	when	Diet
even	party	an	demons	other	Mets	forget	et
often	parka	made	salamander	same	stronghold	and	Gross
again	Rodgers	a	planners	now	Motorsports	personally	pinball
you	Ballerina	the	steak	others	Saint	want	receipt
when	Motocross	put	wines	Also	van	did	WTF
turns	sandcastles	be	Vatican	the	guys	I	cowboy
where	Bank	of	Mykonos	out	Heisman	almost	Jersey
there	recipe	where	cops	then	Hospital	be	Street

Table 11: Top 10 words most associated with each embedding by mean gate activation value within the top 10K most frequent words for each dataset.

English	French	German
1	5, moins, 1	o, 12, 1
airplane	airline, avion, plane	british-airways-flugzeug, flieger, plane
graffiti	graffiti, murales, graffitis	graffiti, graffitti, grafittis
motorcycle	kawasaki, yamaha, harley	motorräder, motorrad, kawasaki
not	interdit, 10, 1	stop, kein, not
tourist	touriste, dorothy, déguise	tourist, dorothy, mexikaner
zebra	zébrés, zébré, tigre	zebra, tiger, gähnender

Table 12: The nearest French or German neighbours for an English word in the Picturebook space.

and c) the dropout rates. All models were trained with a minibatch size of 64 and Adam with an initial learning rate of 0.0003. The learning rate was exponential decayed by a factor of 2 over roughly 5 epochs on each dataset. Early stopping is employed with the best development set model used to report test set results on all tasks.

B.4 Image-Sentence Ranking

All of our VSE++ models are trained in two stages, as in Faghri et al. (2017). In the first stage, the Inception-V3 parameters are frozen and the BiLSTM-Max and embedding matrices are learned. In the second stage, the whole model is fine-tuned end to end. For the first training stage, our models are ran for 250K iterations using a minibatch size of 64. We use Adam with an initial learning rate of 0.0003 and exponentially decay it by a factor of 2 over 50K iterations. The second stage training is done for an additional 250K iterations using SGD with a fixed learning rate of 0.0005. Following Faghri et al. (2017), we use random image crops at training time and evaluate on the center crop at test time. All models use a 1024-dimensional embedding space as in Faghri et al. (2017). Gradient clipping is used with a norm of 5. We also experiment with adding recurrent dropout without memory loss (Semeniuta et al., 2016) to the sentence encoder but found it did not help.

B.5 Machine Translation

For our Multi30k MT experiments, we used the standard seq2seq model with content-based attention attention and with 2 Layer Normalized (Ba et al., 2016) LSTM (Hochreiter and Schmidhuber, 1997) layers in the encoder and decoder, dropout rate of 50% and label smoothing $\epsilon_{ls} = 0.1$ (Pereyra et al., 2017). We train with a minibatch size of 8 and Adam (Kingma and Ba, 2015) optimizer with $\epsilon = 1\text{e}{-4}$ and learning rate of $1\text{e}{-3}$ annealed to $1\text{e}{-4}$. We use the same hyperparameters for our IWSLT experiments except we use a dropout rate of 20% and a minibatch size of 32.