CMPUT 600 Project Report: Automatically Selecting Highly Pertinent Images For BabelNet Concepts With BabelCLIP

Michael Ogezi
Department of Computing Science,
University of Alberta
ogezi@ualberta.ca

Abstract

This work tackles the problem of selecting a basket of k highly pertinent images for a BabelNet concept. To this end, we propose BabelCLIP, which encodes concept-extracted text and images and then ranks the images with an image-concept similarity measure. Our results show that this approach outperforms prior state-of-the-art (SOTA) methods.

1 Introduction

BabelNet (Navigli and Ponzetto, 2010) is an expansive multilingual semantic network. However, it was generated automatically with resources from many sources, including WordNet (Miller, 1995), Wikipedia, OmegaWiki (Meijssen, 2009), GeoNames, ImageNet (Deng et al., 2009), Open Multilingual WordNet and VerbAtlas (Di Fabio et al., 2019). And therein lies the problem, since BabelNet is automatically generated by crawling the aforementioned resources, it is prone to errors such as missing concepts, mistranslated words and unsuitable image attachments. In this work, we focus on the last problem.

For our task, the input is a BabelNet concept c and the output is a basket of images pertinent to that concept b. We employ **C**ontrastive **L**anguage-**I**mage **P**re-training (CLIP) (Radford et al., 2021), a language-image model trained with a contrastive setup on 400 million image-text pairs, in identifying the images that best represent BabelNet concepts. CLIP's pre-training approach allows it to deeply associate language semantics with visual semantics.

The contributions this work makes are two-fold:

1. To provide BabelCLIP, a framework for automatically aligning BabelNet concepts to a relevant and semi-distinct basket of images that effectively convey semantic meaning.
2. To show why this approach can help machines understand the world better.

Here, semi-distinct means that the images are related by a shared concept but are not byte-for-byte the same or minor transformations of each other such as crops and greyscales.

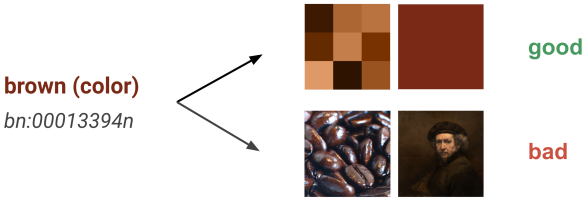

Figure 1: *Good* vs *bad* top-2 concept-image attachment for the BabelNet concept of BROWN (*bn:00013394n*)

As Figure 1 succinctly captures, BabelNet's automatic concept-image attachments are imperfect. For BROWN, the lower images of coffee beans and a man are much less pertinent than the upper images depicting shades of brown, however, they would all be automatically attached within BabelNet.

2 Related Work

With increased computational capabilities in recent years, work related to multimodal models has received increased attention. On the other hand, researchers have been working with semantic networks such as BabelNet and WordNet since these resources were initially released decades ago.

2.1 CLIP and ALIGN

Although their architectures vary, CLIP (Radford et al., 2021) and **A** **L**arge-scale **I**ma**G**e and **N**oisy-Text Embedding (**ALIGN**) (Jia et al., 2021) attempt to align features computed from language data and image data in a common embedding space, leading the model to output similar features for related text-image pairs and different ones for unrelated pairs. This leads to the emergence of capabilities such as zero-shot image classification and text-to-image generation.

2.2 BabelNet

BabelNet (Navigli and Ponzetto, 2010) is an automatically generated multilingual semantic network. However, in the context of our work, we focus on the concept-image attachments specifically.

2.3 BabelPic

Calabrese et al. (2020) released BabelPic,"a hand-labelled dataset built by cleaning the image-synset association found within the BabelNet Lexical Knowledge Base." They produced an expert-curated gold dataset alongside a silver dataset generated with a vision-language model trained to predict whether images were pertinent to concepts. We speak more extensively about this in Section 3.2.

3 Datasets

The datasets that we use come from BabelPic and BabelNet. From BabelPic, we retrieve the gold and silver datasets. These are structured as a mapping from a concept to a bucket of one or more selected images. Similarly, with BabelNet, we assemble a set of all the images attached to a concept. This is a superset of the BabelPic-selected images. Table 1 presents detailed statistics of our BabelPic-BabelNet combination dataset.

3.1 Preprocessing

This section mainly concerns our image data. As is usually the case with automatically-aggregated or web-sourced data, noise abounds. A minority of the files (under 1%) were not even in an image format. We decided to handle only JPGs, PNGs, SVGs, and GIFs because these formats, when combined, offered a coverage rate of **98.77%**.

3.1.1 Filtering

Additionally, we attempted to filter out exact duplicates and near duplicates. In doing this, we employed three progressive steps. First, with URL-based deduplication, we removed images with the same URL. Second, with cryptographic hash-based deduplication, we removed SHA512-hash matches. Finally, with perceptual hash-based deduplication (Zauner, 2010), we removed matches. We note that inasmuch as these steps greatly aided deduplication, the process was imperfect.

3.1.2 Format conversion

In the case of SVGs, we converted them to the PNG format. For GIFs, we took the middle frame of the GIF and converted that to a PNG. Curiously, the other formats included video, audio and obscure image file-formats including OGGs, OGVs, TIFFs, PDFs, MIDIs, MP3s, FLACs etc. These were dropped.

	Gold	Silver
Instances		
Synsets (S)	2,756	9,819
Selected Images (I_S)	15,140	65,497
BN Images (I_{BN})	99,463	496,486
Nouns (N)	2,616	9,819
Verbs (V)	140	-
Adjectives	-	-
Adverbs	-	-
I_S per S	5.49	6.67
I_{BN} per S	36.09	50.56
I_S per N	5.68	6.67
I_S per V	2.02	-
I_{BN} per N	37.80	50.56
I_{BN} per V	4.04	-
Formats		
JPGs	67,904 (68.27%)	313,145 (63.07%)
PNGs	12,306 (12.37%)	71,117 (14.32%)
SVGs	16,095 (16.18%)	98,077 (19.75%)
GIFs	1,871 (1.88%)	8,076 (1.63%)
Others	1,287 (1.29%)	6,071 (1.22%)
Size	6.42 GB	82.28 GB

Table 1: These are detailed statistics of the gold and silver datasets from BabelPic along with our extra data sourced from BabelNet. I_S refers to the set of selected images for a concept while I_{BN} refers to the set of all the images aggregated in BabelNet for a concept. $I_S \subset I_{BN}$.

3.2 BabelPic's Gold and Silver Datasets

The gold dataset was produced by human experts in the work presented in BabelPic. Here, BabelPic gold selected a mean of **15.22%** of the images as highly pertinent.

In the case of the silver dataset, this was generated by a doubly fine-tuned **V**ision-**L**anguage **P**re-training (VLP) (Luowei Zhou, 2019) model. It had earlier been fine-tuned on Visual Question Answering after being trained on the Conceptual Captions dataset (Sharma et al., 2018). The BabelPic authors provided an image-synset ($I - S$) pair and asked the VLP model, "Does the image depict S?". On average, BabelPic silver selected

13.19% of the images as highly pertinent.

4 BabelCLIP

Our framework, aptly named BabelCLIP, employs the CLIP model in selecting a basket of images that is highly pertinent to a concept. It consists of both a text encoder transformer and an image encoder transformer. Our framework consists of the following steps:

1. Select a BabelNet concept eg. *bn:00015267n*.
2. Retrieve the set of all t images that have been automatically attached to the concept within BabelNet.
3. Select a basket of the n most pertinent images out of t. Strictly, $n \leq t$.

The third step warrants further explanation. We use the following mathematical notations:

- d refers to the size of the embedding vectors.
- t refers to the total number of images attached to a BabelNet concept. This is variable across concepts.
- $E_T \in R^d$ refers to the embeddings generated for a piece of text.
- $\mathbf{E_I} \in R^{d \times t}$ refers to the embeddings generated for a group of images. We generate embeddings $E_I \in R^d$ for each image and stack them t times to form $\mathbf{E_I}$.

Selecting the most pertinent images consists of two parallel substeps. On one hand, we use a text encoder to produce E_T for some text associated with a concept. Section 5.1 explores the methodology employed in selecting which text to use more extensively. On the other hand, we use an image encoder to produce $\mathbf{E_I}$ for all the images provided by BabelNet. Finally, we compute the softmax of $\mathbf{E_I}^\top E_T \in R^t$ and pick the n results (out of t images) with the highest probabilities. We may then retrieve the images that correspond to these n probabilities in order.

5 Experiments

We ran our experiments on twin NVIDIA GeForce RTX 3090 GPUs with 10,496 1.4GHz CUDA cores and 24GBs RAM.

5.1 Text Embeddings

To compute how pertinent a concept is to an image, we extract some text data from the concept. Some examples of text data available from BabelNet concepts include lemmas, glosses and examples.

We also note that our text encoder had a maximum token sequence length of 77. This means that after a concept's extracted text data was tokenized, we truncated the token list if it was longer than 77. We speak more about the negative effects of this in Section 4. We further outline strategies used to generate text embeddings for concepts.

1. **Lemmas-based**: Here, we produce our embedding from a text string built by concatenating all the lemmas associated with a concept. Lemmas are words or short phrases that can help explain a concept.
2. **Gloss-based**: A gloss is a definition for a concept. We use this, as a string, without any extra transformation.
3. **Example-based**: Some concepts have examples attached to them. These entail a sentence that portrays an example of how a concept is traditionally used.
4. **Example-based with Lemmas fallback (Ex+Lemmas)**: When a concept has no example, we fall back to the lemmas-based strategy.
5. **Example-based with Gloss fallback (Ex+Gloss)**: When a concept has no example, we fall back to the gloss-based strategy.

From our manual evaluation, the fourth strategy performed the best as it was in the preferred group of baskets **84.62%** of the time. We theorize that this is because examples offer a stronger indicator of how a concept is represented in the real world and is thus more similar to the text data CLIP was pre-trained on. Additionally, since examples are sparse, lemmas provide a decent fallback.

5.2 Image Encoder Configurations

Our image encoder is based on a **Vi**sion **T**ransformer (ViT) (Dosovitskiy et al., 2020). The CLIP image encoder was trained on 224×224 images. The considered configurations are the base models with 32×32 patches (B/32), and 16×16 patches (B/16) as well as the large model with 14×14 patches (L/14). Note that the patch dimensions are all factors of 224. These patches are passed to the encoder as visual tokens.

When selecting images, since the number of images selected by BabelPic datasets is almost always smaller than the total number of images attached to the concept, we show the top-k images where

	CLIP		
Hyperparams	B/32	B/16	L/14
Parameters	86 M	86 M	307 M
Datasets			
ImageNet	76.10%	80.20%	**83.90%**
CIFAR-10	95.10%	96.20%	**98.00%**
CIFAR-100	80.50%	83.10%	**87.50%**

Table 2: Comparison of the three image encoder configurations across hyperparameters and linear probing-based evaluation accuracy. Aggregated from Radford et al. (2021) and Dosovitskiy et al. (2020).

$k = min(n, s)$ and s is the number of BabelPic-selected images while n retains its earlier definition.

Table 2 shows us that the best performing image encoder is L/14. Thus, that is what we employ going forward.

5.3 Results

We randomly sampled the concept-image pairs to construct a **blind preference test** wherein we provided a human evaluator with two unattributed image baskets, one selected by BabelPic and the other by BabelCLIP. The evaluator then selected one choice from: prefer *first*, prefer *second* or prefer *both/neither*. Our scoring metric, the blind preference rate ρ is expressed by this formula:

$$\rho(a) = \frac{n(prefer\ a) + n(prefer\ both/neither)}{sample\ size} \quad (1)$$

We observe that we outperform BabelPic silver but underperform BabelPic gold. However, we impressively remain competitive with it.

	on Gold	on Silver
BabelPic	**83.33%**	75.00%
BabelCLIP	80.56%	**80.00%**
Both/Neither	63.89%	55.00%

Table 3: Comparison of blind preference rates over a sample of BabelPic- and BabelCLIP-selected images.

The fact that the preference of both/neither on silver is better than chance shows that automated methods, to some extent, suffer from similar pitfalls.

5.4 Error Analysis

1. **Image Format Conversion**: SVGs tend to be less natural than images of other formats. BabelCLIP sometimes got confused by images that had been converted from SVGs. These images had larger than average proportions of white or black coming from their unnatural backgrounds. PNGs and GIF-converted images suffered from this, albeit to a lesser degree. We, however, decided against filtering out these images because they sometimes provided very good matches to concepts.
2. **Image Deduplication**: The active steps we took to counteract duplicates (see Section 3.1.1), had commendable but limited success. This was particularly problematic since duplicates provide no marginal utility to the basket.
3. **Concept-Text Selection**: Although we settled on the ex+lemmas strategy, there were cases where other strategies outperformed it. In the future, we may look towards ensemble methods.
4. **Text Token Truncation**: The token limit of 77 proved problematic for concepts with longer glosses or many lemmas. In further work, we look to increase the capacity of the text encoder to countervail this.
5. **Concept Concreteness**: The BabelPic dataset focuses on non-concrete (NC) concepts as opposed to concrete (C) ones. Love and religion versus hug and temple are respective examples of NC vs C. We note that our performance here may not perfectly extrapolate to concrete concepts, although we do not expect particularly poor performance since NC concepts are generally the harder group to decipher.
6. **Concept Esotericity**: We had issues evaluating jargony concepts such as esoteric animal and plant species. Further work may benefit from working with people more well versed in Botany and Biology.

6 Conclusion

In BabelCLIP, we propose a framework for automatically generating a pertinent basket of images for a concept. Our results are positive, and significantly improve the SOTA, indicating that we can semi-reliably use this framework to automatically attach images to concepts. In future work, we look to counteract some of our observed sources of error and improve overall system integrity. Our code is available at https://github.com/okibeogezi/BabelCLIP

Acknowledgements

This final report was put together with guidance from Professor Greg Kondrak, who proposed the idea of using BabelNet images to better understand concepts. Additionally, he suggested the idea of blind preference tests and performed some tests.

This work also owes a debt of gratitude to Greg Kondrak, Bradley Hauer and Karim Ali for providing and administering the compute resources that made a course project of this level of quality possible.

References

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2020. Fatality killed the cat or: BabelPic, a multimodal dataset for non-concrete concepts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4680–4686, Online. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Andrea Di Fabio, Simone Conia, and Roberto Navigli. 2019. VerbAtlas: a novel large-scale verbal semantic resource and its application to semantic role labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 627–637, Hong Kong, China. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Lei Zhang Houdong Hu Jason J. Corso Jianfeng Gao Luowei Zhou, Hamid Palangi. 2019. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*.

Gerard Meijssen. 2009. The philosophy behind omegawiki. *Lexicography at a crossroads: Dictionaries and encyclopedias today, lexicographical tools tomorrow*, 90:91.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, page 216–225, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

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

Christoph Zauner. 2010. Implementation and benchmarking of perceptual image hash functions.