C-CLIP: Contrastive Image-Text Encoders to Close the Descriptive-Commentative Gap

William Theisen
University of Notre Dame
wtheisen@nd.edu

Walter Scheirer
University of Notre Dame

Abstract

The interplay between the image and comment on a social media post is one of high importance for understanding its overall message. Recent strides in multimodal embedding models, namely CLIP, have provided an avenue forward in relating image and text. However the current training regime for CLIP models is insufficient for matching content found on social media, regardless of site or language. Current CLIP training data is based on what we call "descriptive" text: text in which an image is merely described. This is something rarely seen on social media, where the vast majority of text content is "commentative" in nature. The captions provide commentary and broader context related to the image, rather than describing what is in it. Current CLIP models perform poorly on retrieval tasks where image-caption pairs display a commentative relationship. Closing this gap would be beneficial for several important application areas related to social media. For instance, it would allow groups focused on Open-Source Intelligence Operations (OSINT) to further aid efforts during disaster events, such as the ongoing Russian invasion of Ukraine, by easily exposing data to non-technical users for discovery and analysis. In order to close this gap we demonstrate that training contrastive image-text encoders on explicitly commentative pairs results in large improvements in retrieval results, with the results extending across a variety of non-English languages.

1. Introduction

Current publicly available Constrastive Language-Image Pre-Training (CLIP) [21] models suffer from a hidden problem which we term the "Descriptive-Commentative Gap" (hereafter DCG). CLIP models are trained on text we define as descriptive, the text is "presenting observations about the characteristics of someone or something", with the thing in this case being the image it is paired with. If the picture is an elephant, the text will simply say "an elephant standing next to a tree in a field", it provides no additional information or context beyond what the objects in the image represent. However, image-text pairs encountered in everyday life are rarely of a descriptive relationship. Social media image-text pairs are often of a commentative nature, with the text giving "an expression of opinions or offering of explanations about an event or situation" while the image is of the event or situation at hand. Given another photo of an elephant, a simple commentative text pairing could be "Kenya Hails Anti-Poaching Efforts in First Wildlife Census" [1]. The image itself is not directly reflected in the text. CLIP models often report accuracies of up to 99% on retrieval tasks where the text is description. When tested on posts where the text is commentary rather than description the accuracy drops by 60% on average. The transferability of the models to the "wild-west" of commentative captioning is poor, thus lowering the models abilities to be used by others for

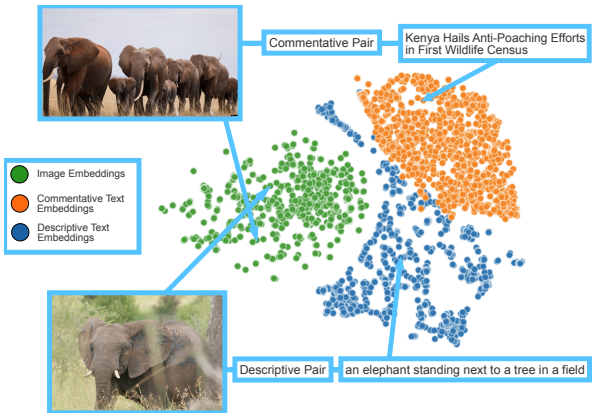

Figure 1. A TSNE [29] reduction of text and image embeddings from a baseline CLIP model. The two groups of texts are taken from the MSCOCO [17] data set commonly used to train CLIP models and taken from a data set of social media posts on the website Telegram [28]. While the images from the two groups share an area of the latent space (green) the text embeddings map to two distinct regions, showing the fundamental difference between commentative text (orange) and descriptive text (blue), which we term the description-commentary gap.

Figure 2. Example images and their caption(s) from each of the four data sets. The four data sets are split into two groups, descriptive (left) and commentative (right). There are five languages reflected in these data sets: English, Spanish, Portuguese, Ukrainian, and Russian.

downstream tasks.

Open-Source intelligence has gained an increasing amount of publicity as the amount of data available has continued to grow larger. Current estimates value the OSINT market at $6.25B and with a Compound Annual Growth Rate (CAGR) of 25 percent, valuations in 2033 reach $58.21B [13]. The invasion of Ukraine has provided many recent examples of OSINT in action, The Economist provides several examples; during the counter-offensive "amateur analysts on Twitter tracked the Ukrainian advance, almost in real time, by 'geo-locating' the images contrasted with a Russian soldier posting pictures from the frontlines." "His post included a geo-tag of the exact location. Ukrainian missiles later struck it" [2]. Of vital importance to OSINT operators is the ability to sift through massive amounts of data. Much of this work is done manually, but with the recent release of high quality multimodal models such as CLIP [21], there is hope that much of this work can be automated in the near-future, as unfortunately the DCG prevents us from doing so now.

In Fig. 1 one can see a visualization of the DCG. In orange are the commentative captions from a data set of Russian Telegram posts [28]. In blue are the descriptive captions from the MSCOCO [17] dataset. The two groups of texts are contained in two clearly demarcated regions of the latent space, while the images (in green) share a grouping in the space. Due to the differences in style between commentary and description, the embeddings of the two classes of text are different, leading to decreased accuracy on downstream tasks when the task is operating on commentary style data. Little discussion of the DCG has been had in technical literature to date. To help with this problem the paper makes the following contributions:

1. Defines and quantifies the Description-Commentary Gap.
2. Quantifies the gap across 5 languages and 3 social media sites.
3. Proposes solutions for closing the gap.
4. Lists experiments for testing models resulting in...
5. Several newly trained models achieving state-of-the-art retrieval accuracies on social media related downstream tasks, open-sourced on Huggingface [33].

2. Related Work

Several papers have raised the issue that not all social media comments are descriptions of their accompanying image, but aside from recognizing this issue it appears an open problem. Vempala and Preotiuc-Pietro claim that " little is known about how textual content is related to the images with which they appear" and describe a grid by which they go on to categorize text-image relations [30]. An image may either add to a tweets meaning or not, and text may be represented in an image or not. Sosea et al. also recognize that tweets and their images may not be of a descriptive character, with their categories consisting of: unrelated, similar, and complementary [26]. While both works recognize that this issue exists they make no attempt to formalize it nor theorize that current model training regimes fail to capture this difference. Additionally both works were published prior to the publication of CLIP and therefore do not and cannot, explore the possibilities this family of encoders provides.

The publication of the Contrastive Language-Image Pre-Training (CLIP) by Radford et al. was a leap forward in

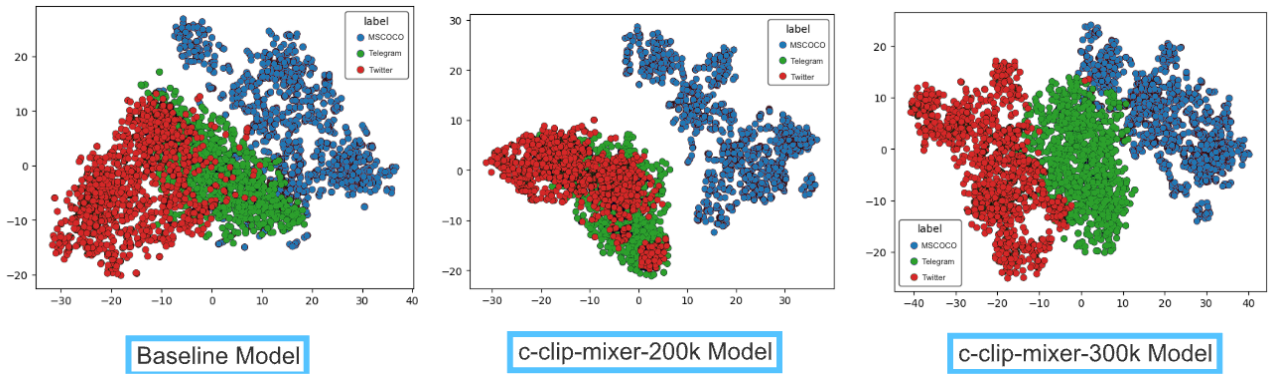

Figure 3. TSNE plots for the latent spaces of 3 models. On the left is a baseline model, showing a clear separation between the two commentative data sets and the descriptive data. In the middle can be seen what happens when the projection layers are trained on a combined set of Twitter and Telegram commentative data, a convergence of the two groups happens, which is expected. Surprisingly, when a model is trained on a mixed set of commentative and descriptive data a convergence is not seen, and instead order is applied to the three groups but they are kept essentially distinct.

multimodal modeling [21]. By jointly training the image and text encoders together on the pairs they easily achieved state-of-the-art performance on downstream tasks such as image retrieval. However, many of these initial models were closed-source and were entirely focused on English. To allay these concerns Carlsson et al. published several multilingual CLIP models, including one which they claim beats the original English-only CLIP model on several benchmarks [7]. Nils Riemers and Iryna Gurevych also published a multilingual model, using siamese bert networks [22].

The models used in this paper are merely built on top of pre-existing work. For the vision model we focus on transformers primarily as introduced by Dosovitskiy et al. [11]. More specifically we use the original clip-vit-base-patch32 model as released by openai [21]. We pair with this a variety of Bert models, originally introduced by Devlin et al. [10]. The primary focus was on RoBerta models [8] and distilBert models [24]. The highest accuracy was achieved using a distlBert model that had been trained with multilingual knowledge distillation as introduced by Nils Reimers and Iryna Gurevych [23].

Downstream multilingual tasks have been discussed in works such as "Towards Zero-shot Cross-lingual Image Retrieval" by Aggarwal and Kale [3] and by Nascimento et al [20]. Aggarwal and Kale discuss training regimes for expanding the models beyond English using pre-training on the text encoder, but as their data set is primarily an extension of MS-COCO [17], they focus only on descriptive pairs. Nascimento et al. focuses on social media data, Brazilian tweets surrounding a series of protests, but they focus on data curation to attempt to reduce the need for labelled data when training GNNs.

The description-commentary gap rears its head primarily when a CLIP model is applied to data collected from social media. [3] and [20] both focus on tweets, but social media at large, and especially memes, have taken a much larger seat at the table of online content understanding.

When considering social media data, image captioning is a closely related downstream task. Coming to the forefront in 2015 with the papers by Xu et al. [34] and Vinyals et al. [31], and more recently Cornia et al. [9]; the goal of image captioning is, given an image, generate caption for this image. MSCOCO features heavily in these papers, and as a result the captions generated are of a descriptive nature. By utilizing object detection frameworks, the papers all simply focus on the recognized objects in the image, yielding captions such as "a little girl in a pink hat is blowing bubbles." More recently, several papers have begun focusing on image captioning specifically as it relates to social media. Wang et al. [32] state that "the social image, associated with a set of user-contributed tags, has been rarely investigated for a similar task." Unfotunately they again simply fall into taking objects as "tags" for their tasks, resulting in image objects described in text as the returned captions. We maintain that the average post on social media does not take this form of image-caption relationship, and CLIP models provide a road forward in improving this modeling.

Further downstream tasks such as retrieval and clustering for social media has also been well treated in prior literature. Zannettou et al. [35], Dubey et al. [12] both cover the clustering of image macros. Unfortunately they fail to bridge the gap of multi-modality, with both raising it as an area of future work. Beskow et al. [4] attempts to bridge the gap between multiple modalities using deep-neural networks but restricts their definition of meme to a picture with superimposed white text in impact font and/or text placed in a white space over a picture. The three previous approaches are all supervised methods requiring strict categorization of the image content, Theisen et al. [28] states that an unsupervised approach yields better results when treating with

social media data as virtually no categories of images can be assumed and new categories are constantly popping up, especially surrounding new events. It is with this in mind that we propose training on large unsorted collections of social media image-pairs in order to understand and close the descriptive-commentary gap.

3. Methodology

Descriptive-Commentative Gap: We hypothesize that there is a quantifiable difference between text used for training CLIP models and text that would appear in downstream tasks relating to social media. We call this difference the "Description-Commentary Gap", hereafter DCG. In Figure 2 we can see two examples of image-text pairs, one taken from the COCO data set frequently used to train CLIP models and another taken from the social media site Telegram. The left pair, from COCO, is what could be described as a "descriptive" pair. The text simply describes or lists the objects that appear in the image. Contrasted to this is the right pair with the text being, at the surface level, almost completely disconnected from the image. They share no objects. This text is of a "commentative" nature with the text adding additionally context to the image and relating it in some way that is not immediately obvious. These two pairs illustrate the DCG.

We hypothesize that the existence of the DCG is non-trivial and significantly reduces accuracies on downstream tasks involving social media data, which is almost entirely of a commentative nature. Humans can intuit this gap, but it has yet to be actually quantified in prior literature and is therefore worth exploring further. Figure 1 shows that the DCG also appears in the latent space of CLIP text embeddings, spanning across a number of different possible text models, languages, and datasets. Descrpitive embeddings are a discrete group showing a degree of separation from commentative embeddings.

Theisen et al. ends their paper with the belief "that using all available context is of the utmost importance for future studies in this area." The release of CLIP models provides an avenue forward in connecting the multimodal aspects of social media posts. However the DCG is a yet unanswered question in the way, reducing accuracy on tasks relating to social media data and preventing non-technical works from benefiting from this technology. To help alleviate this problem a number of different, publicly available, models were trained which we hope can be of use to those wishing to model social media content multimodally.

Data and Procedures: In order to establish that the DCG exists, we test three baseline models. OpenAI's publicly available CLIP model [21], a multilingual model from the M-CLIP team [7], and a multilingual model from the SentenceTransformers group [22]. For this task we use four data sets, two descriptive data sets, one in English and a multilingual set and two commentative data sets, in languages included in the multilingual models training data. We include three non-english language data sets to demonstrate that the DCG is language-agnostic. The two descriptive data sets are Microsoft's Common Objects in COntext (COCO) [18], a standard image-caption data set and Cross-modal 3600 (XM3600) [27], a data set from google released specifically to test multilingual multimodal models in a variety of languages. For our task we choose captions from five languages: English, Spanish, Ukrainian, Russian, and Portuguese.

Training Argument	Value
Epochs	50
Early Stopping	True
Batch Size	32
Learning Rate	5e-5
Optimizer	Adam [16]
Max Seq. Length	128

Table 1. Basic training parameters used in all C-CLIP models.

The two commentative data sets are tweets surrounding a series of protests in Brazil [20] and Telegram posts related to the ongoing war in Ukraine [28]. Figure 2 shows several examples from both data sets and hopefully allows a reader to intuit the fundamental difference between the caption-image pairings across the two categories. These data sets are chosen for a variety of reasons. In addition to their essentially commentative captioning, being directly scraped from social media, they both span a variety of languages. Additionally, they both cover events that have been of particular interest in the OSINT space. According to the Verge, "Telegram has become a window into the war." and while not a common source of social media in western countries "in Russia, Telegram has become sometimes the only source of information amid stifling government censorship. Across the border, the platform has become a lifeline for Ukrainians trying to keep safe from Russia's attacks and track troop movements. And for the rest of the world, Telegram has become the window into a war that has destabilized the world." [6] The January 8th protests that rocked Brazil were another flashpoint for OSINT operators, with people working on twitter data to help authorities track down and unmask protesters [19]. We hope that by training models on data sets of already high importance in the OSINT community we can immediately begin helping these efforts.

The data sets are also large enough to allow for the training of multiple models and a full suite of testing to be run. The Brazilian Twitter data set has 203,781 image-text pairs. The Russian Telegram data set is much larger, yielding 4,766,631 image-text pairs.

After demonstrating the existence of the DCG qualita-

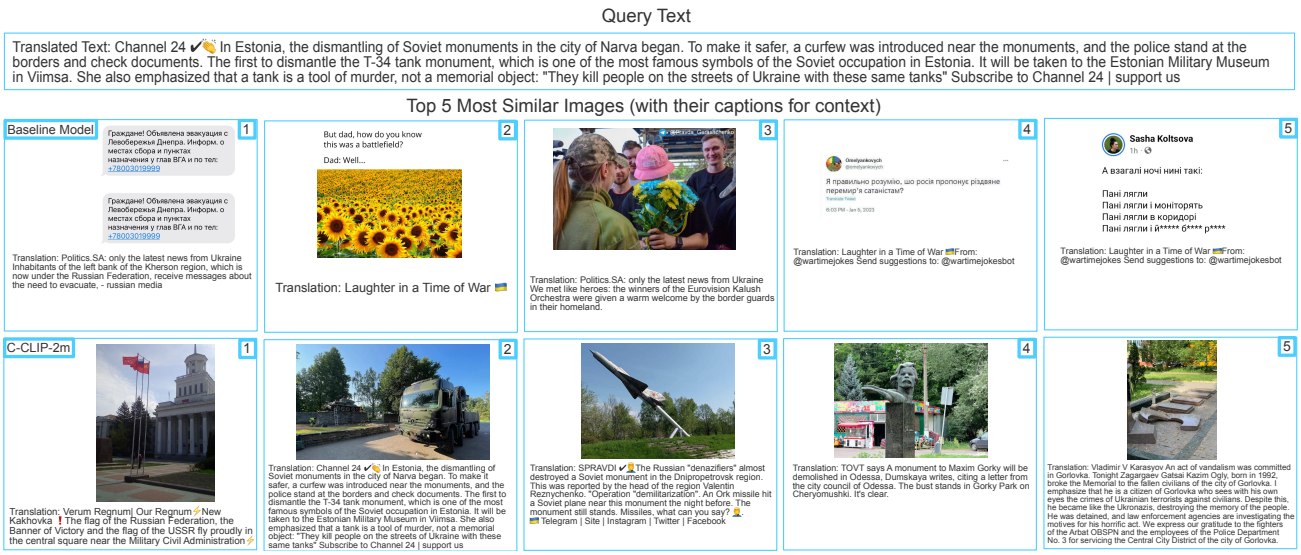

Figure 4. The top 5 most similar images as measured by two CLIP models from a population of 1000 image-text pairs for a query text. The baseline model returns only spurious results, with no clear connection between the query text (top) and the returned images (middle). However the C-CLIP-2m model (bottom) returns 4 images of monuments and 1 image of a soviet symbol (the flag), both of which are mentioned in the first query text. Qualitatively it can be seen that C-CLIP-2m learns quite well the relations between commentative text and their images. We provide also the translated captions that accompanied each image, to give context for the reader. The original, untranslated text, is available in the supplementary material.

tively and quantitatevly, we trained a number of models with the intent to improve performance of CLIP models on tasks using commentative data. Using the VisionTextDualEncoder (VTDE) [15] module from HuggingFace one can easily pair pre-trained image and text models.

The vision model is the same used in prior works [21] [7] [22], being a ViT with an output dimension of 768. This is paired with the DistilBert model from Riemers et al. [22], also with an output dimension of 768. The VDTE forward function had to be modified to correctly use the DistilBert model. Instead of directly using the built in text_projection function, the last hidden state had to be averaged over the number of the tokens in the input string prior to the call. These were then used to calculate the loss. Via the two projection layers added on top of the dual models the output dimensions were projected into the shared latent space and reduced from the original dimension of 768 to 512.

The VTDE then places on top of the two models two projection layers, which then must be further trained on a downstream task.

To train the models, we use HuggingFace's Trainer class [14] with the arguments outlined in Table 1. These settings are rather boilerplate and were based off the work done in Bianchi et al. [5] where the Italian authors fine-tuned a CLIP model to extend to their language. and the Github repository [25].

Several training regimes were tested. The first is training a single model on a single data set style, for example training a Russian Telegram model specifically on collections of Russian Telegram data of varying sizes. Additionally tested was mixing training data in both even and uneven splits, attempting to understand whether the DCG is truly cross-site and multilingual, and what the impacts on performance may be if a model was trained on a mixture of Brazilian Tweets and Russian Telegram, or Russian Telegram and descriptive data.

Models were trained at 10,000 image-text pairs and 100,000 image-text pairs. For Russian Telegram, due to the large size of the data set, 1 million and 2 million pairs were trained as well. In addition to these training splits, validation splits of 20,000 pairs were left out. A final set of 111,000 pairs was used for testing. 1,000 for 10 trials at a population of 100 pairs, 10,000 for 10 trials at 1,000 pairs, and 100,000 for 10 trials with 10,000 pairs. The largest model (C-CLIP-tele-2m) took approximately 28 hours to train. As a rough rule of thumb, every 10,000 pairs added 10 minutes to the training time (on a Titan X with 12GB of VRAM).

To establish the general performance of our models we then calculate the retrieval accuracy of each model. For population sizes of 100, 1000, and 10000 we compute the pairwise cosine similarity of all image text pairs. Then at recalls of 1, 5, 10, and 25 we compute the retrieval accuracy for each artifact. These accuracy results were averaged across ten trials for each population size. These were then

Model Retrieval	MS-COCO				XM3600				Russian Telegram				Brazilian Tweets			
Accuracy (all-pairs)	@1	@5	@10	@25	@1	@5	@10	@25	@1	@5	@10	@25	@1	@5	@10	@25
OpenAI CLIP [21]	57.3%	96%	99.2%	100%	-	-	-	-	-	-	-	-	-	-	-	-
Sen.-Tran. CLIP [22]	65.89%	89.1%	96.1%	99.70%	41.06%	69.13%	78.86%	88.73%	6.06%	13.18%	17.08%	24.08%	10.40%	19.53%	24.55%	32.79%
M-CLIP [7]	64.93%	74.44%	83.69%	92.19%	72.6%	92.89%	96.36%	98.87%	12.99%	24.55%	30.75%	39.80%	5.71%	11.78%	15.32%	21.28%
C-CLIP-2M	4.4%	7.7%	13.5%	24.5%	3.8%	13.27%	19.99%	34.9%	**29.55%**	**57.25%**	**67.26%**	**79.41%**	1.37%	4.93%	7.63%	13.8%
C-CLIP-Mixer-200k	3.6%	6.3%	9.2%	18.8%	3%	10.87%	17.49%	30.93%	7.34%	21.30%	29.14%	42.73%	**13.69%**	**34.28%**	**45.35%**	**60.99%**
C-CLIP-Mixer-300k	18.4%	24.8%	34.5%	51.9%	6.4%	18%	28.5%	45.6%	7.9%	18.7%	27.8%	37.7%	12.37%	30.83%	41.93%	56.87%

Table 2. Retrieval accuracy results for top-performing models across the four data sets, with a population size of 1000 image-text pairs. The three baseline models achieve fantastic results on the descriptive data sets of MS-COCO and XM3600, but dismal accuracies on the two commentative data sets. Contrasted to this are the two C-CLIP models getting the highest results on the two commentative data sets. It shows that the commentative image-text pair space can be learned, but does not extend back to the descriptive space. There appears to be a trade-off required between the two types of data sets. The bottom row is our attempt at training a model to perform equally well on all data sets.

compared against the three baseline models, tested in the same manner. One limitation of the original CLIP model is that it is only trained on English data and thus the performance on MSCOCO is only to establish that the multilingual baseline models achieve similiar accuracies and therefore allow for a reasonable point of comparison to our work.

4. Experiments and Results

With the four datasets chosen for testing the initial hypotheses as outlined in the Methodology the results in Table 2 are produced. They show a clear degradation in performance for baseline CLIP models when they move from a descriptive task to a commentative one. Additionally, these problems exist both across languages and social media sites, demonstrating that the DCG describes something fundamentally present in how social media is used differently when compared to the types of image-text pairs that CLIP models are commonly trained on.

The degradation in results is larger than one might expect, with both multilingual baseline models dropping 61.78% and 65.61% percentage points at a recall accuracy @10 for the Russian Telegram data and 54.31% and 81.04% percentage points on the Brazilian Tweet data set when compared to XM3600. All results are the average accuracy over 10 trials, with a population size of 1,000 image-text pairs, expect for XM3600, which having only 3600 pairs means that at a population size of 1,000 produces only 3 test splits. The standard deviation across all trials on all data sets on all experiments averaged 2.34% at recall 10, making the difference between 3 trials and 10 trials negligible when considering differences of approximately 30 percentage points.

The first three rows in Table 2 demonstrates rather conclusively that the DCG results in decreased accuracy of models trained purely on descriptive data when applied to commentative tasks, estimates the magnitude of its effect, and shows that it persists across languages and (at least two) social media sites.

These baselines show that there is much need for improvement on downstream tasks vital to operators in OSINT spaces.

Our best results on the two commentative data sets are shown in the bottom two rows. Training on the specific data set massively increases performance. While perhaps not particularly surprising, it's important to show that the commentative image-text difference is possible to learn. Our best result on the Russian Telegram data is 36.51 percentage points higher (again @10) and for the Brazilian data, 28.2 percentage points higher than the best baseline.

Quantitative results give only one side of the story, the qualitative side is also important. The retrieval accuracy @5 is only 57.25% on the Russian Telegram data set. However if one considers Figure 5 it can be seen that the other 4 results that are not a direct match are still highly relevant to the query text, in both sets of retrieval results. The top results contain 4 images of monuments and one of a soviet symbol, all of which are mentioned in the query text. The second set of results is for a piece of text about a petition that is being sent to President Zelensky to review. 4 out of the 5 returned results are about various petitions that have been filed since the invasion began. It seems reasonable that even if the directly matching image is not returned, OSINT operators would find useful information in non-match results. If a journalist or reporter were writing a piece on junk petitions filed during the on-going invasion, having a tool to find all posts related to petitions seems a useful tool. The achieved accuracy scores are state-of-the-art, but we contend that the results are even more useful than these scores imply, especially to real-world operators in OSINT.

We also report the "backwards" accuracy of the new models, I.E. how the models trained on commentative data perform on descriptive tasks. In this we can see the DCG happening in reverse, demonstrating that there is indeed a gap. Training on descriptive data only means low accuracy when tested on commentative data and training on commentative data only results in low accuracy when tested on descriptive data. The DCG appears to be a two way street.

In addition to experiments exploring the nature of the DCG, we also report results on a variety of training variations used on our models. Table 3 shows the results of all

Telegram Retrieval	Pop=100				Pop=1,000				Pop=10,000			
Accuracy (all-pairs)	@1	@5	@10	@25	@1	@5	@10	@25	@1	@5	@10	@25
M-CLIP (Baseline) [7]	23.40%	38.80%	47.70%	66.70%	6.06%	13.18%	17.08%	24.08%	2.88%	5.85%	7.82%	11.02%
C-CLIP-tele-2m	**54.50%**	**81.30%**	**88%**	**96.29%**	**29.55%**	**57.25%**	**67.26%**	**79.41%**	**10.25%**	**26.87%**	**36.59%**	**51.11%**
C-CLIP-tele-1m	39.20%	63.40%	73.30%	86.10%	23.30%	46.47%	56.14%	67.90%	8.02%	21.17%	29.22%	41.48%
C-CLIP-tele-100k	28.90%	51.10%	63.40%	79.20%	13.40%	30.60%	40.05%	53.55%	3.32%	10.40%	15.50%	24.72%
C-CLIP-tele-10k	16.09%	34.89%	48.59%	68.10%	4.60%	14.29%	21.29%	33.72%	0.86%	3.17%	5.33%	10.12%

Twitter Retrieval	Pop=100				Pop=1,000				Pop=10,000			
Accuracy (all-pairs)	@1	@5	@10	@25	@1	@5	@10	@25	@1	@5	@10	@25
M-CLIP (Baseline) [7]	18.9%	38.6%	48.2%	68.9%	5.71%	11.78%	15.32%	21.28%	4.11%	7.58%	9.9%	13.39%
C-CLIP-twit-100k	28%	59.9%	73.5%	90.1%	7.27%	21.47%	31%	46.7%	1.14%	3.93%	6.35%	11.39%
C-Clip-twit-10k	14.40%	38.5%	53.5%	76.1%	4.73%	11.87%	16.4%	27.97%	0.84%	2.57%	4.05%	6.88%
C-CLIP-Mixer-300k	38.09%	69.4%	82.19%	92.9%	12.37%	30.83%	41.93%	56.87%	2.15%	6.77%	10.15%	16.67%
C-CLIP-Mixer-200k	**39.4%**	**71.7%**	**82.7%**	**93.5%**	**13.26%**	**33.3%**	**45.13%**	**60.47%**	**2.26%**	**7.29%**	**11.15%**	**18.94%**
C-CLIP-desc-mixer-200k	38.6%	71%	81.7%	92.89%	13.3%	31.76%	43.36%	58.56%	2.18%	7.07%	10.82%	17.8%

Table 3. Retrieval accuracy results for all models trained on the two commentative data sets (top: Telegram, bottom: Twitter). The best performing model on the telegram data was C-CLIP-tele-2m, perhaps unsurprisingly. Increasing the size of the training set appears to yield increasing results. The highest accuracy for the twitter data set was actually a heterogenous mixture of twitter and telegram data. Mixing twitter and descriptive data also did well, but not as well as supplementing with commentative data.

models on the Russian Telegram data set and the Brazilian Twitter data set. The first row shows the baseline with the highest score on the data set for reference. For Russian Telegram the highest accuracy was achieved by the model that was given the most training data, in this case 2 million image-text pairs. All training pairs were filtered to ensure that the associated text had at least 5 "words" (quotations being used on words as there were many emojis in the data, potentially another interesting avenue to explore).

Varying the amount of training data was not the only vector through which accuracy could be increased. The selection of training data is an active area of research and experiments were performed to see to what extent mixing and matching training data would have on downstream accuracy. Due to the relatively small size of the Brazilian Twitter data set (to successfully run 10 distinct trials of the 3 populations you need 111,000 test pairs) the maximum number of Brazilian twitter pairs the model could be trained on was limited to 100,000. This resulted in an accuracy of only 31% (pop=1000, recall=10). While this result is nearly twice the baseline accuracy, it seemed likely that supplementing the training data with other image-caption pairs could lead to an increase. Four different models were tested: one with 100,000 Brazilian tweets and 100,000 Russian telegram posts, a training set with 100,000 Brazilian tweets and 100,000 descriptive pairs, a set with 100,000 Brazilian tweets, 100,000 descriptive pairs, and 100,000 telegram pairs, and one with 100,000 Brazilian tweets, 100,000 descriptive pairs, and 2m telegram pairs. Out of these four, the model that was trained on an equal split of Brazilian Twitter and Russian Telegram data performed the best, increasing to 45.13%. This was closely followed by the other two models that had even data splits, the one trained on 100,000 pairs of both descriptive and Brazilian data and the model trained on Brazilian, Russian, and descriptive data. Interestingly, the model that saw the most data, but had a large skew in the data away from the Brazilian twitter data set performed the worse of the four (though again still better than the baseline). This seems to imply that while descriptive and commentative data are two distinct categories, there exist sub categories of the two that are more specific to certain languages and/or sites. Supplementing with data of a similar type (commentative/descriptive) may help but you can't entirely replace the in-task data with data of the same class.

The model that was trained on Brazilian twitter, Russian Telegram, and descriptive data was tested to see how well its accuracy extended across all four data sets. The goal is a model that achieves universally, uniformly, high accuracies on all tasks. Unfortunately this doesn't seem to be the case. As can be seen in Table 2, the C-CLIP-mixer-300k (last line) doesn't achieve particularly high accuracy results. While one could argue that the differences in amount of training data could potentially make up for this, the accuracy on specifically the Russian Telegram data set is much lower than its counterpart model that was trained on the Russian data only. Compared to the C-CLIP-tele-100k model seen in Table 3, the C-CLIP-mixer-300k got 12.25 percentage points lower retrieval accuracy (pop=1000, recall=10). This is rather surprising, as supplementing with more data increased accuracy on the Brazilian Twitter data set, but seems to lower it on the Russian Telegram data set. As hypothesized above, there seem to be sub-categories in the overarching classes of descriptive and commentative that contain their own quirks. While our models achieve state-of-the-art accuracy scores across the board, further research into this area is greatly needed.

Brief experiments were also run to explore the extensibility of models on data sets they were not trained on. The results can be seen in Table 4. Shown are the differences between the cosine similarity of a true pair and the average

of all false matches, for 1000 image-text pairs. Unsurprisingly we see an increase in this difference as C-CLIP-tele is trained on increasingly more data and this is reflected in increasing accuracy scores as can be seen in Table 3. What is interesting is the growth (or lack-there-of) in the differences on the MSCOCO and Brazilian Twitter data sets. If types commentative data were truly indistinguishable from themselves then we could expect the model to improve its results on the Brazilian data set at a similar pace as the Telegram data set. However we instead see no such increase. This seems to support the intuition stated above that there exist sub-sections of data falling under the commentative super-class. This is reflected qualitatively in the TSNE plots shown in Figure 3 but required the additional quantitative justification provided here. Therefore it seems that if one wishes to have an accurate C-CLIP model the best data to train on is data specific to their task.

If one has access to a limited amount of data for that task, supplementation of data does help. These results can be seen in the bottom three rows of Table 4. Training on a mixture of balanced commentative data yields the best result, and is actually the highest performing model on the Brazilian Twitter data set (shown in 2). Supplementing this model further with descriptive data actually decreases the difference between positive and negative similarities. With the Twitter models we also see the trend of a lack of increase in the models performance on other commentative data sets, further supporting the notion that there exist subtle differences in commentative data. The cause of these differences are unknown but could be as far ranging as language specific to the event or region, or just different emoji usage by different cultures. Further work in this area is required.

The DCG implies that there should not be an increase on the MSCOCO data set when training on commentative data, but we instead see a small increase in the differences on the MSCOCO data set, even when a model is not trained on descriptive data. We believe that the underlying models already having a strong baseline in descriptive tasks is able to slowly appear as the projection layers get more and more finely tuned with increasingly large amounts of commentative data. However as the mixer-300k and desc-mixer-200k models show, the best way to increase the difference is simply to have descriptive data present in the training data.

5. Conclusions

By training CLIP models on commentative text rather than only descriptive text we can improve the accuracy of these models in downstream retrieval tasks on social media data. While we report state-of-the-art retrieval accuracies when compared to baseline multilingual clip models, the usefulness of the results is better than the accuracy implies. Allowing non-technical people to explore large collections of unsorted and unlabelled data and discover semantically similar results with higher speed and accuracy is key to improving OSINT operations. Figure 5 shows that results that are not an exact match, as determined by the actual pairing from social media, are still highly relevant and further integration of C-CLIP models into pre-existing downstream tech is an exciting prospect.

C-CLIP Model	MSCOCO	Telegram	Twitter
tele-10k	0.1172	0.2387	0.1081
tele-100k	0.2578	0.3422	0.1359
tele-1m	0.2589	0.3847	0.1147
tele-2m	0.2056	0.4297	0.0906
twit-10k	0.0964	0.0764	0.1910
twit-100k	0.0866	0.0491	0.2499
desc-mixer-200k	0.4826	0.0800	0.2981
mixer-200k	0.1723	0.2771	0.3438
mixer-300k	0.4763	0.2468	0.3083

Table 4. The average difference between the cosine similarity of a correctly matching pair and the average of all incorrect pairs for a population size of 1000. The ideal model has a large difference between the similarities of a positive pair and a negative pair and displays this difference across all three data sets.

Limitations and Future Work. When training models the limitations are of course primarily data related. Training data for the C-CLIP models was randomly selected from the data set, and a better curation method would likely lead to better results. Additionally the creation of more social media data sets in a variety of languages is crucial to improving understanding in languages outside of English. This paper merely trains the models and measures baselines, much future work is yet to be done in integrating these models into OSINT tools in order to aid those involved.

References

[1] Kenya Hails Anti-Poaching Efforts in First Wildlife Census — voanews.com. https://www.voanews.com/a/africa_kenya-hails-anti-poaching-efforts-first-wildlife-census/6210170.html. [Accessed 15-08-2023]. 1

[2] Open-source intelligence is piercing the fog of war in ukraine, Jan 2023. 2

[3] Pranav Aggarwal, Ritiz Tambi, and Ajinkya Kale. Towards zero-shot cross-lingual image retrieval and tagging, 2021. 3

[4] D. Beskow, S. Kumar, and K. M. Carley. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing and Management*, 57(2), 2020. 3

[5] Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. Contrastive language-image pre-training for the italian language. *arXiv preprint arXiv:2108.08688*, 2021. 5

[6] Masha Borak. Telegram has become a window into war, Jul 2023. 4

[7] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France, June 2022. European Language Resources Association. 3, 4, 5, 6, 7

[8] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020. 3

[9] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 3

[12] A. Dubey, E. Moro, M. Cebrian, and I. Rahwan. Memesequencer: Sparse matching for embedding image macros. In *In Proceedings of the International World Wide Web Conference*, 2018. 3

[13] Future Market Insights Global and Consulting Pvt. Ltd. Open-source intelligence (osint) market is anticipated to surpass us$ 58.21 billion by 2033, at a cagr of 25%: Latest expert analysis by future market insights, inc.., Apr 2023. 2

[14] HuggingFace. 5

[15] HuggingFace. Visiontextdualencoder. 5

[16] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 4

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 1, 2, 3

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 4

[19] Maverick. Bad trip: how osint tools unmasked people involved in brazil's january 8 riots, Jan 2023. 4

[20] José Nascimento, João Phillipe Cardenuto, Jing Yang, and Anderson Rocha. Few-shot learning for multi-modal social media event filtering, 2022. 3, 4

[21] 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. Learning transferable visual models from natural language supervision, 2021. 1, 2, 3, 4, 5, 6

[22] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 3, 4, 5, 6

[23] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation, 2020. 3

[24] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. 3

[25] sgugger. Visiontextdualencoder and clip model training examples, 2023. 5

[26] Tiberiu Sosea, Iustin Sirbu, Cornelia Caragea, Doina Caragea, and Traian Rebedea. Using the image-text relationship to improve multimodal disaster tweet classification. In *ISCRAM 2021 Conference Proceedings – 18th International Conference on Information Systems for Crisis Response and Management (pp. 691–704)*, 2021. 2

[27] Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset, 2022. 4

[28] William Theisen, Daniel Gonzalez Cedre, Zachariah Carmichael, Daniel Moreira, Tim Weninger, and Walter Scheirer. Motif mining: Finding and summarizing remixed image content. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1319–1328, 2023. 1, 2, 3, 4

[29] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 1

[30] Alakananda Vempala and Daniel Preoţiuc-Pietro. Categorizing and inferring the relationship between the text and image of Twitter posts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2830–2840, Florence, Italy, July 2019. Association for Computational Linguistics. 2

[31] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014. 3

[32] Leiquan Wang, Xiaoliang Chu, Weishan Zhang, Yiwei Wei, Weichen Sun, and Chunlei Wu. Social image captioning: Exploring visual attention and user attention. *Sensors (Basel)*, 18(2):646, Feb. 2018. 3

[33] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. 2

[34] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015. 3

[35] S. Zanneettou, T. Caulfield, J. Blackburn, E. D. Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil. On the origins of memes by means of fringe web communities. In *ACM Internet Measurement Conference*, 2018. 3

Figure 5. The qualitative results with the original, untranslated captions that were posted with the image. Google translate was used for the translation, so slight inconsistencies may exist. Hence the need to include captions in the original.