Sky + Fire = Sunset
Exploring Parallels Between Visually Grounded Metaphors and Image Classifiers

Yuri Bizzoni
Saarland University
yuri.bizzoni@uni-saarland.de

Simon Dobnik
University of Gothenburg
simon.dobnik@gu.se

Abstract

This work explores the differences and similarities between neural image classifiers' mis-categorisations and visually grounded metaphors - that we could conceive as intentional mis-categorisations. We discuss the possibility of using automatic image classifiers to approximate human metaphoric behaviours, and the limitations of such frame. We report two pilot experiments to study grounded metaphoricity. In the first we represent metaphors as a form of visual mis-categorisation. In the second we model metaphors as a more flexible, compositional operation in a continuous visual space generated from automatic classification systems.

1 Introduction

Visually grounded metaphors are metaphors that function on the visual similarity between two objects. Humans use visually grounded metaphors to describe scenes or objects in a vivid way: we could call a large person an elephant, describe somebody's blue eyes as a clear summer sky; and so forth. The basic idea is that using a metaphoric element to "overlay" on the described object creates, in the imagination of the reader or listener of the metaphor, a stronger and effective mind picture. At the same time, the mechanisms and workings underlying metaphors remain unclear. Grounded metaphors are among several categories of metaphors that work on the interplay of all five senses plus the abstract dimension of language. Some of these metaphors and similes are harder to model: the precise reason why a *cold voice* creates the idea of a specific tone and sound of voice, or why some other synesthetic expressions like *a blue music* work in the human brain is difficult to define. Visually grounded metaphors, unlike metaphors that draw from several senses, could be easier to model: and we could try to study them by applying the visual models used for image captioning in NLP. The central question of this paper is whether, and to to what degree human visual metaphors can be reproduced using image features trained in the image captioning scenario. This pilot study defines some possible lines of exploration through two small scale experiments that lead to further research in this area of connecting language and vision and modelling of metaphorical language. More data and analyses will be necessary to further deepen the topic.

2 State of the Art

A large bibliography has discussed the relation between language and perception and the way linguistic meaning is expressed linguistically and non-linguistically. This is done by grounding word meaning in visual perception (Siskind, 2001) and testing the compositionality of visually-enriched language representations (Gorniak and Roy, 2004), and more recently even using sensory-motor robotics to model lexico-grammatical patterns (Zhong et al., 2019), often for the construction of multimodal or dialogue agents (Roy, 2005; Roy and Reiter, 2005), under the general assumption that many aspects of language cannot be captured without extra-linguistic information (Barsalou, 2008). Despite this and the existence of consolidated linguistic work about the mechanisms of the underlying metaphor processing (Lakoff and Johnson, 2008), the research on the topic of metaphor and grounded language models is quite scarce and has focused on the use of visual features to identify metaphors (Shutova et al., 2016). The common practice in the domain is currently to attempt metaphor identification and modelling through linguistic data only (Zhang and Barnden, 2013; Do Dinh and Gurevych, 2016; Bizzoni and Lappin, 2018) Neurolinguistic studies such as De-

Proceedings of the Second Workshop on Figurative Language Processing, pages 126–135

https://doi.org/10.18653/v1/P17

sai et al. (2013) show that the literal properties of terms are still activated if those terms are used metaphorically, confirming the idea that metaphors create a compositional feature transfer between source and target: for example, senso-motory verbs used in metaphorical ways still activate their "normal" senso-motory paths in the brain, a behaviour that distinguishes metaphors from idioms (Lai and Curran, 2013; Kemmerer, 2015). Other studies, such as Zanolie et al. (2012), show that some abstract concept, like *power*, seem linked, in the brain, to a spatial feature, like the vertical up-down dimension, which would, according to the study, account for the spatial metaphors of power that tend to visualise hierarchy on a vertical axis. In short, several studies in neurolinguistics support the idea that many metaphors do indeed rely on sensory knowledge (Lacey et al., 2012). Reproducing these mechanisms in computational models is the main idea behind our project. The goal of our study is to examine to what degree these observations are reflected in the performance of image classification deep neural models trained on images. The way image classification models both represent and categorise pictures can help us understand better to what extent grounded compositional metaphors are actually grounded (Section 4) and compositional (Section 6).

3 Models

For our study we use pre-trained visual object classification models available in Keras (Chollet et al., 2015). The two main models used in this work are ResNet50 (He et al., 2016) and VGG16 (Simonyan and Zisserman, 2014). For comparison we also use VGG19 (Simonyan and Zisserman, 2014) and InceptionResNetV2 (Szegedy et al., 2016). In the versions we use for this study, the networks were pre-trained on 1,000 object categories from ImageNet. Unlike the systems used for articulated image description, these models generate single token captions. (Deng et al., 2009). All models operate through in three steps: (i) the model takes an image as input; (ii) using the pre-trained weights the model transforms the image into a vector that represents its main features; (iii) the model's final layer operates a prediction or classification over such a vector. Since the categories are limited (here to 1,000) and the models imperfect, mis-categorisations occur.

4 Mis-categorisations

A classifier presented with an image will output a list of possible captions or descriptions, with confidence scores that indicate the similarity between the output category and the image. For example, if presented with a picture of a bird, ResNet50 will output the following probabilities:

1. brambling 0.473
2. house-finch 0.155
3. water ouzel 0.090
4. junco 0.005
5. robin 0.053

If presented with an airliner, the model will output captions like *airliner* (0.93), *warplane* (0.03), *airship* (0.001). If we use a good object classifier on a clear instance of an object that is well presented in its training data, the result is a high-probability prediction of the object category, followed by low-probability categories that share a decreasing number of features with the target object in the image. For example, if presented with a picture of an Indian elephant, ResNet50 outputs the following probabilities over categories:

1. Indian elephant 0.95
2. tusker 0.03
3. African elephant 0.01
4. triceratops 2.1814798e-05
5. water buffalo 1.0476451e-05
6. warthog 6.76768e-06
7. hippopotamus 6.4546807e-06
8. ice bear 3.6104445e-06

The gap in probabilities between the first and the second prediction is large, and the probabilities after the 4th item are insignificant. It is possible to notice how in all cases the model's predictions are based on the main features of the elephant's shape: the output's classes share some visual similarities with the Indian elephant, in a decreasing order of overlap. The model predicts, in order, the Indian elephants; other kinds of elephants; and other animals that in ResNet50's ontology share important properties with the Indian elephant. The reader can observe that the suggested alternative species have common characteristics of being massive four-legged animals and in many cases displaying prominent tusks. We can compare the network's behaviour with the strategies a human

would adopt in the attempt to describe a specific animal to someone who has never seen it before: looking for other animals sharing some similarities with it. This mechanism becomes even more evident if we present the network with a category that was absent from its training data, or that was too rare in the training to allow the model to generalise on its features. Let's take for example Figure 1, an image of a fire on a dark background.

Figure 1: Fire!

The Keras' pre-trained model of ResNet50 lacks the category "fire" in its training dataset, and if presented with this picture, the model returns the following list of probabilities over categories: *stove* (0.85), *fire screen* (0.14), *dutch oven* (0.002).

The network identifies categories of objects in which fire is a likely component. Being unable to figure out the object's category with confidence, the model returns captions drawn from categories of objects that share some properties with the presented image. The picture's background and style play important roles as well. For example, ResNet50 categorises the leftmost image in Figure 2, a drawing representing a dragon, as *comic book* (0.28) or *laptop* (0.08), and the rightmost image Figure 2, a statue of dragon, as *pedestal* (0.61), *fountain* (0.38) or *palace* (0.0005). In both cases the classifier focuses on the "style" of the object - a drawing in the first case, a statue in the second case - to attempt a low-probability classification.

Figure 2: Two dragons

Mis-categorisations can happen if the object appears in a new or unusual way that confuses the network. For example, ResNet50 has various *bridge* categories in its ontology such as *steel arch bridge*, but it tends to classify pictures of small bridges mirrored in the water as *viaduct* or *lakeside*, being confused by the elements present in the image. Similarly, it labels an aerial picture of a large modern bridge crossing the sea as *bannister* (0.44) or *dam* (0.03). As in the previous cases, the model looks for objects that pertain to the same field or conceptual area of the difficult image: to categorise bridges the models look for dams, viaducts and so on. In the same way, if presented with a picture of a church, the model suggests *church* (0.93), *bell cote* (0.02), *monastery* (0.02), and it tries to describe the Burj Khalifa as a *mosque* (0.13), a *palace* (0.08), a *bell cote* (0.07) or an *airship* (0.06). The last association is of particular interest for our study, since the model moves out of the conceptual domain of the picture to look for similarities in different areas. But what happens if the model is presented with an image belonging to a conceptual domain completely unknown? For example, "our" ResNet50 lacks in its pre-trained categories every thing pertaining to the sky: clouds, planets and stars are absent from its ontology. If presented with a classical image of Saturn, the model predicts *candle* (0.81). We suspect that the reason of this unexpected elaboration lies in the colour of Saturn's atmosphere, that is similar, in some pictures, to a candle's wax colour.

Figure 3: A guy.

In other situations, the network focuses on background elements that it recognises. For example, this model is also not trained on people: it cannot label a person as *person*. If presented with the image of a person in a bathtub, the suggested captions are *bathtub* (0.08), *bath towel* (0.07), *tub* (0.06). For similar reasons, and showing some politically incorrect bias, the model labels the person in Figure 3 as *prison* (0.05), *jean* (0.02) and *barrow* (0.06). In a picture representing a breaking storm over the sea, ResNet50 - not "knowing" what a storm is - pics the peripheral elements: *breakwater* and *pier*. The abundant literature on object classification has

duly noted these systematic mistakes (Wang et al., 2009; Xu et al., 2015).

This behaviour also applies to different models. Many of VGG16's predictions in front of unknown or puzzling objects mirror those of ResNet50: *dam* (0.22) for the modern bridge picture, *pedestal* (0.55) for the dragon in Figure 2, *viaduct* (0.24) after *triumphal arch* (0.33) for the bridge mirrored in the water. For known objects, VGG16's second and third best guesses align with the ones produced by ResNet50: for example, the *church* prediction for the church is followed by *monastery* (0.02) and *bell cote* (0.02). In the case of the Burj Khalifa, the first prediction remains *mosque* (0.33), but VGG16 seems quicker to move out of the "buildings domain" to seek objects having the Burj's shape: *obelisk* (0.10), *missile* (0.09) and *projectile* (0.05). This possible predilection of VGG16 for shapes of elements, colour or "style" appears in other examples: Saturn is a *spotlight* (0.08) and a *ping pong ball* (0.07) rather than a *candle* (0.03), and the leftmost dragon in Figure 2 is labelled as *jersey* (0.68). Many of these mis-classifications are similar, in principle, to the operations underlying visually rooted metaphors. These metaphors give the listener or reader a clearer mind picture of a given element or scene through the parallel with something having similar visual properties, but pertaining to a different domain. To stick with the models' mistakes, it wouldn't be hard to imagine someone describing the Burj Khalifa as a "missile pointing to the sky" or the gigantic "obelisk of Dubai". Others of our models' mistakes, though, sound less natural to our sensibility: for example, describing Saturn as a candle in the sky or a ping pong ball in the sky is a less effective metaphor. ResNet50 captions an image of the setting Sun as a *ping pong ball* (0.58) and a *spotlight* (0.05) and only with lower confidence predicts similes used by humans to describe the sun in similar scenes, such as *orange* (0.017) and *balloon* (0.014). To give the reader a first hand idea of the descriptive qualities of these mis-categorisations, we present in Figure 4 a series of pictures with the first 5 categories assigned by the VGG16 model. Table 1 provides a small comparison of the mis-categorisations by VGG16 and ResNet50 on the same pictures.

The main intuition of this study is that some of these mis-categorisations seem to make a metaphoric sense for a human reader. For example, ResNet50 classifies a picture of lightnings as *spider web*. Although this may seem like an unexpected comparison, there are several pictures of lightnings on the Internet that are described by human annotators as *spider web lightning* due to their thread-like and branching shape. On the other hand, if a model labels the picture of a man sitting by a fire as a *volcano*, the metaphoric power of the classification becomes doubtful (although not necessarily absent).

To explore the parallels and differences between human visually rooted metaphors and visual models' mis-categorisations, we collected through manual online search a dataset of 100 pictures that human users had described with a metaphor or simile as shown in Figure 5. We exclusively selected metaphors that had *only* their target in the 1,000 ImageNet categories present in our models' ontology: for example, images of lightning (source, absent from the ontology) described as spider webs (target, present in the ontology), or fireworks (source, absent from the ontology) described as sea urchins (target, present in the ontology). With such a dataset it is possible to see, to a limited extent, whether the mis-categorisations performed by the models confronted with unforeseen elements go in the direction of the metaphors and similes humans conceive to provide a vivid description of an object.

If a human captioner described the image of a *sponge* (absent from the models' ontology) as a *harp* (present in the models' ontology), and our models categorise the same image as a *harp*, there is an overlap between the two frames. If, as in this case, our models categorise the same image as something else, e.g. a *barn spider*, there is a difference between the two frames. Table 2 shows the performance of four pre-trained models on metaphorical mis-categorisation. If we consider the first retrieved category for each picture, most models achieve an F-score between 0.23 and 0.26, with the exception of InceptionResNet that achieves an F-score of 0.0. If we relax the boundaries and take into consideration the first 5 results for each picture, most models' performance ranges between 0.3 and 0.4, and if we include the first 20 results they reach F-scores higher than 0.5. Considering the complexity of the task, we see F-scores of 0.3 and higher for the first 5 answers as an interesting result. At the same time, it is clear that this experimental frame remains limited: we were only able to use specific kinds of metaphors to work on the models' restricted ontology, and the metaphors had to be of

(a) Bubble, Alp, fireboat, mountain tent, fountain.

(b) Golf ball, gong, tick, chambered nautilus, spotlight.

(c) Matchstick, hook, spotlight, safety pin, flatworm.

(d) Comic book, jigsaw puzzle, book jacket, theater curtain, shower curtain.

(e) Shower curtain, pajama, ballpoint, maraca, rubber eraser.

(f) Volcano, fountain, seashore, space shuttle, lakeside.

Figure 4: The first 5 output categories of VGG16 for various pictures. Most of these pictures represent objects the network was not trained on. The reader can notice that some of the mis-classifications could work as metaphoric descriptions (as in b or f) more than others (as in a or e).

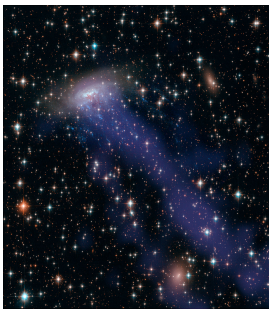


Figure 5: Two elements from our dataset. A galaxy described in the human-generated caption as a jellyfish, and a building described (and commonly known) as a cucumber/gherkin.

the single-word-to-single-word kind: the compositionality and flexibility present in many visually rooted metaphors, such as "the lawn is a *green* carpet" or "the snowflakes were *falling* dancers", are out of the scope of this kind of test. In the rest of the paper, we explore a different frame that allows more flexibility in the study of visually grounded metaphors.

5 Visual Spaces

The category of *cigarette* is absent from the ontology of the models we use in this study. If presented with a cigarette, most models see a *ruler* or a *band aid*. The classification step happens in the last layer of the networks: before that final layer, each model transforms its input picture in a 1x224x244x3 tensor that encodes the relevant visual features of the picture as a 4-dimensional set of weights. From such tensor the model draws a 1x1000 vector that represents the probability of such tensor to fall in each one of the 1000 categories.

This characteristic, shared by most neural classifiers, opens the possibility of exploring visually rooted metaphors as operations in a continuum, by using the final tensor extracted by each picture as a vector in a multi-dimensional space. Returning to the cigarette example, our VGG16 model cannot recognise any of the objects or symbols present in Figure 6.

The last picture confuses our model which miscategories it in a different way than the previous two images: the model fails to pick the similarities evident to a human eye. But if we flatten the pre-categorisation final tensors created by the model to represent these three images and compute their cosine similarity, we might be able to overcome the rigidity of the classification step. Here are the results of such trial: the cosine similarity between the cigarette (a) and the danger sign (b) is 0.5, as the two pictures are different. But the similarity between (b) and the cigarettes are dangerous sign (c) is 0.68, while the similarity between (a) and (c) is as high as 0.83. In other words, the visual similarity that the multi-class classification frame kept latent clearly emerges.

Object	VGG16	ResNet50
Burj Khalifa	Mosque, obelisk, missile	Mosque, palace, bell cote
Mountain	Alp, valley, mountain tent	Alp, valley, mountain tent
Galaxy	Jellyfish, fountain, window screen	Volcano, ski mask, jellyfish
Mushrooms	Water tower, lampshade, table lamp	Mushroom, hen of the woods, fountain
Blanket clouds	Seashore, fountain, sandbar	Wing, seashore, sandbar
Sun(drawing)	Ping pong ball, envelope, maraca	Wall clock, analog clock, web site
Ballerinas	Spiny lobster, hoopskirt, fountain	Fountain, king crab, pole
Belt	Buckle, muzzle, hair slide	Buckle, muzzle, hair slide
Cloud looking like a bird	Geyser, lakeside, valley	Valley, lakeside, worm fence

Table 1: Comparing mis-categorisations between VGG16 and ResNet50. The first column names the object presented in the picture, the remaining two columns present the first 3 captions offered by the two models.

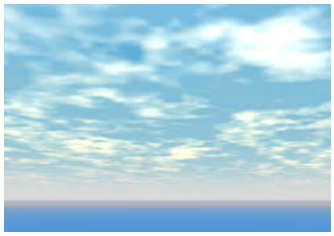
(a) ruler, band aid (b) hatchet, electric guitar (c) reel, croquet ball

Figure 6: Three pictures representing elements absent from our VGG16 ontology. The last picture shares obvious similarities with the first two images, but VGG16 classifies each picture in a completely different way.

Model	Top 1	Top 5	Top 20
VGG16	0.25	0.39	**0.57**
VGG19	0.23	0.33	0.54
InceptionResNet	0.00	0.01	0.05
ResNet50	**0.26**	0.39	0.53

Table 2: F1 scores for human-like metaphorical classification of 4 models considering the first 1, 5 and 20 results of each.

To strengthen this frame, we create visual vectors that represent several images of the same concept in order to create "conceptual" clusters or, in other words, new "classes" for our experiment without the need of a full new training set. For example, if we sum the flattened output tensors for two danger signs we obtain a new vector that "represents" both danger signs' relevant features. This approach seems to return better results. If we compute the cosine similarity between three different danger signs' tensors, we obtain an average similarity of 0.64. But if we sum two danger signs' tensors and compute the cosine similarity of the resulting tensor with the left-out picture's tensor, the average similarity rises to 0.73. In other words, by summing two danger signs' tensors we created a visually meaningful centroid in the feature space that represents better than any single image the essential appearance of a generic *danger sign*. It is possible to imagine that adding more pictures would make a more consistent representation. However, the most interesting aspect of this approach for our study is the possibility of obtaining a reasonable effect without the need of collecting large datasets or training the model from scratch. The same effect happens with the cigarette pictures: if the cosine similarity between two simple pictures of cigarettes as in Figure 6(a) is 0.95, the cosine similarity of a single cigarette vector with the summed vector of two other images of cigarettes rises to 0.99. Now we can compute the similarity between cigarettes, danger signs and cigarettes are dangerous disclaimers



(a) Sky

(b) Fire

(c) Sunset

Figure 7: Samples from the three compound visual vectors we create to serve as source, target and modifier of the metaphor *Sunset* is *Sky* on *Fire*.

of the kind presented in Figure 6 as a cosine similarity between vectors. Through this simple operation the similarity between the danger signs' vector and the disclaimers' vector rises up to 0.81, and the similarity between the cigarettes' vector and the disclaimers' vector rises to 0.87. If we sum the cigarettes' vector and the danger signs' vector and compare the result with the anti-smoke disclaimers' vector, the cosine similarity goes to 0.92. This high similarity does not seem to be the effect of noise: the similarity of the anti-smoke disclaimers' vector and a vector of an unrelated picture, such as an image of a firework, is -0.8.The essential visual similarities that make the symbolic disclaimer in Figure 6(c) understandable for humans are retrievable from the visual feature space. To prove the concept, we computed the cosine similarity between the danger signs' vector and the individual vectors of all the pictures present in the dataset described in this study. Despite more than 100 confounders, the three pictures of danger signs came with the highest ranking, followed by the two anti-smoke disclaimers. We then repeated the operation with the (danger+cigarette) vector, that retrieved all the cigarettes as most similar pictures. This leads to our final investigation.

6 Adding Fire to the Sky: Compositionality in Visually Grounded Metaphors

A visually grounded metaphor sometimes used to describe an impressive sunset is *the sky is on fire*. This is a simple and effective grounded metaphor: the reader or listener "adds" the colours and intensity of fire to the sky in order to imagine a vivid sunset. If this metaphor is indeed rooted in visual data and visual data only, this is the operation we should be able to perform in the visual space to "create" a sunset vector. Through online manual search we collected 8 pictures of sunsets described as *Sky on fire* by their captioners. The individual vectors of some of these pictures already present the similarities necessary for the metaphoric shift: out of 8 pictures described as *Sky on fire* 2 retrieved as most similar picture in our dataset a picture of a fire with an average cosine similarity of 0.6, and another 4 had pictures of fire among the first ten most similar elements, with an average cosine similarity of 0.5.

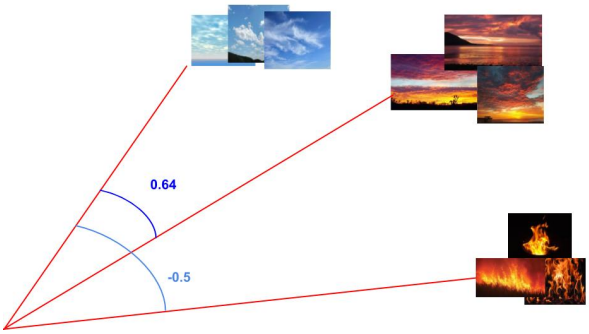


Figure 8: A schematic visualisation of the cosine similarities between the *sky*, the *fire* and the *sky on fire* vectors. The *sky* vector is relatively similar to the *sky on fire* vector and further away from the *fire* vector.

But is it possible to reproduce the compositionality of this metaphor in the visual space? To answer this question, we created a *sky* vector out of 10 pictures of (mainly blue) skies such as the one in Figure 7(a). The average cosine similarity of these pictures is around 0 (the lowest possible cosine similarity is -1). This relatively low similarity between sky pictures is probably due to the lack of stable and recognisable shapes in the "concept of sky": most of our sky pictures featured a of varying hue background sometimes with some clouds, and the clouds' shapes varied constantly. We then created a *fire* vector out of 13 pictures of fire such as the one in Figure 7(b) with average cosine similarity

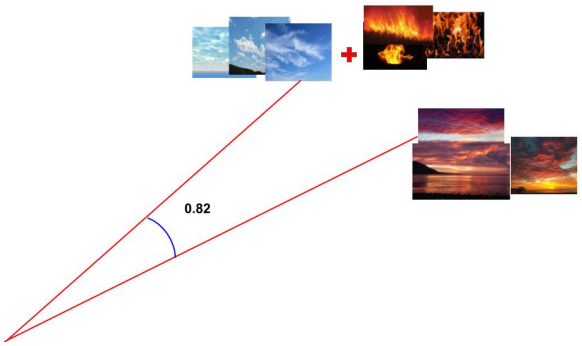


Figure 9: A visually grounded metaphor in the visual space. If we sum the *sky* vector with the *fire* vector, we create a "metaphoric" new vector which is closer to the *sky on fire* vector than the simple *sky* vector was: adding fire to the sky seems an effective way to recreate a sunset.

of 0.6. Finally, we created a *sunset* vector out of 7 pictures of sunsets captioned by humans as *sky on fire*, as the one in Figure 7(c). The average cosine similarity between the pictures of this group is 0.74. As represented in Figure 8, the cosine similarity between the *sky* vector and the *fire* vector is low: -0.5. The objects, sky and fire, have little in common in terms of visual features. The cosine similarity between the *sky* and *sky on fire* vectors that represent the same object under different conditions is higher: 0.64. If we sum the vectors of *sky* and *fire*, we create a *sky-fire* vector as shown in Figure 9. The cosine similarity of this new vector with the *sky on fire* vector rises to 0.82. In other words, adding the metaphoric *fire* to the literal *sky* made the sky vector closer to the sunset vector. The compositional *sky on fire* metaphor seems to work in our visual space (see Figure 9 for a visualisation). Although this is a particularly clear case of visual compositionality, metaphoric compositionality of this kind seems to be present also in other examples. In Table 3 we give an overview of some of the metaphors we tried. We tried to select on metaphors that could be strongly visual, and that would not rely on excessively complex shapes or hues.

Collecting pictures that represent such metaphors from online data is a difficult task. The metaphors have to be compositional, visually grounded and included in captions. We thus collected a tiny dataset of 22 such metaphors (some of which we include in Table 3) with an average of 5 images associated with each element: source, target and modifier. To make negative examples

Metaphor	**Sim ST**	**Sim (S+M)T**
Sunset is sky on fire	0.64	0.82
Blonde hair are river of gold	0.01	0.1
Snow is white carpet	0.82	0.90
Lawn is green carpet	-0.3	0.4
Hair are white waterfall	-0.5	0.62

Table 3: Compositional metaphors: the similarity between two visual vectors representing the (S)ource and the (T)arget of a metaphor increases if a modifier's vector is added to the source - (S+M)T. For example, the cosine similarity of the *blonde hair* vector and *river* in the second row is 0.01. If we sum *river* with a vector representing its modifier *golden* (which is a vector created out of several pictures of gold and golden elements) the similarity goes up to 0.1.

we created a "balancing" list of 22 false metaphors by randomly shuffling targets and modifiers. To keep the experiment clear in its scope, we tried to avoid for this negative counterpart borderline compositions that could work as unusual but still valid or evocative metaphors, since that would create a fascinating but hard to define grey area. For each of these metaphors, we operated the following steps:

1. We measured the cosine similarity of the source and target visual vectors: for example, the similarity of the Sky vector with the Sunset vector.

2. We either summed or multiplied the modifier's vector to the source's vector; for example, we added the Fire vector to the Sky vector.

3. We measured the cosine similarity of the new "modified source" vector with the target vector: the similarity of Sky+Fire with Sunset.

Every time the modifier increases the similarity between a true metaphor's source and target, we count it as a true positive. Every time the same happens for a false metaphor, we count it as a false positive . We show the performance of the InceptionResNet and ResNet50 models on this dataset in Table 4.[1] We find that most real metaphors improve through the addition of the modifier's vector, while most

[1]The other two models returned very similar results and therefore we do not discuss them specifically.

false metaphors do not, achieving the best F-score of 0.68 using ResNet50.

For both classifiers, the majority of the real metaphors improved (in terms of source-target cosine similarity) if we added the modifier with the source: in other words, in over 60 per cent of the cases we reproduced the mere visual compositionality of these metaphors. At the same time, the majority of the false metaphors got worse (in terms of souce-target cosine similarity) if we added the modifier to the source, confirming that the effect is linked to the visual compositionality of such expressions.

Multiply	**Precision**	**Recall**	**F1**
InceptionResNet	0.55	0.68	0.61
ResNet50	0.63	0.31	0.42
Sum	**Precision**	**Recall**	**F1**
InceptionResNet	0.60	0.64	0.62
ResNet50	0.64	0.73	**0.68**

Table 4: Precision, Recall and F1 for two models' metaphorical compositionality. The F-score measures to what extent the modifier improves the similarity between source and target in real metaphors (precision), but not in false metaphors (recall). We test the composition of visual vectors using multiplication (top) or sum (bottom).

7 Conclusions and Future Works

We conceived this study as an exploration of visually grounded metaphors in two experiments and two tiny datasets. In the first experiment we focus on categorisation: two image captioning models classified pictures of previously unseen elements and compared them with human-generated metaphors for the same pictures, returning a low overlap between human generated metaphors and models' mis-classifications. It is important to keep in mind that metaphors are flexible and diverse, and some of the mis-classifications of the models might be valid metaphors for humans - they were just absent from the specific dataset we collected. In this respect, an overlap of 30% between human metaphors and the first 5 captions produced by the models is encouraging. In a number of cases, the mis-classifications of the models do not seem to align with anything similar to a human-like metaphor, especially if variables like background or peripheral elements come into play. This doesn't mean that the so-called visually rooted metaphors are not visually rooted: but they might rely on either more complex similarity that were not captured by our models, or on a composition of visual and extra-visual world knowledge. In the second experiment we focus on unsupervised composition using a multi-dimensional visual feature space that offers a flexible representation for our domain. Using an unsupervised approach we cannot produce a comparison against a labelled dataset as in the classification experiment. However, we do show that the apparent visual compositionality in several metaphors can be predicted in the visual feature space. In most cases adding the metaphoric modifier to the metaphor's source made the source and the target closer than they were before. This shift indicates that the metaphors are grounded in the visual features that are encoded by image classification models. The visual elements present in those metaphors are working in the visual space and account for the effectiveness and flexibility of the metaphoric expressions.

There are several ways to continue from this study. First of all, our datasets are very small. Collecting larger corpora of visually grounded metaphors and relative pictures would be necessary to expand our study. This would also imply finding more systematic ways of selecting both the metaphors and their pictures, since for these studies we used our sole sensibility to select and collect the examples. It would also be interesting to add more complex cases, especially when operating on the continuum visual space, and to compare the compositional efficiency of different metaphors. While such selection and collection steps are challenging due to the complex nature of the problem we study, the ever-growing wealth of annotated pictures makes us optimist about its feasability. An open question remains to what degree different metaphors are grounded visually and to what degree they can be predicted from the language models. It would also be interesting to check which visual clues are most useful to the classifiers when they reproduce human metaphorical combinations. Finally, even if no human ever produced a specific metaphor, this doesn't mean that such metaphor is bad: it would be interesting in the future to measure the level of human appreciation of visually grounded metaphors generated through image captioning. These will be the foci of our future work.

References

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55.

François Chollet et al. 2015. Keras.

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

Rutvik H Desai, Lisa L Conant, Jeffrey R Binder, Haeil Park, and Mark S Seidenberg. 2013. A piece of the action: modulation of sensory-motor regions by action idioms and metaphors. *NeuroImage*, 83:862–869.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33.

Peter Gorniak and Deb Roy. 2004. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

David Kemmerer. 2015. Are the motor features of verb meanings represented in the precentral motor cortices? yes, but within the context of a flexible, multilevel architecture for conceptual knowledge. *Psychonomic Bulletin & Review*, 22(4):1068–1075.

Simon Lacey, Randall Stilla, and Krish Sathian. 2012. Metaphorically feeling: comprehending textural metaphors activates somatosensory cortex. *Brain and language*, 120(3):416–421.

Vicky Tzuyin Lai and Tim Curran. 2013. Erp evidence for conceptual mappings and comparison processes during the comprehension of conventional and novel metaphors. *Brain and Language*, 127(3):484–496.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Deb Roy. 2005. Grounding words in perception and action: computational insights. *Trends in cognitive sciences*, 9(8):389–396.

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1-2):1–12.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Jeffrey Mark Siskind. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of artificial intelligence research*, 15:31–90.

Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261.

Josiah Wang, Katja Markert, and Mark Everingham. 2009. Learning models for object recognition from natural language descriptions.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Kiki Zanolie, Saskia van Dantzig, Inge Boot, Jasper Wijnen, Thomas W Schubert, Steffen R Giessner, and Diane Pecher. 2012. Mighty metaphors: Behavioral and erp evidence that power shifts attention on a vertical dimension. *Brain and cognition*, 78(1):50–58.

Li Zhang and John Barnden. 2013. Towards a semantic-based approach for affect and metaphor detection. *International Journal of Distance Education Technologies (IJDET)*, 11(2):48–65.

Junpei Zhong, Martin Peniak, Jun Tani, Tetsuya Ogata, and Angelo Cangelosi. 2019. Sensorimotor input as a language generalisation tool: A neurorobotics model for generation and generalisation of noun-verb combinations with sensorimotor inputs. *Autonomous Robots*, 43(5):1271–1290.