Final Project. Image categorization systems as models of visually grounded metaphors

yuri.bizzoni

October 2018

1 Introduction

In this project I will study the differences and similarities between neural image classifiers' mis-categorizations human visually grounded metaphors - that could be conceived as intentional mis-categorizations.

A visually grounded metaphor is often used to describe a scene or an object in a particularly vivid way. These metaphors can come in several shapes: a large person could be described as an elephant; somebody's blue eyes can be described as clear as a summer sky; and so on. 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 more effective mind picture.

At the same time, the mechanisms and workings underlying metaphors are not completely understood yet. Visually grounded metaphors are only one of 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 how some other synesthetic expressions like *a blue music* work in our brain is not clear yet.

But in the case of visually grounded metaphors, unlike metaphors that draw from several senses, it might be possible to start modelling them using the visual NLP models used for image captioning.

So the question is if, and to what extent, modern image captioning models can be used to replicate the workings of visually grounded metaphors and similes. Since this is a short pilot project, I will sketch some possible lines of exploration together with two small scale experiments: more data and analyses would be necessary to deepen the topic.

While the problems and advantages of visually grounding language in computational models are widely discussed in a large bibliography (Siskind; 2001; Gorniak and Roy; 2004; Roy; 2005; Roy and Reiter; 2005; Barsalou; 2008), and the mechanisms underlying metaphor processing have been amply explored (Lakoff and Johnson; 2008), the research on the topic of metaphor and grounded computational models is still scarce (Shutova et al.; 2016).

Neurolinguistics has shown that the the literal properties of terms are still activated when those terms are used metaphorically: for example, senso-motory verbs activate the senso-motory memory in the brain even when those verbs are used metaphorically, while the same does not happen with idioms: the idea that metaphors create a compositional feature transfer between the source and the target seems to be confirmed (Desai et al.; 2013). Other studies have also shown that some abstract concept, like power, seem to be linked in the brain to some concrete spacial dimension, like the vertical up-down dimension (Zanolie et al.; 2012), which would account for the spacial metaphors of power, very often visualized on a vertical axis. In short, many studies in neurolinguistics are proving that metaphors are indeed sensory-based (Lacey et al.; 2012). Reproducing these mechanisms in computational models is the main idea behind this project.

For this project I will use two of keras' (Chollet et al.; 2015) pre-trained models. These models are ResNet50 (He et al.; 2016) and VGG16 (Simonyan and Zisserman; 2014).

As with other keras implementations, these models come with its pre-trained set of weights and categories. They are a classifiers, not tools for articulated image description: the captions generated by these models are always single tokens.

In the pre-trained version of these models, the networks have been trained on 1000 categories from ImageNet (Deng et al.; 2009).

Both models operate through three obvious, macro-level phases. First, the model takes an image as input. Second, it transforms that image into a vector of weights that should represent that image's most relevant features. Finally, it operates a prediction or classification over such vector. Since the categories are relatively few and the models are not perfect, mis-categorizations occur.

2 Miscategorizations

When presented with an image, a classifier will output a list of possible captions for that image with their scores (that indicate the similarity between the category and the image). So for example when presented with a picture of a bird, ResNet50 will output a list like the following:

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

and so on.

When presented with an airliner, the model will output captions like *airliner* (0.93), *warplane* (0.03), *airship* (0.001) etc.

When it's presented with a clear instance of an element it has been trained upon, the classifier usually produces a very high-probability prediction, followed by categories that progressively share less features with the observed image.

For example, when presented with a picture of an Indian elephant, ResNet50 outputs the following list of captions:

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

Even if the gap in probability between the first prediction and the second is very large, and the probability of the predictions from the 4th on are very small, we can see that the model is not predicting randomly: the output's classes share some features with the Indian elephant, in a decreasing order of overlap. First other kinds of elephants are suggested; then, other animals that, in ResNet50's ontology, share important properties with the Indian elephant. It is possible to observe how the suggested alternative species have in common the characteristic of being large, massive animals, four-legged and often with prominent tusks.

What the network is doing here is in some way similar to what a human would do to describe an specific animal to someone who has never seen it: it is looking for other animals sharing some similarities with the one it has to describe.

This mechanism is even more evident, of course, when the network is presented with a picture of "something" it has never seen before, or that it has not been trained to categorize properly.

Let's take for example Figure 1, an image of a fire in a non specified environment. Keras' pretrained version of ResNet50 does not have "fire" among its categories.

So when presented with this picture, the model returns the following list of possibilities:

1. stove 0.85
2. fire screen 0.14
3. dutch oven 0.002

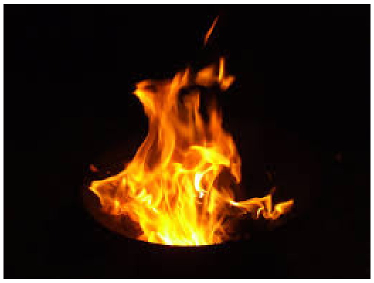
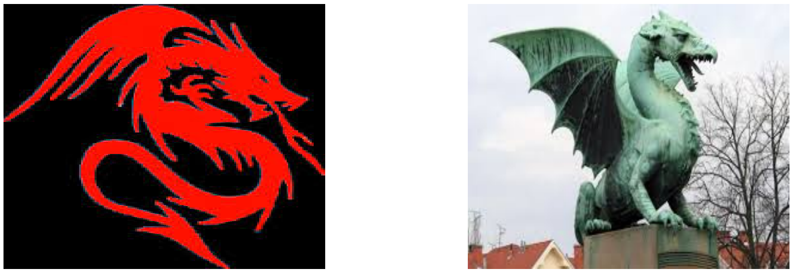


Figure 1: Fire!

4. frying pan 0.0007

It seems evident that the network is looking for categories in which fire is probably a component. Not being able to figure out what it is looking at with a 0.9+ confidence, it returns captions drawn from categories of pictures that share some properties with the presented image.

Sometimes, the background or style of the picture also plays a role. For example, Figure 2.a, a dragon, is categorized as *comic book* (0.28) or *laptop* (0.08), while Figure 2.b, still representing a dragon, is labelled as *pedestal* (0.61), *fountain* (0.38) or *palace* (0.0005). In both cases the classifier has focused on the "style" of the object - a drawing in the first place, a statue in the second one - to attempt a low-probability classification.



(a) a dragon

(b) another dragon

Figure 2: Two dragons.

Mis-categorizations can also happen when the object is a depicted in a way that confuses the network. For example, ResNet50 has various *bridge* categories (such as *steel arch bridge* and so on) in its ontology. But when presented with the picture of a small bridge mirrored in a lake it classifies it as *viaduct* (0.46) or *lakeside* (0.21), clearly confused by the two elements present in the image, and when presented with an aerial picture of a large modern bridge crossing the sea it labels it as *bannister* (0.44) or *dam* (0.03).

Once again, the model is looking into objects that pertain somehow to the same field or conceptual area of the difficult image: for the bridges, it looks for dams, viaducts and so on. In the same way, when presented with a picture of a church, the model suggests *church* (0.93), *bell cote* (0.02), *monastery* (0.02). Also, the model tries to describe the Burj Khalifa, which it has never seen before, 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 me since it moves out of the conceptual domain of the picture to look for similarities in different areas.

But what happens if the model is present with an image belonging to a conceptual domain completely unknown? For example, this particular implementation of ResNet50 has not been trained to anything pertaining to the sky: clouds, planets and stars are absent from its ontology.

When presented with a classical image of Saturn, the model predicts *candle* (0.81). The reason of this unexpected elaboration lies, I suspect, in the color of Saturn's atmosphere, that is somewhat similar to a candle's wax color. [1]

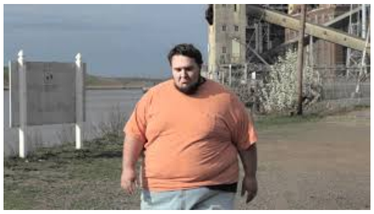


Figure 3: A guy.

In other situations, the network focuses instead on background elements that it can recognize. For example, this model is also not trained on people: it cannot label a person as *person.* So when presented with the image of a person in a bathtub, it captions it as *bathtub* (0.08), *bath towel* (0.07), *tub* (0.06). For similar reasons, and showing some non politically correct bias, the person in Figure 3 is described as *prison* (0.05), *jean* (0.02) and *barrow* (0.06). Again, in a picture representing a breaking storm over the sea, ResNet50 - not "knowing" what a storm is - pics the peripheral, scenic elements: *breakwater*, *pier* and so on. These mistakes have been duly noted by the bibliography (Wang et al.; 2009; Xu et al.; 2015).

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.b, *viaduct* (0.24), after *triumphal arch* (0.33), for the bridge mirrored in the water. Also for known objects, its second and third best guests often align with the ones produced by ResNet50: for example, the *church* prediction for the church is again followed by *monastery* (0.02) and *bell cote* (0.02).

In the case of the Burj Khalifa, the first prediction is again *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).

[1] Following this mis-categorization I actually found online a kind of wax called "Saturn yellow".

This possible predilection of VGG16 for shapes of elements like color or "style" appears in other examples: Saturn is categorized first as a *spotlight* (0.08) and a *ping pong ball* (0.07) rather than a *candle* (0.03), and the 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 are meant to 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, would sound less natural to the human sensibility: for example, describing Saturn as a candle in the sky, or a ping pong ball in the sky, might be a less effective metaphor.

The same model also captions an image of the setting Sun as a *ping pong ball* (0.58) and a *spotlight* (0.05) and only later it recurs to smiles more 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-categorizations, I present in Figure 4 a series of pictures with the first 5 labels assigned to them by the VGG16 model.

In Table 1 I provide a small comparison of the miscategorizations operated by VGG16 and ResNet50 on the same pictures.

The examples could continue for a long time. The main starting point, or intuition, of my study is that sometimes these mis-categorizations seem to make a metaphoric sense for a human reader, and sometimes they don't. For example, a picture of lightnings is captioned by ResNet50 as *spider web*. Despite this could seem like an unexpected comparison, there are several pictures of lightinings on the Internet that have been captioned, by their human annotators, as *spider web lightning* due to their thread-like and branching shape. On the other hand, when a picture of a man sitting by a fire is labelled as *volcano*, the metaphoric power of the caption becomes more doubtful [2].

It's important to remember that we are talking here of completely visual similarities. In this sense, the categories produced by the models are not necessarily equal to those that can be retrieved in a classical semantic space. In Table 2 I show some examples of this difference.

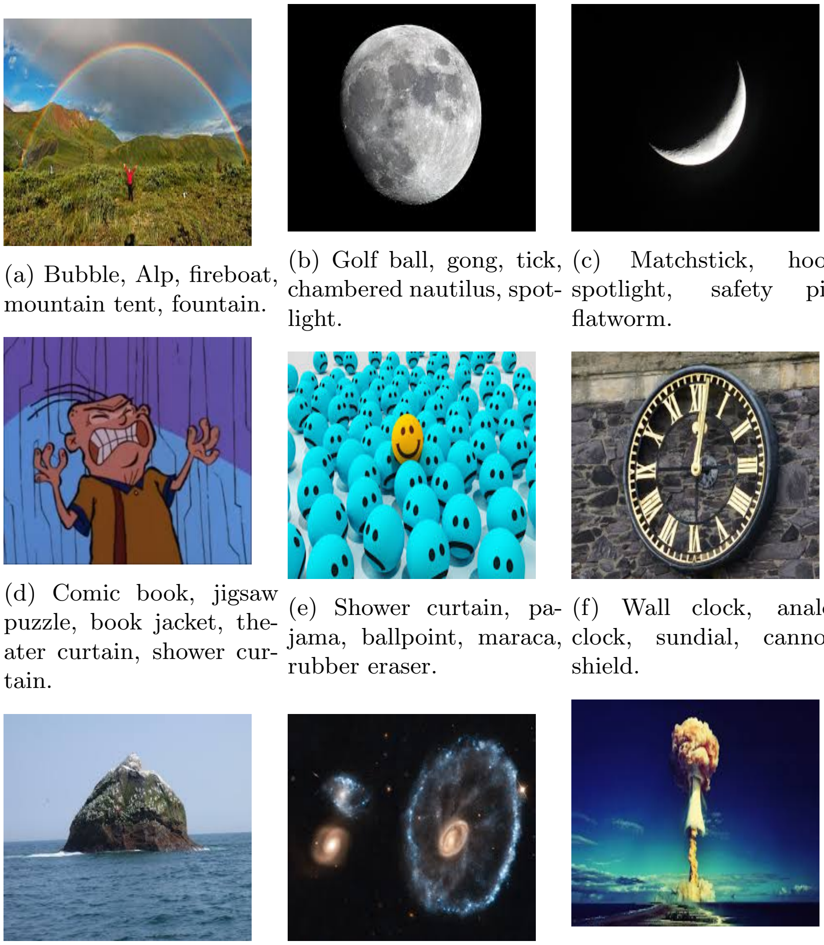
A simple way of exploring the parallels and differences between visually rooted metaphors and models' mis-categorizations is to collect a small corpus of visually rooted metaphors present in pictures' captions.

That's what I did: I collected, through basic online search, 100 pictures that were described by the users with some kind of metaphor. I exclusively selected metaphors that had *only* the target in the 1000 ImageNet categories present in my models' ontology: for example, images of lightning (source, not present in the ontology) described as spider web (target, present in the ontology) or

[2] Although not necessarily absent

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
Atomic mushroom	Volcano, knot, fountain	Volcano, cauliflower, geyser
Submarine	Space shuttle, scuba diver, hammerhead	Submarine, scuba diver, tiger shark

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



(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) Wall clock, analog clock, sundial, cannon, shield.

(g) Promontory, seashore, cliff, wreck, pirate.

(h) Nematode, jellyfish, tick, frying pan, fig.

(i) 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 has not been trained upon. The reader can notice how some of the mis-classifications could work as metaphoric descriptions (as in b,h,i), while others could not (as in a,d,e).

fireworks (source, not present in the ontology) described as sea urchins (target, present in the ontology).

In this way it is possible to see, to a small extent, whether the mis-categorizations performed by the models when confronted with unforeseen elements go in the direction of the metaphors and similes humans conceive when they want to provide a more vivid description of an object.

If the image of a *sponge* (not present in the models' ontology) described

Pair	**ResNet50**	**Word2Vec**
Tank - Cannon	0.05	0.33
Satan - Mask	0.83	0.09
Jupiter - Chambered Nautilus	0.53	0.30
Solar flare - Volcano	0.25	0.26
T-rex - Triceratops	0.08	0.65
Diplodocus - Triceratops	0.02	0.54

Table 2: Differences between distributional and visual similarities. In the first column I show an object as classified by ResNet50. For example, an image of a tank was classified by ResNet50 as a *cannon.* I then show the similarity between the two objects as computed by ResNet50 (second column) and the distributional similairity between the two words (*tank* and *cannon*) as found in a Word2Vec implementation. When the semantic similarity of two concepts does not entail any kind of visual similarity, as in the case of *triceratops* and *T-rex* or *diplodocus*, the differences in similarity appear quite clearly.

by the human captioner as a *harp* (present in the models' ontology) is also categorized by the models as *harp*, there is an overlap between the two frames.

If instead, as in this case, the same image is categorized by the models as something else (in this example, *barn spider*), there is a difference between the two frames.

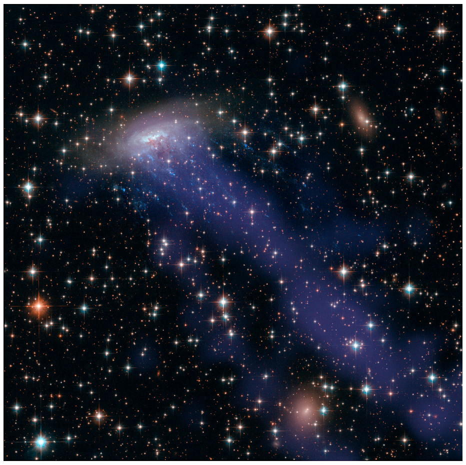


Figure 5: An element from my dataset. A galaxy described in the human-generated caption as a jellyfish.

If I take only the first retrieved category for each pictures, both models achieve an overlap of 0.15 (15 matches over 100) with human metaphoric de-

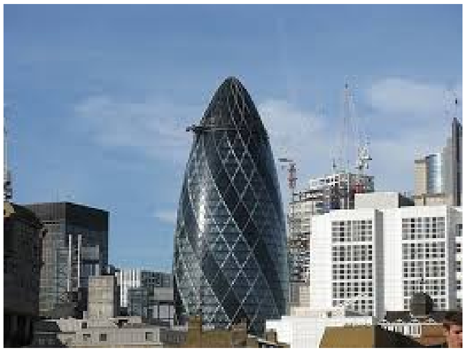


Figure 6: An element from my dataset. A building described in the human-generated caption as a cucumber. It is also commonly known as *the cucumber.*

scriptions.

If I relax the boundaries and take into consideration the first 5 results for each picture, ResNet50 reaches an overlap of 0.32, while VGG16 an overlap of 0.3.

Considered the complexity of the task, a 30 percent overlap within the first 5 answers can be seen as already an interesting result.

At the same time, it is clear that this frame of experiment is still limited: I only could use specific kinds of metaphors to work on the models' restricted ontology and the metaphors had to be of 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 set.

In the rest of the project I will explore a different frame that can allow way more flexibility in the study of visually rooted metaphors.

3 Visual spaces

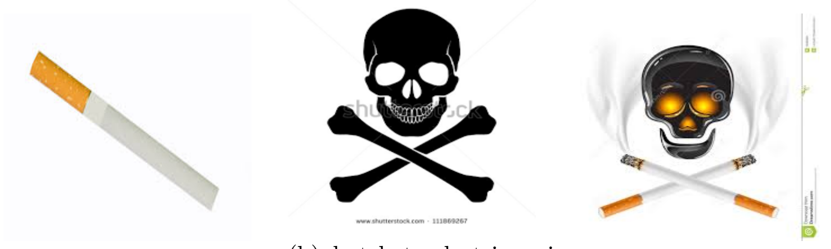
My implementation of VGG16 was not trained on the *cigarette* category. When presented with a cigarette, the model sees it as a *ruler* or a *band aid*.

Anyway, the classification step happens only in the last layer of the model: before that, the model transforms any picture in a 1x224x244x3 tensor that contains, or should contain, the relevant visual features of the picture encoded as a 4-dimensional set of weights.

From such tensor the model then draws a 1x1000 vector that represents the probability of such tensor to fall in each one of the 1000 categories.

This might open the possibility of exploring visually rooted metaphors in a less supervised way, by directly using the final tensor extracted by each picture in a multi-dimensional space.

To come back to the cigarette example, my VGG16 model cannot recognize any of the objects or symbols present in Figure 7.



(a) rule, band aid

(b) hatchet, electric guitar

(c) reel, croquet ball

Figure 7: Three pictures representing elements not present in my VGG16 ontology. The last picture clearly contains features of the first two, but the model still classifies them in completely different ways.

Apparently, the model is confused enough by the last picture to mis-categorize it in a different way than it did with the previous two: it does not see the similarities evident at a human eye.

But what happens if I flatten the pre-categorization final tensors of these three elements created by the model and compute their simple cosine similarity?

These are the result of the trial: the cosine similarity between the cigarette (a) and the danger sign (b) is 0.5, and the two pictures are actually quite different. But the similarity between (b) and the *cigarettes are dangerous* sign (c) rises up to 0.68, and the similarity between (a) and (c) goes up to 0.83.

In other terms, the visual similarity between (a) and (c), and (b) and (c), starts emerging, even if the multi-class classification frame did not make it evident.

To strengthen this frame, we could create visual vectors that represent several images of the same concept, in order to create "conceptual" clusters or, in other terms, new "classes" for our experiment without the need of a full new training set.

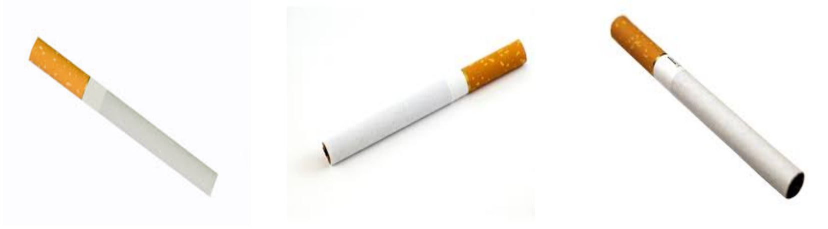
For example, if I sum the flattened output tensors for two danger signs, I obtain a new vector that could "represent" the danger sign relevant features better.

This approach seems to work better. If I compute the cosine similarity of three different danger signs, I have an average similarity of 0.64. But if I sum two of them and then compute the cosine similarity of the resulting vector with the left out picture, the average similarity rises to 0.73. In other terms, I created a visually meaningful centroid in the space, that represents better than a single image how a *danger sign* looks like. It's possible to imagine that adding more pictures would make it even more consistent, but what interests me here 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 cigarette pictures: even if the cosine similarity between two individual pictures of cigarettes like the one in Figure 7.a is already up to 0.95, the cosine similarity of a single cigarette vector with the summed up vector of 2 images of cigarettes rises to 0.99.

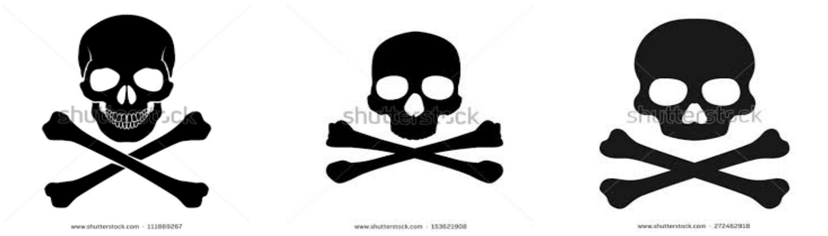
Now, I can re-compute the similarity between cigarettes, danger signs and *cigarettes are dangerous* disclaimers of the kind presented in Figure 7 as a cosine similarity between summed vectors. The effect is evident: the similarity between the danger signs vector and the disclaimers vector rises up to 0.81, and the similarity between the cigarettes vector and the disclaimers vector is now 0.87.

Finally, if I sum the cigarettes vector and the danger signs vector together and compare it with the anti-smoke disclaimer vector, the cosine similarity goes to 0.92, while the similarity of the anti-smoke disclaimer vector with a completely unrelated picture, such as an image of a firework, is -0.8. The essential visual similarities that make the symbolic disclaimer in Figure 7.c understandable for humans are now clearly retrievable in the visual space.



(a) a cigarette (b) another cigarette (c) another cigarette

Figure 8: The cigarette vector is the sum of the individual vectors of these three images.



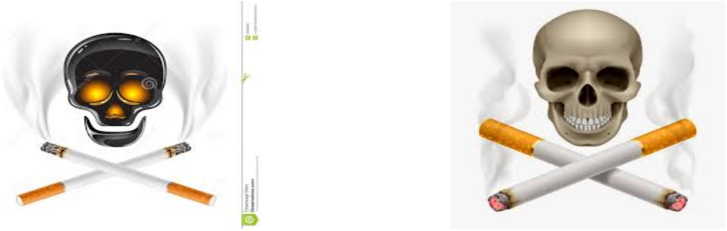
(a) a danger sign (b) another danger sign (c) another danger sign

Figure 9: The danger vector is the sum of the individual vectors of these three images.

To prove the concept further, I computed the cosine similarity between the danger signs vector and the individual vectors of all the pictures present in the small variegated dataset I collected for this project (circa 300 pictures).

The three pictures of danger signs where the highest ranking, immediately followed by the two anti-smoke disclaimers, while the disclaimers' vector retrieved all the cigarettes as very similar pictures.

This frame could lead to a final investigation.



(a) an anti-smoke disclaimer

(b) another anti-smoke disclaimer

Figure 10: The disclaimer vector is the sum of the individual vectors of these two images.

4 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*. The colors and intensity of fire are to be "added" to the sky by the reader or listener of the metaphor in order to imagine a vivid sunset. If this metaphor is completely rooted in visual data, this is precisely the operation we should be able to perform in the visual space to "create" a sunset vector.

Looking online it is possible to find several pictures of sunsets described as *sky on fire*.

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 element in my dataset of pictures, a picture of a fire, with an average cosine similarity of 0.6, and 4 more had fire pictures among the first ten most similar elements, with an average cosine similarity of 0.5.

But is it possible to reproduce the compositionality of this metaphor in the visual space?

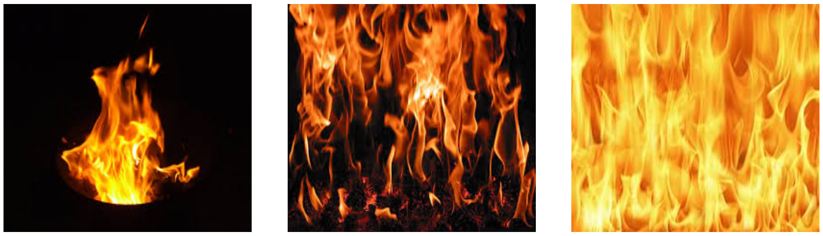
This is how one could proceed to verify this hypothesis.

First, I created a *sky* vector out of 10 pictures of (mainly blue) skies. The average cosine similarity of the three pictures is around 0 (the lowest possible cosine similarity is -1). I then created a *fire* vector out of 13 pictures of fire, with inner cosine similarity of 0.6. Finally I created a *sunset* vector out of 7 pictures of sunsets that were captioned by their users as *sky on fire*. The inner average cosine similarity of this group is 0.74.

The cosine similarity between the *sky* vector and the *fire* vector is expectedly very low: -0.5. The objects, sky and fire, have very little in common in terms of visual features.

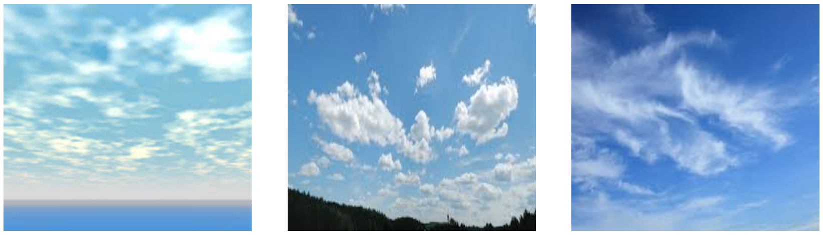
The cosine similarity between the *sky* and *sky on fire* vectors is obviously higher: 0.64. They represent the same object under different conditions.

Finally, if I sum the two vectors of *sky* and *fire*, I can create a *sky-fire* vector. The cosine similarity of this new vector with the *sky on fire* vector rises to 0.82.



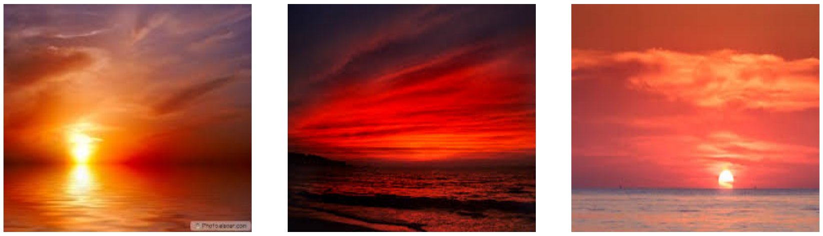
(a) a fire (b) another fire (c) another fire

Figure 11: Some of the components of the fire vector.



(a) the sky (b) another sky picture (c) another sky picture

Figure 12: Some of the components of the sky vector.



(a) the sky is on fire (b) another sky on fire (c) another sky on fire

Figure 13: Some of the components of the sky-on-fire vector.

In other terms, compositionally adding the metaphoric *fire* to the literal *sky* made the sky vector closer to the sunset vector: the *sky on fire* metaphor seems to work in our space.

Although this doesn't always work that well, the compositionality of similar metaphors seems to be present in various examples. In Table 3 I offer an overview of some of the metaphors I have tried.

Collecting similar metaphors from online data is not an easy task. They have to be compositional, visually grounded and present in captions.

I collected a small corpus of 22 such metaphors (some of which I present in Table 3), with an average of 5 images for each element - source, targed and modifier.

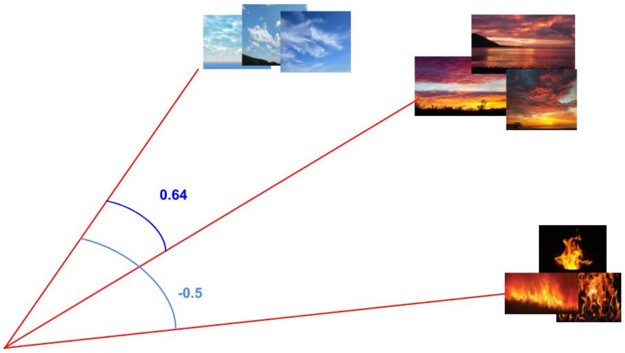


Figure 14: A schematic visualization of the cosine similarities between the *sky*, the *fire* and the *sky on fire* vectors. The *sky* vector is relatively similar to the *sky on fire* vector, and far from the *fire* vector.

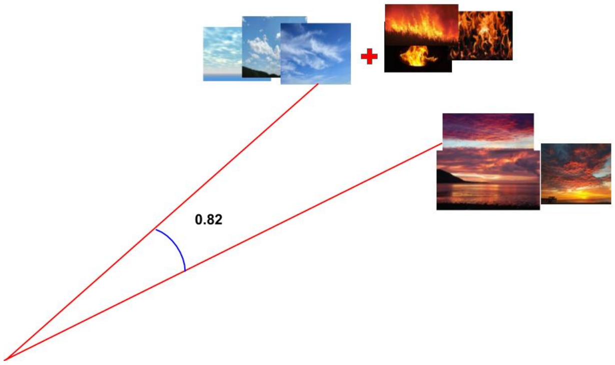


Figure 15: A visually grounded metaphor in the visual space. If I sum the *sky* vector with the *fire* vector, I create a "metaphoric" new vector which is closer to the *sky on fire* vector than the simple *sky* vector was: adding fire to the sky seems an effective way to recreate a sunset.

Out of 22 visually rooted metaphors of this kind, 16 "improved" (in terms of cosine similarity) when the modifier was added: in other words, in over 72 per cent of the cases the merely visual compositionality of these metaphors was reproduced in the space.

5 Conclusions

I conceived this study as an exploration in the little explored field of visually rooted metaphors. While my datasets are necessarily small, I think some elements already appear with clarity.

In the categorization frame, in which I asked two image captioning models to

Metaphor	**Similarity between source and target**	**Similarity between source and target+modifier**
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 way the similarity of two visual vectors increases when the modifier's vector is added. For example, the cosine similarity of the *blonde hair* vector and *river* in the second row is 0.01. If I 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.

categorize pictures of unforeseen elements, the overlap between human generated metaphors and models' mis-categorizations was quite low. At the same time, it is important to keep in mind that metaphor is a very flexible and diverse phenomenon and some of the mis-categorizations of the models might have been used by humans as valid metaphors - they were just not present in the specific dataset I collected. In this sense, an overlap of 30 per cent between human metaphors and the first 5 captions produced by the models could be seen as encouraging.

In a number of cases, nonetheless, the mis-categorizations of the models don't seem to align with anything similar to a human-like simile, especially when variables like the background or peripheral elements come into play.

This doesn't mean that the so-called visually rooted metaphors are not visually rooted, though.

The visual space I propose offers a more flexible frame to work on the topic. Being a completely unsupervised frame I couldn't reproduce a classification approach like the one I used for the first experiment, but I show that the compositionality that seems to operate in several metaphors can be apparently reproduced in the space: in 72 per cent of the cases, adding the metaphoric modifier to the metaphor's target actually made source and target closer in the visual space. This shift indicates that the visual elements present in those metaphors is actually working in the visual space itself and accounting for the effectiveness and flexibility of the human generated metaphoric expressions.

References

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

Chollet, F. et al. (2015). Keras.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database, *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, Ieee, pp. 248–255.

Desai, R. H., Conant, L. L., Binder, J. R., Park, H. and Seidenberg, M. S. (2013). A piece of the action: modulation of sensory-motor regions by action idioms and metaphors, *NeuroImage* **83**: 862–869.

Gorniak, P. and Roy, D. (2004). Grounded semantic composition for visual scenes, *Journal of Artificial Intelligence Research* **21**: 429–470.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

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

Lakoff, G. and Johnson, M. (2008). *Metaphors we live by*, University of Chicago press.

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

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

Shutova, E., Kiela, D. and Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 160–170.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* .

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

Wang, J., Markert, K. and Everingham, M. (2009). Learning models for object recognition from natural language descriptions.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention, *International conference on machine learning*, pp. 2048–2057.

Zanolie, K., van Dantzig, S., Boot, I., Wijnen, J., Schubert, T. W., Giessner, S. R. and Pecher, D. (2012). Mighty metaphors: Behavioral and erp evidence that power shifts attention on a vertical dimension, *Brain and cognition* **78**(1): 50–58.