Visual Conceptual Blending with Large-scale Language and Vision Models

Songwei Ge[1] **and Devi Parikh**[2]
[1]University of Maryland, College Park
[2]Facebook AI Research & Georgia Tech

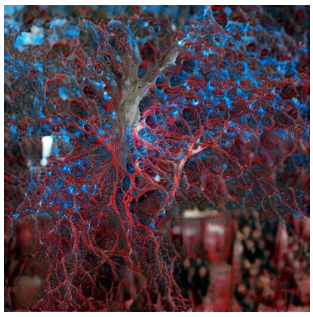
Abstract

We ask the question: to what extent can recent large-scale language and image generation models blend visual concepts? Given an arbitrary object, we identify a relevant object and generate a single-sentence description of the blend of the two using a language model. We then generate a visual depiction of the blend using a text-based image generation model. Quantitative and qualitative evaluations demonstrate the superiority of language models over classical methods for conceptual blending, and of recent large-scale image generation models over prior models for the visual depiction.

Introduction

Throughout the development of human civilization, our unique capacity to blend unfamiliar concepts has led to innovation of advanced tools, invention of new art styles, and breakthroughs in science. Machines demonstrating this ability is considered to be one of the hallmarks of creativity and intelligence. Such systems could help understand human creativity. Moreover, they can assist humans in exploring the inexhaustible space of combinations of different concepts. This has been an area of research for decades (Fauconnier and Turner 1998), which has led to both theoretical work (Cunha, Martins, and Machado 2020) as well as prototypes of support tools to assist users (Karimi et al. 2018; Chilton, Petridis, and Agrawala 2019). In the meantime, deep learning has achieved exceptional success in many areas where humans excelled, from beating the best professional player[(Silver et al. 2016)] in Go to making creative advertising designs[(Brown et al. 2020)].

In this paper, we examine deep neural networks trained on large-scale data in a general scenario of visual conceptual blending: given a single object as input (e.g., moon), can a relevant object [that is conceptually grounded (Cunha, Martins, and Machado 2020)] be identified (e.g., an orange), can a relevant property that a blend can hinge on be identified (e.g., sliced), and finally, can an image be generated to depict the blend (e.g., "the moon sliced like an orange")? We use prompt-engineering with language models for the reasoning phase (identifying a relevant object and property), and text-based image generation models for the visualization phase. See Figure 1 for example outputs.



(a) "A tree made of blue and red blood vessels".

(b) "The moon that is sliced like an orange".

Figure 1: Visual conceptual blends generated by our framework using large-scale language and vision models.

We compare our approach quantitatively and qualitatively to representative existing approaches. To evaluate the ability to associate concepts, we compare our approach to traditional knowledge bases on a simile dataset. To evaluate the visual generation, we compare our approach to an existing GAN approach via human studies. We show that large-scale models significantly outperform these baseline models. In general, we find that an appropriate composition of recent large-scale models results in encouraging creative abilities like visual conceptual blending.

Related Work

Visual Conceptual Blending Fauconnier and Turner first proposed the idea of conceptual blending and pointed out its indispensability in human development (Fauconnier and Turner 1998; 2008). Cognitive and neural scientists have been fascinated by the human ability to blend concepts and view such an ability as a milestone for AI development (Eppe et al. 2018). More practically, the idea of visual conceptual blending has been applied in many commercial areas from advertising, journalism, to public service announcements (Chilton, Petridis, and Agrawala 2019). In this section, we discuss the recent progress in developing systems that automatically blend visual concepts and the studies that measure the success of conceptual blending.

[Computational approaches to conceptual blending such as Divago (Pereira and Cardoso 2006) and COINVENT (Schorlemmer et al. 2014; Eppe et al. 2018) follow

the seminal idea based on Mental Spaces Theory (Fauconnier 1994)]. Many systems developed by these studies act as support tools for augmenting human creativity. (Chilton, Petridis, and Agrawala 2019) [present] a workflow where users identify the associated concepts, retrieve appropriate images, and label the analogous parts of the objects while the system automatically blends the images by combining these common parts. [Vismantic (Xiao, Linkola, and others 2015) on the other hand retrieve and preprocess the images for given words, ask a human to pick ideal photos, and automatically combine the images in fixed ways.] (Karimi et al. 2018) [explore] visual conceptual blends in the context of sketching by leveraging the idea of concept shifts. (Cunha et al. 2017) [propose] a description-based method that can blend sketches using detailed annotations. See (Cunha, Martins, and Machado 2020) for a road map of visual conceptual blending. (McCaig, DiPaola, and Gabora 2016; Berov and Kuhnberger 2016) [apply] style transfer models and the deep dream algorithm to render an image in a particular artistic style. (Sbai, Couprie, and Aubry 2021) [study] placing objects in uncommon contexts using a search-and-compose method. Measuring the creativity of visual blends is known to be difficult. Fauconnier and Turner proposed several optimality principles to guide the conceptual blending (Fauconnier and Turner 1998). (Martins et al. 2015) [analyze] what makes a good blend using 15 hybrid animal images and a questionnaire.

Analogical Reasoning with Language Models Language models were first proposed to model the sequential nature of language (Mikolov and Zweig 2012). With the increasing sizes of training data and model capacities, large-scale language models such as BERT (Devlin et al. 2018) fine-tuned on the downstream tasks have dominated standard leaderboards. Interestingly, several recent studies use language models as knowledge bases to solve different problems without training on the task of interest (Petroni et al. 2019; Jiang et al. 2020). These methods rely on task-specific prompts – converting the task of interest to that of language modeling. Letting the language model predict masked parts from the prompt then becomes equivalent to the model solving the task of interest (Petroni et al. 2019; Jiang et al. 2020). We propose to apply a similar idea to concept blending – we design appropriate prompts to identify relevant concepts and properties along which to blend the concepts. Analogical reasoning has also been approached with large-scale knowledge bases (Liu, Wu, and Yang 2017). However, knowledge bases are known to be incomplete and rigid. We argue that this makes them less suitable for associating concepts in flexible ways (Cunha, Martins, and Machado 2020).

Deep Generative Models for Images Most state-of-the-art image generation methods are built on either Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) or Variational AutoEncoders (VAEs) (Kingma and Welling 2014). In this paper, we are primarily interested in generating conceptually blended objects. (Bau et al. 2020) [propose] to modify the images through manipulating the intermediate layers in GANs which admits the possibility to blend concepts. In this work we use a textual description of the blend to guide the generation. Text-based image generation models (Reed et al. 2016; Zhu et al. 2019; Tao et al. 2020) are relevant. DALL·E (Ramesh et al. 2021) is one such recent model that uses a pretrained discrete VAE to compress images into low-dimensional vectors and then models the joint distribution of the vectors with text embeddings autoregressively.

Table 1: Top 5 concepts relevant to *moon*, and associated properties using simile-inducing prompts to a BERT model.

concept	property				
ghost	dead	killed	gone	alive	murdered
dream	over	real	complete	gone	broken
rainbow	broken	colorful	green	white	black
beacon	lit	active	red	closed	automated
jewel	lost	gone	precious	beautiful	gold

Approach

Next, we describe how we use large language and image generation models to produce conceptually blended images given an input object. We decompose the visual conceptual blending process into two phases: reasoning and generation.

Reasoning Phase The reasoning phase produces a textual description of the blend. We formulate the problem as follows: given an input object, the model identifies a relevant object and generates a description of the blend of the two. Note that our setting is more general than one where both concepts to be blended are given as input (Cunha, Martins, and Machado 2020). [We explore two prompt engineering approaches, simile-inducing and property-guided prompts, which connect the input objects to other objects that are either generally relevant, or in terms of a specific property.] We use *moon* as the example input to explain the details of our prompt engineering approach.

To identify a relevant object, we use a simile-inducing input: "the moon is like a [MASK]" and ask the language model to predict the masked word. The language model produces *ghost*, i.e. "the moon is like a ghost". Next, we utilize the prompt "the ghost has the property of [MASK]", where the language model predicts the word *dead*. We plug the predictions into a template and produce the description of the blend "a moon that is dead like a ghost". Other concepts and their properties identified using a pretrained BERT (Devlin et al. 2018) model are shown in Table 1. Sometimes the retrieved objects are semantically similar rather than visually similar to the *moon* such as *ghost* and *dream*. We see some interesting blends such as "a moon that is lit like a beacon" and "a moon that is broken like a rainbow".

Shape is often recognized as the bridge to connect different visual concepts (Steinbrück 2013; Chilton, Petridis, and Agrawala 2019). This motivates a shape-guided prompt to identify relevant objects. Specifically, we first use language models to predict the shape of the *moon* with the prompt "The shape of the moon is [MASK1]". The language model outputs *spherical*. This gives us "The shape of the moon is spherical". Then we plug the word *spherical* into the prompt

Table 2: Top 5 concepts relevant *moon*, and associated properties using shape-guided prompts to a BERT model.

concept	property				
shell	white	smooth	thin	brown	small
head	rounded	black	white	brown	small
fruit	edible	white	yellow	red	purple
egg	white	yellow	laid	blue	red
eye	open	closed	small	red	black

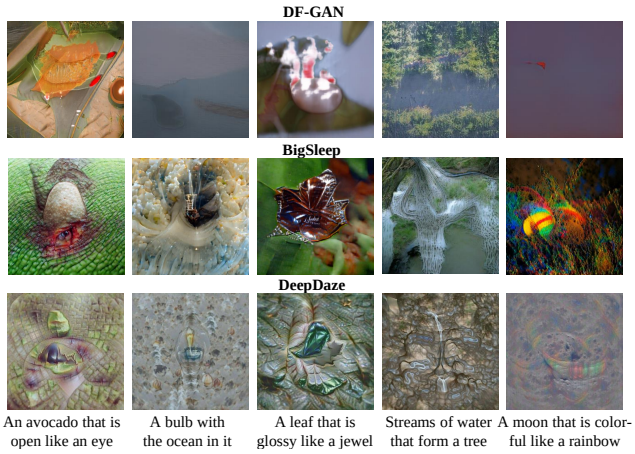


Figure 2: Visual blends generated using different methods using blend descriptions shown at the bottom as input.

"The shape of the [MASK2] is spherical", and the language model predicts the relevant object *shell*, i.e. "The shape of the shell is spherical". This leads to a blend description "a moon that is smooth like a shell" with the property *smooth* of the *shell*. More identified concepts and their properties are shown in Table 2. We find that the candidate concepts we obtain are visually similar to the *moon* in terms of shape. Some interesting descriptions include "a moon that is laid like a egg" and "a moon that is edible like a fruit". In practice, shape can be replaced by other properties that connect visual concepts. For example, speed connects *bullet* and *runner* and reflection connects *mirror* and *lake*.

Generation Phase In this phase we generate an image based on the description output by the reasoning phase. To demonstrate the ability of large models in realizing the blends, we explore BigSleep[1] and DeepDaze[2] which utilize the CLIP model (Radford et al. 2021) to guide the BigGAN (Brock, Donahue, and Simonyan 2019) and SIREN (Sitzmann et al. 2020) models for text-based image generation.

Specifically, suppose we are given a trained CLIP model $f_\theta(x_i, x_s)$ which takes an image x_i and a sentence x_s as input and outputs the similarity, and a trained BigGAN model $g_\phi(z)$ which takes a random Gaussian vector z as input and outputs an image. We first sample a vector z_0 from a standard Gaussian distribution. z is iteratively updated to maximize the similarity of the generation $g_\phi(z_t)$ and the text x_s as computed by the CLIP model $f_\theta(x_i, x_s)$. DeepDaze adopts a similar process with BigGAN replaced by SIREN.

Overall, we now have a full pipeline to go from an input concept (e.g., *moon*) to a description of its blend with a related concept (e.g., "a moon that is sliced like an orange") to an image that depicts this blend (e.g., Figure 1).

Evaluation

Reasoning Phase To evaluate how well language models blend concepts, we evaluate on the simile dataset (Chakrabarty, Muresan, and Peng 2020). It contains pairs of literal input and its simile version in the form of <*Source*, *Target*>, e.g. <The city was beautiful, The city was like a painting>. It evaluates the model's ability to identify "painting" based on "the beautiful city". However, we found that the language is inconsistent across the dataset. For instance, many pairs lack a subject or use a pronoun as subject, e.g. <Felt worthless, Felt like a low budget film>. We instead focus our evaluation on the model's ability to accomplish the core reasoning step – predicting the property "worthless" based on the object "a low budget film". Using heuristics for pre-processing, we extracted $66,442$ property-objects pairs for evaluation.

We compare language models to knowledge bases. For the language model we use the prompt "a low budget film is [MASK]" as the input and ask the model to generate candidate predictions for the masked word. We consider 4 trained language models: ELMO (Peters et al. 2018), $\text{BERT}_{\text{Base}}$ and $\text{BERT}_{\text{Large}}$ (Devlin et al. 2018), and GPT (Radford et al. 2018). For knowledge base, we use ConceptNet (Speer, Chin, and Havasi 2017) which contains relations including "IsA", "HasA", "HasAProperty", etc., which form candidate predictions for properties relevant to the object.

Note that sometimes the object in our dataset is described as a phrase including qualifiers (e.g., "a low budget film") while ConceptNet only contains the root objects. We use dependency parsing to find the root of the phrase and use it to query ConceptNet. In our example, "film" instead of "a low budget film" is used. After this processing, 96.34% of objects from our evaluation set can be found in ConceptNet.

For each method, we produced 1000 candidates, and report the precision, i.e. percentage of time that the property (e.g., "worthless") is in the top 10, 100, 1000 candidates. Note that the ConceptNet API does not offer a straightforward way to request an exact number of relations for an object. Different objects have different number of properties associated with them. When requesting 1000 relations for objects in our evaluation set, 688.90 were returned on average. As shown in Table 3, the precision using ConceptNet is significantly lower than using language models.

Additionally, we notice that using larger language models can further improve the precision. In general, these results demonstrate that language models are better at associating concepts than knowledge bases. We hypothesize this is due to their flexibility and comprehensiveness.

Generation Phase We collect 20 text descriptions of blends (see Figure 2 for examples) – half generated with our reasoning approaches and rest by us. We use these descriptions

[1]https://github.com/lucidrains/deep-daze
[2]https://github.com/lucidrains/big-sleep

Table 3: Precision of language models and knowledge base on the simile dataset.

	P@10	P@100	P@1000
ConceptNet	1.12	2.70	5.90
Elmo	0.13	7.69	37.33
$\text{BERT}_{\text{Base}}$	1.59	15.72	53.08
$\text{BERT}_{\text{Large}}$	1.42	15.89	46.56
GPT	2.59	24.84	66.38

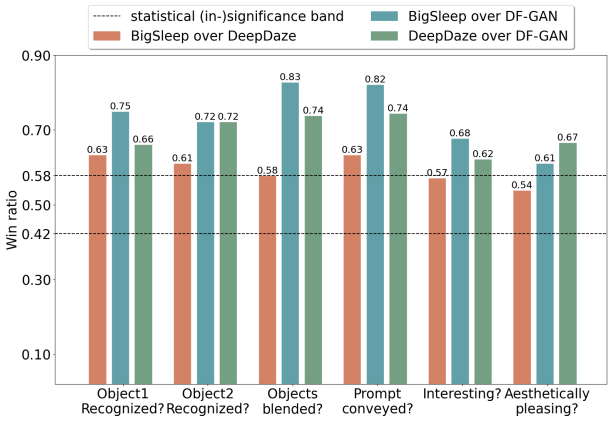


Figure 3: Human preference for different methods w.r.t. different questions. Values outside the band between the dashed lines are statistically significant at 95% confidence.

as input to the large-scale BigSleep and DeepDaze models described earlier, as well as a recent DF-GAN (Tao et al. 2020) model. We run human evaluation on Amazon Mechanical Turk (AMT). We show subjects a pair of images generated by different methods to depict the visual blend of two objects and ask six questions: 1. *In which image do you recognize* OBJECT1 *more?* 2. *In which image do you recognize* OBJECT2 *more?* 3. *Which image blends the two objects better?* 4. *Which image conveys the* DESCRIPTION *better?* 5. *Which image looks more interesting to you?* 6. *Which image looks more aesthetically pleasing to you?* These are designed using the optimality principles for concept blending [(Fauconnier and Turner 1998; Cunha, Martins, and Machado 2020)]. Specifically, 1 and 2 relate to the unpacking principle, 3 and 4 to the integration principle, and 5 and 6 to general quality. Each question (6) for every pairwise comparison of models (3) and every textual description (20) is answered by 9 unique subjects.

See results in Figure 3. The CLIP-based models (BigSleep and DeepDaze) significantly outperform DF-GAN, demonstrating the superiority of large models in generating visual blends. BigSleep is preferred over DeepDaze. We conjecture that this is because BigGAN learns a better prior on the image distribution than SIREN.

Conclusion

In this paper, we apply large-scale language and image generation models to a classic computational creativity problem – visual conceptual blending. Our experiments show that these models allow us to use simple yet effective ways to generate visual blends that are significantly better than previous methods. Future work includes engineering novel prompts to connect concepts and developing more complex blending strategies given the identified concepts. For example, the classic blend of boat and house (houseboat) – "a man lives in a house that is built on the water like a boat" – considers structural relationships of the objects and includes two different properties from the two objects – a place of accommodation (from house) and being on water (from boat).

References

Bau, D.; Liu, S.; Wang, T.; Zhu, J.-Y.; and Torralba, A. 2020. Rewriting a deep generative model. In *ECCV*.

Berov, L., and Kuhnberger, K.-U. 2016. Visual hallucination for computational creation. In *ICCC*. Citeseer.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large scale gan training for high fidelity natural image synthesis.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv*.

Chakrabarty, T.; Muresan, S.; and Peng, N. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *EMNLP*.

Chilton, L. B.; Petridis, S.; and Agrawala, M. 2019. Visiblends: A flexible workflow for visual blends. In *CHI*.

Cunha, J. M.; Gonçalves, J.; Martins, P.; Machado, P.; and Cardoso, A. 2017. A pig, an angel and a cactus walk into a blender: A descriptive approach to visual blending. *ICCC*.

Cunha, J.; Martins, P.; and Machado, P. 2020. Let's figure this out: A roadmap for visual conceptual blending. In *ICCC*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*.

Eppe, M.; Maclean, E.; Confalonieri, R.; Kutz, O.; Schorlemmer, M.; Plaza, E.; and Kühnberger, K.-U. 2018. A computational framework for conceptual blending. *A.I.*

Fauconnier, G., and Turner, M. 1998. Conceptual integration networks. *Cognitive science*.

Fauconnier, G., and Turner, M. 2008. *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

Fauconnier, G. 1994. *Mental spaces: Aspects of meaning construction in natural language*. Cambridge University Press.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial networks. *NIPS*.

Jiang, Z.; Xu, F. F.; Araki, J.; and Neubig, G. 2020. How can we know what language models know? *TACL*.

Karimi, P.; Maher, M. L.; Grace, K.; and Davis, N. 2018. A computational model for visual conceptual blends. *IBM JRD*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. *ICLR*.

Liu, H.; Wu, Y.; and Yang, Y. 2017. Analogical inference for multi-relational embeddings. In *ICML*. PMLR.

Martins, P.; Urbancic, T.; Pollak, S.; Lavrac, N.; and Cardoso, A. 2015. The good, the bad, and the aha! blends. In *ICCC*.

McCaig, G.; DiPaola, S.; and Gabora, L. 2016. Deep convolutional networks as models of generalization and blending within visual creativity. *arXiv*.

Mikolov, T., and Zweig, G. 2012. Context dependent recurrent neural network language model. In *SLT*. IEEE.

Pereira, F. C., and Cardoso, A. 2006. Experiments with free concept generation in divago. *Knowledge-Based Systems* 19(7):459–470.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language models as knowledge bases? In *EMNLP*.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. *arXiv*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *ICML*. PMLR.

Sbai, O.; Couprie, C.; and Aubry, M. 2021. Surprising image compositions. In *CVPR Workshops*.

Schorlemmer, M.; Smaill, A.; Kühnberger, K.-U.; Kutz, O.; Colton, S.; Cambouropoulos, E.; and Pease, A. 2014. Coinvent: Towards a computational concept invention theory.

Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature* 529(7587):484–489.

Sitzmann, V.; Martel, J.; Bergman, A.; Lindell, D.; and Wetzstein, G. 2020. Implicit neural representations with periodic activation functions. *NeurIPS*.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Steinbrück, A. 2013. Conceptual blending for the visual domain. *Ph. D. dissertation, Masters thesis*.

Tao, M.; Tang, H.; Wu, S.; Sebe, N.; Wu, F.; and Jing, X.-Y. 2020. Df-gan: Deep fusion generative adversarial networks for text-to-image synthesis. *arXiv*.

Xiao, P.; Linkola, S. M.; et al. 2015. Vismantic: Meaning-making with images. In *Proceedings of the Sixth International Conference on Computational Creativity*. Brigham Young University.

Zhu, M.; Pan, P.; Chen, W.; and Yang, Y. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*.