Scientific and Creative Analogies in Pretrained Language Models

Tamara Czinczoll[♣♡] **Helen Yannakoudakis**[♠] **Pushkar Mishra**[♢] **Ekaterina Shutova**[♣]
[♣]ILLC, University of Amsterdam, the Netherlands
[♢]Meta AI, London, United Kingdom
[♠]Dept. of Informatics, King's College London, United Kingdom
[♡]Hasso Plattner Institute/University of Potsdam, Germany
tamara.czinczoll@hpi.de, helen.yannakoudakis@kcl.ac.uk, pushkarmishra@meta.com, e.shutova@uva.nl

Abstract

This paper examines the encoding of analogy in large-scale pretrained language models, such as BERT and GPT-2. Existing analogy datasets typically focus on a limited set of analogical relations, with a high similarity of the two domains between which the analogy holds. As a more realistic setup, we introduce the **S**cientific and **C**reative **An**alogy dataset (SCAN), a novel analogy dataset containing systematic mappings of multiple attributes and relational structures across dissimilar domains. Using this dataset, we test the analogical reasoning capabilities of several widely-used pretrained language models (LMs). We find that state-of-the-art LMs achieve low performance on these complex analogy tasks, highlighting the challenges still posed by analogy understanding.

1 Introduction

Analogy-making is a cornerstone of human intelligence (Gentner et al., 2001), allowing us to acquire new knowledge and creatively explore new concepts. According to Gentner (1983)'s *Structure-Mapping Theory*, analogical reasoning is different from surface similarity. Instead, the attributes of a familiar concept (the source domain) are mapped to something less familiar (the target domain) if their *relational structures* are similar enough. For example, while not directly similar in their attributes, the underlying relational structure of the *solar system* matches that of an *atom*. The relationship between two source domain attributes (e.g. *sun* and *planet*) helps us to understand that between their target domain counterparts (*nucleus* and *electron*).

Within natural language processing (NLP), the word analogy task (Mikolov et al., 2013), has been widely used to demonstrate analogical reasoning capabilities of pretrained word embedding models. The task involves solving analogies of the form *A:B :: C:D* (i.e., *A is to B as C is to D*) by exploiting (local) linear properties of word vectors (vector offsets). Subsequently, word analogy became one of the standardized tasks for intrinsic evaluation of word embedding quality. However, Gladkova et al. (2016) showed that the vector offset method was not sufficient for most types of analogical relations, and Rogers et al. (2017) pointed out shortcuts that the models were taking. Existing word analogy datasets focus on a limited set of analogical relations, include words that are semantically similar and do not require the model to relate distinct concepts via systematic comparison of their relational structures, all of which is necessary for human-like analogy making. In parallel, the field has seen the development of large-scale pretrained sentence encoders, whose analogical reasoning capabilities have not yet been fully tested.

To address these issues, we devise and release a new dataset – the **S**cientific and **C**reative **An**alogy dataset (**SCAN**) – comprising holistic analogies between concepts from semantically distant domains. It draws on metaphorical and scientific analogies. Resolving these analogies requires the models to identify systematic ontological correspondences between two distinct semantic domains, such as in the *solar system – atom* example. Our contributions are threefold: 1) We present the **SCAN** analogy evaluation task and dataset, which we make publicly available to the research community; 2) We systematically evaluate current state-of-the-art LMs on the established BATS dataset (Gladkova et al., 2016), which consists of a large number of traditional word analogies, as well as the novel SCAN dataset. We show that in the latter case the models exhibit severe limitations in understanding analogies; 3) We show that a high performance on BATS is not indicative of how well the models solve the complex SCAN analogies, supporting our hypothesis that BATS does not require full analogical reasoning.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2094–2100
December 7-11, 2022

2 Related Work

Turney (2008) presented an algorithm for analogy solving and tested it on 20 scientific and metaphorical examples, where a source domain is mapped to a different target domain, along with a number of its attributes (e.g. *waves* to *sound*). While few, these examples were true to human analogy-making, representing a wide range of semantic relationships. Using a linear offset method (*3CosAdd*), Mikolov et al. (2013) demonstrated that their word embeddings automatically capture analogical information about word relationships, so that $emb(king) - emb(man) + emb(woman) \approx emb(queen)$, where emb is the embedding function represented by the neural network. Gladkova et al. (2016), however, showed that the pretrained word embeddings at the time could only reliably complete word analogies for inflectional morphology categories while struggling on many semantic categories. They released a balanced, larger and more diverse dataset than Mikolov et al. (2013)'s (40 vs 15 relation types), the *Bigger Analogy Test Set* (BATS), demonstrating that Word2Vec was not able to solve most types of word analogies. In particular, a larger semantic distance between the source and target domains resulted in low performance (Rogers et al., 2017).

Transformer language models such as BERT (Devlin et al., 2019) have pushed the state-of-the-art on a number of NLP tasks. But since 3CosAdd cannot easily be applied to them, due to their word embeddings not being fixed but dynamically calculated, their analogical capabilities have not been investigated much. While some headway has been made in that regard (Li and Zhao (2020), Zhu and de Melo (2020), Ushio et al. (2021)), the focus has been on transferring the traditional word analogy datasets to the sentence level. This does, however, also transfer their limitations. In general, the low performance of models in Gladkova et al. (2016) and Zhu and de Melo (2020) suggest that completing word analogies is challenging for state-of-the-art LMs, even when structure mapping across distinct domains is not explicitly tested.

3 SCAN Dataset

Our dataset contains 449 analogy instances, clustered into 65 full concept mappings. A source concept is mapped to a target concept along with a number of related attributes. Table 1 provides two examples. When mapping from the source concept *War* to the target concept *Argument*, a number of relevant attributes' correspondences are given. The number of attributes per cross-domain mapping is not fixed. The dataset includes the 20 mappings from Turney (2008) (10 scientific and 10 metaphorical) and extends them by another 43 metaphorical mappings and 2 scientific ones. The new metaphorical mappings include conceptual metaphors from the Master Metaphor List (Lakoff et al., 1991) and other conceptual metaphors widely-discussed in linguistic literature (Lakoff and Johnson, 1980; Musolff, 2000; Lakoff and Wehling, 2012). Each conceptual metaphor was then annotated for attribute correspondences by three metaphor experts. First, the semantic frames of the source and target domains were identified and then the correspondences between individual frame elements were established (see Tab. 1). We build a word analogy task from this data by defining the cross-domain mappings (e.g., *Argument* and *War*) as the first word pair, and the attribute mappings (in this case *Debater* and *Combatant*) as the to-be-completed second word pair. Since each concept includes multiple attributes, a total of 449 word analogies are constructed.

SCAN offers richer and more holistic analogies than traditional word analogy datasets. Taking a statistical view, the chances of the words in the source and target domains co-occurring are much lower than in BATS. For example, countries and their capitals, animals and the sounds they make, and most grammatical analogy types in BATS are quite likely to occur in the training corpus together. However, the same cannot be said for most SCAN analogies, meaning that true analogical transfer needs to occur.

Additionally, the in-domain words in SCAN are semantically more dissimilar. For example, in the argument domain, *debater* and *topic* are fundamentally different concepts. In BATS, on the other hand, every domain member is another instance of the same concept, e.g. *France* and *Germany* instances of a country.

Lastly, the analogical relationships between domains in SCAN are more abstract than those in BATS. To successfully extract the same relationship from *solar system – atom* and *planet – electron*, more abstraction and inference over the relational structure of the domains is needed than in BATS. In BATS, the relationships between, e.g. *France – Paris* and *Germany – Berlin*, are straightforward

Target	Source	Attribute	mapping
Argument	War	Debater	Combatant
		Topic	Battleground
		Claim	Position
		Criticize	Attack
		Rhetoric	Maneuver
Code	Virus	Malware	Virus
		Replication	Reproduction
		Installation	Infection
		Removal	Eradication
		Antivirus	Vaccine

Table 1: Example mappings in SCAN. For one source concept multiple relevant attributes are mapped to the corresponding target concept's attributes.

and do not require much abstraction.

Overall, SCAN offers more human-like analogies by employing more diverse in-domain words, more abstract mapping relations and by avoiding obvious co-occurrences. Due to its full-concept mappings, SCAN is not confined to the word analogy task. By holistically mapping entire source domains to a new target domain we want to encourage a broader range of analogy representations.

4 Models

We probe the analogical capabilities of several widely used language models: GPT-2, BERT and Multilingual BERT (M-BERT). We use GloVe as a baseline, given it has been shown to outperform language models on some relation types in previous analogy tasks (Zhu and de Melo, 2020).

GPT-2 (Radford et al., 2019) can be viewed as a "true" LM since it is trained to predict the next word in a sequence, and can be used for language generation. It is a transformer-based model with 48 layers and 1542M parameters, trained on a custom dataset, *WebText*, created only from outbound links from Reddit to improve text quality. Due to its predictive nature, GPT-2 is one-directional, i.e. only the context on the left-hand side influences the prediction of the next word.

BERT (Devlin et al., 2019) is a bidirectional language representation model. It is trained with two objectives: masked-token prediction and next-sentence prediction. We use BERT-base with 12 layers in our experiments (110M parameters). Since BERT is bidirectional, it can incorporate information from both sides of the masked token.

M-BERT is a BERT model, trained on a Wikipedia dump of 100 languages. The model performs best on high-resource languages such as English, French and Chinese, since lower-resource languages are underrepresented in the training data. We test whether M-BERT's pre-training on a wide range of languages, and thus a wide range of culture-specific analogies, might enhance the model's general analogy understanding.

5 Experiments

Setup We use pretrained model instances of GPT-2, BERT Base and Multilingual BERT [1]. As BERT and GPT-2 are trained on full sentences, we insert the word analogy quadruple into a placeholder sentence. We use "*If* A *is like* B, *then* C *is like* D.", which was selected from a set of candidates as it performed best on the development set. Similarly to Ettinger (2020), who probed BERT with a number of cloze and negation tasks, the models need to predict the last token of the sentence. We force the models to predict word D by either masking it for the two BERT models, or by cutting the sentence off before it for GPT-2. We report the mean reciprocal rank (MRR) of the first token of the target word (or that of one of the alternative answers) in only the top 10 predicted tokens to reduce compute. If the label is not in the top 10 tokens, its RR is 0. Model performance in terms of accuracy, recall@10 and recall@5 is reported in the Appendix. We use an Nvidia 16GB GPU.

SCAN vs. BATS To evaluate how well the models can solve different types of analogy, we test them on **BATS** in addition to SCAN. We do not fine-tune the models. BATS consists of 98000 examples of balanced relations. There are four main relations – inflectional and derivational morphology, and lexicographic and encyclopedic semantics – each of which consists of ten subcategories. For some examples, multiple correct answers are listed.

Zero-shot vs. One-shot Previous work on analogical reasoning in GPT-3 (Mitchell, 2020; the Latitude Team, 2020) has shown that when the model is given a full example of a word analogy in addition to the incomplete one, the performance on the incomplete analogy substantially increases. We see this as a form of one-shot vs. zero-shot testing and also test the models this way, investigating whether this has an impact on the LMs'

[1] https://huggingface.co/

	BATS	SCAN	Science	Meta.
GloVe	.099	.022	.099	.006
GPT-2	.098	**.057**	.073	**.054**
BERT	**.207**	**.044**	.092	**.034**
M-BERT	**.205**	.041	.088	**.031**

Table 2: Model MRR on BATS and SCAN. Statistically significant differences compared to the GloVe baseline are in bold (two-sided permutation test; $p < 0.05$; #resamples=$10e^5$).

performance on SCAN. We insert a complete version of our template sentence before the incomplete one, ensuring that none of the analogy words from the full example appear in the incomplete analogy. Note that GloVe does not benefit from this setup as the vectors used for 3CosAdd remain the same.

Training Set Effects Lastly, we further investigate the difference between the types of word analogies in BATS and those in SCAN. We split the BATS dataset into a train, validation and test set (70/15/15 ratio), ensuring that each word pair appears in only one of them. We fine-tune the LMs on the training set and take each model's version with the best score on the BATS validation set. We train ($\sim$ 4h) all models with the AdamW (Loshchilov and Hutter, 2019) optimizer, a learning rate of $5e^{-5}$ and a batch size of 16 for 4 epochs (based on manual tuning). If the model has learned about general analogy-making it must understand new analogical relations "on-the-fly" and improve not only on the BATS test set but also on SCAN. We expect there to be strong improvements on BATS compared to its untrained counterpart, but little to none on SCAN, showing them to be inherently different. This is not performed on GloVe, since 3CosAdd outputting an embedding is not part of its original training setup.

6 Results & Discussion

Table 2 shows the accuracy of each of the models on the BATS and SCAN datasets, as well as on the scientific and creative analogies separately. BERT achieves the highest MRR on the BATS dataset, with a strong lead compared to the other models. Similarly to Zhu and de Melo (2020), we find that GloVe can keep up with the other models on BATS, performing similarly to GPT-2. However, this trend is not observed on the SCAN dataset, where GloVe is relegated to last place, indicating that context is important for understanding SCAN analogies. All models perform better on the scientific analogies than on metaphors. This could be due to the fact that their attributes are less abstract and have clearer correspondences in the target domain. The models' MRRs are generally lower on SCAN, which we attribute to the greater semantic dissimilarity between source and target domains. GPT-2 achieves the highest performance, followed by BERT. This, combined with its lower accuracy on the BATS baseline, indicates that GPT-2 is better at modeling more extensive and narrative analogies instead of the more artificial and strictly-defined ones in BATS. Multilingual features only appear to be marginally effective for the task, something which can be explained by the fact that most analogies are language-dependent, an observation also made by Ulčar et al. (2020). Overall, these results indicate that the SCAN analogy task is challenging for state-of-the-art LMs and that their true analogy solving capabilities still need to be improved.

	BATS	SCAN	Science	Meta.
GPT-2	**.121**	.048	.056	.046
BERT	**.095**	.035	.077	.027
M-BERT	**.180**	.036	.112	.020

Table 3: MRR when an example sentence is given. Statistically significant differences (two-sided permutation test; $p < 0.05$; #resamples=$10e^5$) compared to each model's baseline in bold.

	BATS	SCAN	Science	Meta.
GPT-2	**.384**	**.022**	.066	**.012**
BERT	**.592**	**.019**	.061	**.010**
M-BERT	**.499**	**.020**	.087	**.006**

Table 4: MRR on the BATS test set as well as on SCAN after training. Statistical significance compared to the baseline (two-sided permutation test; $p < 0.05$; #resamples=$10e^5$) in bold.

Zero vs. One-Shot Table 3 shows model accuracy when the input contains a complete additional example. Apart from GPT-2 on BATS, this does not help the models better understand the task. This contrasts the examples on GPT-3 from Mitchell (2020), possibly due to the models not identifying the analogical relationships in the example sentence.

Training Set Effects After training on BATS, one could expect that if the models learn about analogical reasoning in general, they would also naturally do better on the SCAN dataset with more complex analogies. However, our results in Table 4 show that the opposite is the case. While training on BATS drastically increases the models' MRR

on the held-out BATS test set, it has an adverse effect on SCAN. This suggests that the two datasets' analogy types are innately different, validating our hypothesis that standard word analogy datasets do not adequately represent human analogy use.

Error Analysis While GloVe scores consistently on all relation types in BATS, this is not the case for the other models. On SCAN, none of the models predict the mappings of all attributes of a concept (or even most of them) correctly. While the models are able to solve some individual mappings, the fact that they cannot apply this to all aspects of the concept indicates that none of them are really able to grasp the workings of analogy. In cases where analogies remain entirely unsolved, it is likely that the required domain knowledge is lacking.

7 Conclusion

Analogical reasoning remains a challenging task even when state-of-the-art Transformer LMs are used. We have shown that, even with models such as BERT and GPT-2, there is large room for improvement on automated reasoning and understanding of realistic analogies. We have introduced a new dataset, SCAN, that is different from existing word analogy datasets in that it is composed of whole concept mappings across semantically dissimilar domains, demonstrating that popular LMs are unable to fully understand these analogies. We further tested whether a full example of the task can help the models, finding that this is not helpful in our setup. Lastly, our results indicate that the SCAN analogies are substantially different from those of traditional word analogy datasets. Improving on them is a line of research we wish to investigate further in the future. We make SCAN and the related code publicly available.[2]

8 Limitations

Our experimental design focuses on evaluating the models' analogical capabilities in a generative setting. We see value in this, as analogical reasoning is inherently a generative cognitive function. The BERT models are, however, not trained to perform left-to-right generation and, furthermore, rely on wordpiece vocabulary for tokenization. The evaluation of its predictions in the analogy task is, therefore, less straightforward and not exactly comparable to other models. We adapt the task for BERT models by letting them only predict the first token of the missing answer. Comparing only the first token leaves some variability, however, when matching the prediction and the right answer. We expect this effect to be limited in English due to its sparse morphology.

Furthermore, the metaphorical analogies come from English literature and cultural background. It would be interesting to compare these with analogies from other languages and cultures to investigate whether the language models' lack of understanding is due to encoding of language-specific properties, missing domain knowledge or the general analogical mapping abilities.

Lastly, some metaphors in SCAN exhibit antiquated gender roles, e.g. the metaphor "government:household :: governor:father". While these relationships are culturally often still relevant for metaphor understanding, the underlying implied gender roles need to be treated carefully and issues of their encoding by neural models investigated further.

[2] https://github.com/taczin/SCAN_analogies

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy*. *Cognitive Science*, 7(2):155–170.

Dedre Gentner, Keith Holyoak, and Boicho Kokinov. 2001. *The Analogical Mind: Perspectives From Cognitive Science*. MIT Press.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: What works and what doesn't. In *Proceedings of the NAACL-HLT SRW*, pages 47–54, San Diego, California, June 12-17, 2016. ACL.

George Lakoff, Jane Espenson, and Alan Schwartz. 1991. The master metaphor list. Technical report, University of California at Berkeley.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago.

George Lakoff and Elisabeth Wehling. 2012. *The Little Blue Book: The Essential Guide to Thinking and Talking Democratic*. Free Press, New York.

Yian Li and Hai Zhao. 2020. Learning universal representations from word to sentence.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Melanie Mitchell. 2020. Can gpt-3 make analogies? https://medium.com/@melaniemitchell.me/can-gpt-3-make-analogies-16436605c446. Accessed: 2021-05-05.

Andreas Musolff. 2000. *Mirror images of Europe: Metaphors in the public debate about Europe in Britain and Germany*. Iudicium, Muenchen.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 135–148, Vancouver, Canada. Association for Computational Linguistics.

the Latitude Team. 2020. World creation by analogy. https://aidungeon.medium.com/world-creation-by-analogy-f26e3791d35f. Accessed: 2022-10-19.

P. D. Turney. 2008. The latent relation mapping engine: Algorithm and experiments. *Journal of Artificial Intelligence Research*, 33:615–655.

Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. 2020. Multilingual culture-independent word analogy datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4074–4080, Marseille, France. European Language Resources Association.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3609–3624, Online. Association for Computational Linguistics.

Xunjie Zhu and Gerard de Melo. 2020. Sentence analogies: Linguistic regularities in sentence embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3389–3400, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Additional Evaluation Metrics

Model	BATS				SCAN				SCAN Science				SCAN Meta.			
	Acc	MRR	Rec@10	Rec@5	Acc	MRR	Rec@10	Rec@5	Acc	MRR	Rec@10	Rec@5	Acc	MRR	Rec@10	Rec@5
GloVe	.004	.099	.240	.192	.018	.022	.031	.024	.091	.099	.110	.097	.003	.006	.014	.009
GPT-2	**.044**	.098	**.250**	**.172**	.020	**.057**	**.167**	**.107**	.026	.073	.195	.143	.019	**.054**	**.161**	**.099**
BERT	**.126**	**.207**	**.401**	**.316**	.009	**.044**	**.134**	**.080**	.026	.092	**.273**	.182	.005	**.034**	**.105**	**.059**
MultBERT	**.141**	**.205**	**.341**	**.289**	.018	.041	**.100**	**.067**	.065	.088	.130	.117	.008	**.031**	**.094**	**.056**

Table 5: Accuracy, MRR, Recall@10 and Recall@5 on the two datasets. Statistically significant differences compared to the GloVe baseline are in bold (two-sided permutation test; $p < 0.05$; #resamples=$10e^5$).

Model	BATS				SCAN				SCAN Science				SCAN Meta.			
	Acc	MRR	Rec@10	Rec@5	Acc	MRR	Rec@10	Rec@5	Acc	MRR	Rec@10	Rec@5	Acc	MRR	Rec@10	Rec@5
GPT-2	**.050**	**.121**	**.282**	**.218**	.018	.048	.147	.087	.013	.056	.169	.104	.019	.046	.142	.083
BERT	**.054**	**.095**	**.207**	**.152**	.016	.035	.107	.060	.052	.077	.143	.117	.008	.027	.099	.048
MultBERT	**.123**	**.180**	**.307**	**.256**	.020	.036	.071	.051	.091	.112	.143	.13	.005	.020	.056	.035

Table 6: Accuracy, MRR, Recall@10 and Recall@5 when an example sentence is given. Statistically significant differences (two-sided permutation test; $p < 0.05$; #resamples=$10e^5$) compared to each model's baseline in bold.

Model	BATS				SCAN				SCAN Science				SCAN Meta.			
	Acc	MRR	Rec@10	Rec@5	Acc	MRR	Rec@10	Rec@5	Acc	MRR	Rec@10	Rec@5	Acc	MRR	Rec@10	Rec@5
GPT-2	**.305**	**.384**	**.550**	**.482**	.011	**.022**	**.051**	**.036**	**.039**	.066	.143	.117	.005	**.012**	**.032**	**.019**
BERT	**.501**	**.592**	**.756**	**.717**	.004	**.019**	**.051**	**.031**	.026	.061	**0.130**	.091	.000	**.010**	**.035**	**.019**
MultBERT	**.420**	**.499**	**.661**	**.604**	.016	**.020**	**.038**	**.022**	.078	.087	.156	.078	.003	**.006**	**.013**	**.011**

Table 7: Accuracy, MRR, Recall@10 and Recall@5 on the BATS test set as well as on SCAN after training. Statistically significant differences (two-sided permutation test; $p < 0.05$; #resamples=$10e^5$) compared to each model's baseline in bold.