One Picture and a Thousand Words

Generative Language+images Models and How to Train Them

Roberto Zamparelli[1]

[1]CIMEC, University of Trento / Corso Bettini 31, Rovereto TN, 38068, Italy

Abstract

Thanks to independent advances in language and image generation we could soon be in the position to have systems that communicate with humans by combining language and images in their *output*, a skill that humans do not possess (we receive, but do not produce images at high speed). The paper explores some of the implications of this idea: which kinds of data sets need to be developed to train such systems, in which cases language and images could be most usefully integrated and which issues could arise on the image generation and language+image integration side.

Keywords

Language generation, image generation

1. Introduction: a superhuman task

Human beings acquire information about their environment in two ways: by witnessing it through the senses, directly or indirectly (*seeing* a car crash, *hearing* the noise, *smelling* gasoline etc.) or by obtaining reports through language (hearing *that* a car has crashed, *when/how* it happened, etc.). Our input may be multimodal, but the main distinction is between symbol-mediated information (language and other communication codes) and any type of (direct) sensory information.

When we turn to the *output*, to the way our species produces information for others to understand, the choice is remarkably narrower. We have evolved to produce language and body postures (gestures), interactively and at high speed; we have *not* evolved to be able to show the specific way in which a car crashed, or imitate the exact noise its breaks made. If we use gestures to try to convey an iconic aspect, the result is understood to be an approximation: the way I move my hand may be indicative of the way the car slid, but a number of other aspects remain unexpressed.

Time and technology have helped us narrow the gap between our inputs and outputs. Given enough time (and skills that many people with a good use of language don't often have) we can draw what happened, make graphs of the car's deceleration or create physical models of all the objects involved. Affordable cell phone technology now allows witnesses to take pictures or videos, vastly expending our possibility of *showing*, not just *telling*, what took place. This, however, does not extend to what *did not* happened, or has not happened *yet*. I can explain

NL4AI 2023: Seventh Workshop on Natural Language for Artificial Intelligence, November 6-7th, 2023, Rome, Italy [1]

roberto.zamparelli@unitn.it (R. Zamparelli)

0000-0002-6890-8723 (R. Zamparelli)

how (maybe, eventually) Starships will land on Mars, but showing this graphically requires professional video makers and special effects.

The goal of this piece is to point out that AI systems from the near future will *not* have these limitations: they could in principle be able to produce language *and* still images (soon, moving images and 'stage noises'), mixing them in any way we can imagine and in many we cannot (yet). Moreover, this mix could be fast enough to take place in *dialogue*: a chatbot could answer my questions with an image accompanied by verbal explanations, adjust the image if it is unclear, answer my questions by expanding details, link referring expressions to points in the picture, render focus by unblurring parts, making sections transparent or adjusting colors, etc. None of these capabilities exist in humans: gestures are too generic, photos are restricted in style and possibly too detailed, drawing is too slow to be usable without interrupting the flow of the conversation. Put otherwise, combining language and images in on-line generation is a task in which AI systems not from the realm of science-fiction (due to independent advances in language generation and image generation, see Sec. 2) could very soon achieve superhuman abilities and produce actually useful output.[1]

This unique status comes, of course, with a number of problems. The aspects where NLP or artificial vision have historically progressed the fastest are those where they could leverage huge datasets, annotated via crowd-sourcing & bootstrapping. Humans do create combinations of images and texts in various contexts and with various purposes, but these contexts and purposes appear to be a small subset of the cases where it would be useful for an AI to generate what I am going to call I+L (Images+Language): none of these contexts and cases is both fast and interactive.

The problem of creating systems that can output both language and images has actually been around since the 80ies (see [2]) often under the name Multimodal Output Generation (MOG, see [3]). In the previous literature, however, it was seen as a procedural problem: finding the bast way to combine textual explanations with either pre-stored materials (e.g. product pictures, see [4], medical images, etc.) or data-generated graphs [2]. The perceived central issue was planning what to serve where, and in which modality. Machine learning, when present, was used to find the best combination of the two modalities [5], not as a way to generate either of them. To take advantage of the additional possibilities given by modern generation models we need to turn our attention to three aspects: *I+L Training*, *I+L Theory* and *I+L Input*. This paper gives an overview of some of the issues that arise within these domains, mostly seen from the vantage point of language and linguistics.

The paper is organized as follows. Sec. 2 gives a landscape of the current AI models the I+L idea builds on, and discusses the (lack of) current datasets. In Sec. 3 I highlight some questions raised by the development of I+L models. Sec. 4 discusses possible ways to develop training materials, while a few linguistic issues arising from the combination of images and language are discussed in Sec. 5.

[1] As a side note, no deepfakes are possible at this level: a system that is seen answering questions by producing intertwined text and images could not be mistaken for a human production simply because no human could accomplish this feat. Of course, the final output could raise the same issues or authorship and attribution as other artificial creations, and give raise to the security concerns raised by its individual constituents.

2. Background and middle-hanging fruits

The possibility of generating I+L builds on recent advances in a number of neighbouring tasks: language generation (OpenAI's GPT3, [6], Google's PaLM, [7], META's OPT, [8], etc.), text-to-image (T2I) models (GLIDE, [9], DALL-E, [10]), text&image-to-image generation [11] and multiple-round image editing with language [12, 13, 14]. Many of these advances have been made possible by a move from annotated images datasets such as ImageNet [15] or MS-COCO [16] to image embeddings like CLIP [17], trained on image-text matching collected on a huge scale. CLIP does not have a limited vocabulary and has been shown to improve VQA tasks [18], including image captioning [19], another task quite relevant for I+L. Other advances come from applying to image generation the textual embeddings developed by very recent, large-scale language models (e.g. T5 in IMAGEN, [20]), combined with image upscaling. Given the strict connection between language and images required by a I+L system, another relevant strand of research is the one that aims to go from images to comprehensive descriptions, far longer than typical user-generated captions, either by associating captions to the individual bounding boxes (*Dense Captioning*, [21]) or by creating extensive human-readable descriptions from images (*Image Paragraph*, [22]). The aim of some of this work is assistive technology for the visually impaired [23], but they could be very useful to map linguistic descriptions of entities back to the images, or as part of a circular flow that goes from language to images and back to language (see Fig. 4).

Overall, the field has seen an extremely fast growth, with new open-source data made available monthly (e.g. the CLIP-based LAION-5B[2] dataset), ongoing efforts to make models trainable at lower costs [24], and the release of open-source T2I software for low-end hardware (Stable Diffusion[3]), which is likely to create a huge pool of testers for future models.

In all the cases considered, however, the goal is to produce *images* (from text or text+images), or *text* which either describes the image (automatic captioning) or uses the image as ground truth (e.g. in VQA, [25]). These tasks are generalizations drawn from innumerable examples of text-image associations (in the form of human-generated captions, textual contexts or image metadata). The goal of the sort of I+L system envisoned here is different: the visual and the linguistic side of the output should provide contents that are both *connected* and *complementary*. There have been attempts at creating datasets where the text and images are more complementary (see BD2BB, [26]), but there are many more ways to be complementary than to coincide, and more systematic ways to generate training materials must be found. At the time of writing, the Task section of the huge Paperswithcode Dataset[4] repository does not contain any task resembling image *and* text generation.

One domain where humans routinely combine voice and images are presentations with slides (Powerpoint©, etc.). Large public slide repositories do exist (e.g. the Slideshare-1M[5]), but no dataset that combines the speaker's voice (or its transcription) with the slide presented. TED Talks, another possible model of a certain way of combining speech and images, have been compiled in various machine-learning-friendly forms, but not in one that systematically

[2] https://laion.ai/blog/laion-5b/
[3] https://stability.ai/blog/stable-diffusion-public-release
[4] https://paperswithcode.com/datasets
[5] https://purl.stanford.edu/mv327tb8364

associates the slides used with the speech (TED Talks[6] has voice and videos of the speaker's body; TED LIUM 3[7] only audio and transcripts; the TED Gesture Dataset[8] could be used to teach systems to accompany speech with gestures, but obviously the potential of I+L generation goes much beyond this).

Next, there are instructional videos (on products, procedures, DIY projects, etc.). These could provide valuable training material, as long as they do no not (mostly) show the speaker. Turning them into I+L-training material would involve reducing the video to maximally informative still frames of the type/resolution that could be matched by current T2I systems, then linking them to the corresponding voice passages. Note that multiple video frames could be used to generate a 3D representation of the objects shown (e.g. using NeRF-like techniques, [27]). In this case, the final dataset would contain a paired combination of text/voice and the 3D embeddings of the entities mentioned at every point. An I+L system could learn to accompany the text passage with a single 3D image, giving the listener the possibility to explore it interactively (asking for rotation, zooming, etc.).

Finally, there are commercial *movies*, and there are *comics*. When the former are combined with storyline datasets like MPST [28] it may be possible to extract still frames that best highlight specific storyline passages and that could be used for I+L training (see Phenaki[9] [29] for an attempt along this line, with videos, minus the textual output). Unlike with videos, it remains to be seen if a movie's storyline is not too coarse to be meaningfully mapped to the still frames. The case of comics is in a way the literary genre that comes closest to the possible output of a type of I+L system. The problem, in this cases, is in the very limited range of topics and themes that the comics literature covers (especially if one wants to exclude copyrighted material). Comics could be useful to train a I+L system that produces just that — comics, but their utility at large remains to be proved.

3. Open questions

As we have need, current AI research has a large variety of datasets available for training models, yet none seems ready and set to be used for I+L development. The questions to address are:

- Which already existing datasets could be repurposed to train I+L?
- Which novel datasets specific for I+L generation should we start developing, and how?
- Can build broad L+I training datasets which can be fine-tuned to specific tasks and applications?

To create new datasets or help the systems use existing ones in the best way we need to develop a *theory* of how I+L should be done. Possible questions to answer are:

- **Use cases** Which applications could profit the most from the integration of language and vision, and which aspects of NLP should stick to words alone? An obvious application

[6] https://paperswithcode.com/dataset/ted-talks
[7] https://paperswithcode.com/dataset/ted-lium-3
[8] https://paperswithcode.com/dataset/ted-gesture-dataset
[9] https://phenaki.video/

is the generation of *fiction*: creating a mix of images and written narrative from a story line, or turning a text-only story into a multimodal narration. We should of course ask how useful this task really is. As discussed in the context of movies-from-books [30], depicting narrative can actually stifle mental imagery, standardize places and characters, and ultimately reduce the appreciation of the story. However, if the generation system is flexible enough to give the human story creator control over the way the narration is depicted or the ability to customize the scenes and the look of the characters, the L+I narration shifts in the direction of film-making, a respectable creative activity. A more radical option is to give the *viewer* the possibility of customizing the imagery (changing style, light, possibly background). This stretches the boundary between creation and fruition, but also between images and textual narrative (if I can change the illustrations, why not the story?). Scene customization is in fact a natural extension of the common practice of letting video-game players choose the likeness of their avatars, though here the choice is done by presenting compositional alternatives to choose from ("Do you want *this* avatar? *This* style of hair"). Still, avatars represent the self — in the case of stories generated via a L+I system, it's much more likely that many viewers would accept passive fruition. Ultimately, the trade off between the complexity of a system that allows broad customization and the number of user interested in using it should be carefully considered.

Storytelling is just only one of many tasks that generation systems could tackle. At least initially, a more likely candidate for practical I+L systems is the creation of how-to instructional videos (e.g. "How to replace the battery of your phone."), descriptions of possible activities (e.g. a mountain walk created by combining 3D map imagery with a commentary on difficult steps or things to see; a road trip guide that shows various possible stops and their attractions), rendering of events for which we have only textual descriptions, renderings of complex objects (e.g. the 3D view of a device extracted from a patent description) and probably many other non-obvious applications that will emerge as the systems mature.

- **The focus of interactive communication.** Where would the dialogic dimension add the most value, compared to a static I+L system? Currently, the domain that best combines NLP and image generation is *image editing*, where the focus is the image itself. Current T2I models create images which may be far from their creator's expectations; the subfield of Multimodal Image Synthesis and Editing (MISE, see [31] for a recent review) deals with the best approach to direct the generation process or edit the result. While many systems use input different from text in the editing stage, high-level linguistic instructions could be a powerful tool to refine them (e.g., in Fig. 2: "Make the teddy bear smaller and move it to the center of the zebras."). The system could answer by modifying the image (see Sec. 2), but it is easy to see the usefulness of interaction, in clarifications ("Should the skate also be smaller? Should I keep the bear on focus?") or explanations (User: "What is the object on the right?"; System: "A shop."; "Can you show a specific type of shop?", "Which one?").

  Despite the importance of MISE, an important yet undeveloped use case remains that of a system that provides textual feedback to explain things *beyond* what an image shows: its context, or any aspect that cannot be cast visually (see e.g. (1) below). This contrasts with

standard image captioning tasks, where the text describes what the image does show. The I+L problem is how text and image can be used *together* to provide information.

- **How to deal with cases where language and images interact?** A case in point is *anaphoric reference* from text to the object in the image. Should for instance objects in a scene flash or light up when mentioned by the system or by the user? In a modality where the system sees the user as he or she looks at a screen, how should *pointing* be detected and interpreted? (*What is this? Can you show the other side of this?* [pointing to a detail in the image]).
- **Image-Language Ratio.** When contents are rendered as I+L, what is the ideal images/language ratio? This question has been debated for a long time (see [3]), but it acquires new depths with the range of possibilities offered by generative AI. Human (off line) examples go from illustrated novels (text, for the most part) to the intertitles that occasionally interrupt silent movies (images, for the most part). The ratio could be a user choice (see [4]), but varying it would forces the system to analyze and structure the contents differently. Moreover, should the text be voiced, written with/before/after the image, or both? A theoretical analysis by Stenning and Oberlander [32] suggests that images, being less abstract, can in many cases facilitate processing, but their conclusions do not keep into account the cognitive cost of images that are too or not enough specific, or consistent.
- **Level of specificity** In many cases, an I+L system could make use of a mix of realistic and computer-generated materials. How can we clarify to the viewer/listener when an image tries to give an accurate rendition of a particular scene or object, and when it is a generic representation, or a mere filler? (In the description of the crash, the picture of the car is specific, the lamp post it hits, made up).
- **Consistency** The images produced by the system should remain consistent throughout the discourse, on pain of confusing the user. For instance, a story about a dog of unspecified breed should prompt the AI to draw a particular type of dog, but the dog must then remain *of that breed* in any subsequent image. There is a trade off between the dialogic dimension and consistency: information that becomes available in the development of the dialogue might contradict a random design choice made by the system early on (for instance, the system might have drawn the mention of a dog as a German shepherd, to later discover it was supposed to be a chihuahua). If the system has the possibility of processing the information from whole text beforehand this pitfall can be avoided, but the range of use cases may be reduced. Similarly, the system should avoid unwanted visual transformation as it transitions from scene to scene; switching to a mirror-image of the previous scene can be confusing, and so can changing the visual style and the appearence of the characters. Yet, variation is the norm in current models, and current solutions around this problem resolve in bags of tricks[10].

Lastly, there are technical aspects related to the *input*. Among them:

- What kind of input should a I+L generation system receive? Current image generators create images from text, with a wide margin of variation (Fig. 1), due to the underspec-

[10] https://mythicalai.substack.com/p/how-to-generate-consistent-characters

Figure 1: Different realizations of "a large ceramic mug with a left-protruding round handle and the drawing of a Japanese wave on the side." Source: Stable Diffusion

ification of textual information and to the non-deterministic nature of the generation algorithms (e.g. diffusion models). Images with a large part made up would be appropriate when the I+L system is meant to provide a *decoration* for the text, not when an image must be a faithful rendering of a description which is grasped more easily in graphic than in textual form. In this case the input should include images of the specific object that needs to be represented (see Sec. 4, esp. [33])

- A textual input to the image generation system needs not be normal human language. It could be language augmented with extremely detailed visual descriptions (on the model of the text prompts[11] used for image generation). But how can this type of input be provided for training? What is the best image+language *input* combination?

The answer to these questions is of course conditioned on what is already (potentially) available on the training side.

4. Dataset-building strategies

It would of course be relatively easy to use existing T2I system to associate images to the text generated by an automatic system (think of this as a DALL-E-over-GPT3-output approach). In principle, the process could be carried out at different levels of granularity. At one end, we could imagine generating images for all noun phrases in the text above a given threshold of imageability [34][12], then selecting some to be presented alongside the text. At the other end of the spectrum, one could pass each generated sentence to the T2I generator, then filter the

[11] https://neuroflash.com/blog/ai-image-generation-prompt-examples/

[12] It does not make much sense to try and generate images of individual nouns, as opposed to noun phrases. The ability to select *existing* images matching a conversation was demoed by Amazon in the re:MARS 2022 keynote.

Figure 2: A pensive Teddy. Based on DALL-E image ©OpenAI

output in some way. Neither of these extremes are likely to be very useful. What is needed instead is a *splitter* that decides which aspects of the input text should be cast in graphic form and which ones as text, starting from a required text/image ratio that will be low for, say, young children literature and high for, e.g., philosophy talks. For illustration, consider (1).

(1) a. A teddy bear is standing on a skateboard in Times Square, **thinking on how lonely it appeared during COVID peaks.**
b. A teddy bear is standing on a skateboard in Times Square, **because it doesn't have a wallet to pay the cab.**

A graphic/textual splitter should presumably decide that the part in **bold** should be best express by language, while the rest can be rendered as an image. Broadly speaking, when the modifier is not imageable ("a **sinful/nostalgic** teddy bear") or its relation with the rest is not transparent ("Teddy is in Times Square **because/when it wants/needs to see people**"), it is likely to be best rendered as text.[13] Modifiers that embed disjunction or negation ("The teddy bear **or stuffed dog** that **didn't come yesterday** reached Times Square **instead of its brother**, **despite having only a skateboard**") are essentially impossible to render as images alone.

In other cases, the linguistic relation could be easily rendered by segmenting the text in multiple images: *before(/after)* in (2) suggest ordering Fig. 2 *before* (respectively, *after*) Fig. 3. *Or* in (2-b) suggests a composition of Fig. 2 and 3, and so forth.

(2) a. A teddy bear is standing on a skateboard in Times Square, **before/after** playing chemistry with a friend.
b. A teddy bear is thinking of skateboarding in Times Square **or** playing chemistry with a friend.

[13] This is not to say that it is impossible to add to the image elements that suggest nostalgia, or a specific reason. There is a ratio between making the information unambiguous and cluttering the image that must be carefully balanced.

Figure 3: Teddy bears doing chemistry. Source: DALL-E

As mentioned above, in the representation of a narrative the text needs to be segmented in a sequence of images, and any image must be generated from the appropriate text segment. But the image generation must be designed to preserve elements that should not vary from scene to scene (the drawing style, the aspect and size of previously introduced objects). Solving image consistency through time is a must for video generation, so it is a topic likely to receive a lot of attention (recent attempts, like [35], start from stills and learn movements from untagged videos), but the problem of connecting the text to the individual entities or events shown in an artificially generated image is specific to I+L. The techniques proposed in the Deep Captioning (see Sec. 2) and the Image Paragraph literature are relevant for this point: a very detailed and structurally explicit textual description would allows reference to the individual objects introduced in the image, and make possible editing instructions like “Remove the skateboard from Fig. 2”. It could also be used to avoid repeating in the text elements that are obvious from the image.

Fig. 4 illustrates a potential training path: a detailed textual description (b) is obtained from an original image (a) via a pretrained I2T, then used by a pretrained T2I system to generate a derived image (c) (at low-resolution, to keep the computation manageable; upscaling techniques could be used for the final output, as in [20], [33]). The I2T phase could now be fine-tuned using as error function the T-T distance between the description of the original image (b) and that of the derived image (d). Alternatively (or concurrently), one could try to minimize a distance metric (I-I) between the original image (a) and the derived one (d). This is obviously more difficult, as the same textual description can generate very different images, but "different" does not necessarily mean "incompatible": two pictures of the same object from different angles can be graphically very different, yet serve the same purpose and be recognized as portraying “the same” content. On the other hand, a nearly-identical framing of the same objects in different styles can be disconcerting.

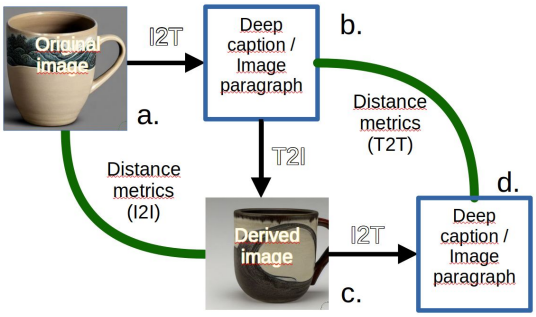

Figure 4: Pipeline schema for exhaustive image descriptions. Red lines are the distances to minimize via I2T and T2I fine-tuning.

Ruiz at al. [33] and Gal et al. [36] describe systems to generate the same object in different context, by associating a small set of specific images to a unique token identifier (for instance, in [33], "[S]" is a unique identifier and the T2I is fine-tuned to associate "The [S] dog" to a particular dog image, while preserving the association of "the dog" to a more diverse set of dog images). The problem is how to do this starting from a single image, one which might belong to the ground truth (the car that crashed in the narration of an actual accident), but also be a decoration that the system has allucinated to fill up a realistic image (the ambulance, the policemen). If the object has a familiar shape, one strategy might be to derived a set of 3D views from it (see the technique in [37], but without its very narrow domain) and use these views to condition the T2I. One obvious limit of this approach is that it would require repeated fine-tuning of large T2I systems, as new objects appear and are picked up in the narration, so it would be too slow for on-line deployment.

It is interesting to consider here the differences and similarities between humans, formal representations of meaning and distribution-based representations, like all NN models. Humans have no problem with the idea of persisting entities, be they physical objects or events, but are also very good at forming general concepts or classes. Formal semantics, on the other hand, has tackled the behavior of general classes of objects (*types*, or *kinds* in the terminology of [38] or [39]) only after dealing with (quantified) *token* objects, and arguably with greater difficulty. Distributional semantics has often been said to be more successful at capturing *generic* knowledge (i.e. types; see [40]) rather than episodic events, or tokens [41]. The same seems to apply in generation: T2I systems have no trouble generating individual tokens, but struggle to generate the *same* token in a changing environment. This time, the problem is shared by humans: a well-known difficulty for artists who start drawing comics is how to make each character *recognizable* in different scenes.

5. Some linguistic issues

Even the ability to neatly split what *can* be rendered as images and to render it consistently across scenes is no guarantee that we are doing a service to communication, or attention. There is a vast cognitive psychology literature on the relation between language and vision (see [42]

for a particular linguistic strand, and [43]), on how images affect attention and how languages could shape perception. This literature should inform some of the issues touched above: what is the best proportion of language and images for any given style (technical instruction, question answering, reports, storytelling, etc.), and whether the text should be written as subtitles, voiced over, voiced between the images or suspended as a balloon, comics-syle.[14]

One important aspect is the possibility of training the I+L system to render *graphically* features that are normally rendered by language alone. Consider for instance *repeated entity reference*. Language is fond of introducing entity in the discourse, then go back to them using pronouns or definite description; interpreting *chains of reference* is a complex issue in language comprehension and production. A mini-discourse like (3) could be easily translated into a correct sequence of pictures, but only if the referents of the boldfaced nouns are clear to the system and can be made recognizable by the user.

(3) **Two teddy bears** are at a picnic table. **One** decides to chase a butterfly. **The second one** runs after **him**. **A third teddy bear** occupies the picnic table left empty by **them**.

Suppose that the initial entities, two teddy bears, are first introduced graphically. Referring back to them in the subsequent images may be done verbally, using a normal pronoun (*they*) or a definite description (*the teddy bears*). In a I+L system, however, these linguistic devices could also be supplemented with arrows or bounding boxes, by making the referent flash or unfocus the background — useful signals if we are dealing with many objects of the same type, or if it is unclear how to designate an object that has only been introduced graphically ("the *thing* on the left").

This is of course not the only way to put graphics to linguistic use. Another possibility is to render the distinction between the true information that the I+L system is trying to convey and any additional material. Consider for instance (4)

(4) Students were protesting against a judge.

Here *a judge* may be a very specific one (part of the ground truth: the face might be recognizable, for instance); the bare plural *students* is non-specific, so the system would probably draw a random assortment of persons. But the system would also be entitled to draw a background (the *place* where the students were protesting, maybe some journalists). These additions are reasonable but entirely made up, and not in the prompt. Do we want to distinguish *graphically* what now looks like a three-way distinction in determination? (specific type and token: the judge; specific type, not token: the students; non specific type or token: everything else). If the answer is positive, the distinction must be carried out consistently in all subsequent scenes. Once again, we do not have a dataset of generated images where objects that were explicitly mentioned in the prompt can be visually distinguished from object that have been added by the generation process, but creating it does not seem to be an impossible task. Going back to the loop in Fig. 4, it should be possible to start from a prompt, generate an image from it (with T2I), then caption it (with I2T). The goal is now to have a captioning system that separates what it guesses to have been in the prompt from what has been added (a *discriminator*). With this in

[14] It is worth noting that voice-over narration is not very popular is modern cinematography, possibly on the account that it breaks the narrative identification. Still, narrative identification may not be a desideratum in many uses of I+L dialogue.

place, the T2I system could be fine-tuned to introduce changes in the image which keep the image recognizable but increase the effectiveness of the discriminator. The desired result would be images where aspects or object that have been hallucinated are recognizable to the naked eye.

6. Conclusions

This piece is a call to arms for a new AI research task, the generation of mixed languages and images in dialogue, which is starting to become possible due to advances in neighboring tasks: text-to-text, image-to-image, text-to-image, image-to-text (in captioning), text+image-to-image (in "guided" language generation), text+image-to-text (in VQA). The missing element, text(+image)-to-text+image, is not absent by chance. While potentially useful (we do use images in public presentations, we do have illustrated books and comics), it is a combination that our species cannot produce at "on-line" speed: a speed sufficient for dialogue. The consequence is that none of the many datasets accumulated by AI researchers can be directly used to train it, though some (instructional videos, comics, film+storyline) could be the base to create useful data. I have suggested that the output of text-to-text generation system (or more image-ready language variants) could be pre-processed in a way that separates those parts that can be converted into images or image sequences and those that need to remain voice or text. This could be carried out, in part, using hand-picked linguistic features, but could maybe also be derived from human-annotated data. The mix of language and images that would result raises a number of difficult but interesting linguistic and psychological questions — in part as a result of entering the territory of superhuman abilities.

7. Acknowledgments

The authors wishes to acknowledge Raffaella Bernardi, Paolo Rota and other CIMeC colleagues for stimulating conversations, and the audience of the EviL seminar series plus two anonymous TACL reviewers for feedback on early versions of these ideas.

References

[1] E. Bassignana, D. Brunato, M. Polignano, A. Ramponi, Preface to the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2023), 2023.

[2] S. F. Roth, J. Mattis, X. Mesnar, Graphics and natural language as components of automatic explanation, SIGCHI Bull. 20 (1988) 76. URL: https://doi.org/10.1145/49103.1046410. doi:10.1145/49103.1046410.

[3] M. Theune, I. van der Sluis, Y. Bachvarova, E. André, The AISB'08 Symposium on Multimodal Output Generation (MOG 2008), The Society for the Study of Artificial Intelligence

and Simulation of Behaviour (AISB), 2008, pp. iii–iv. AISB Symposium on Multimodal Output Generation, MOG 2008, MOG ; Conference date: 03-04-2008 Through 04-04-2008.

[4] G. Kahl, R. Wasinger, T. Schwartz, L. Spassova, Three output planning strategies for use in context-aware computing scenarios (2008) 46–49. AISB Symposium on Multimodal Output Generation, MOG 2008, MOG ; Conference date: 03-04-2008 Through 04-04-2008.

[5] V. Rieser, O. Lemon, Learning effective multimodal dialogue strategies from Wizard-of-Oz data: Bootstrapping and evaluation, in: Proceedings of ACL-08: HLT, Association for Computational Linguistics, Columbus, Ohio, 2008, pp. 638–646. URL: https://aclanthology.org/P08-1073.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, arXiv 2005.14165, 2020. arXiv:2005.14165.

[7] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, , N. Fiedel., Palm: Scaling language modeling with pathways., 2022. ArXiv:2001.08361 and 2022.

[8] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, Opt: Open pre-trained transformer language models, 2022. URL: https://arxiv.org/abs/2205.01068. doi:10.48550/ARXIV.2205.01068.

[9] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen, GLIDE: towards photorealistic image generation and editing with text-guided diffusion models, CoRR abs/2112.10741 (2021). URL: https://arxiv.org/abs/2112.10741. arXiv:2112.10741.

[10] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, 2022. URL: https://arxiv.org/abs/2204.06125. doi:10.48550/ARXIV.2204.06125.

[11] X. Lu, L. Ng, J. Fernandez, H. Zhu, CIGLI: conditional image generation from language & image, CoRR abs/2108.08955 (2021). URL: https://arxiv.org/abs/2108.08955. arXiv:2108.08955.

[12] Y. Zhou, R. Zhang, J. Gu, C. Tensmeyer, T. Yu, C. Chen, J. Xu, T. Sun, Interactive image generation with natural-language feedback, PREPRINT: https://www.aaai.org/AAAI22Papers/AAAI-7081.ZhouY.pdf, 2022.

[13] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, M. Irani, Imagic: Text-based real image editing with diffusion models, 2023. arXiv:2210.09276.

[14] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, E. Xing,

Multimodal image synthesis and editing: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 1–20. doi:10.1109/TPAMI.2023.3305243.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.

[16] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, Microsoft coco: Common objects in context, 2014. URL: https://arxiv.org/abs/1405.0312. doi:10.48550/ARXIV.1405.0312.

[17] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. doi:10.48550/ARXIV.2103.00020.

[18] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, K. Keutzer, How much can clip benefit vision-and-language tasks?, 2021. URL: https://arxiv.org/abs/2107.06383. doi:10.48550/ARXIV.2107.06383.

[19] T. Ghandi, H. Pourreza, H. Mahyar, Deep learning approaches on image captioning: A review, 2022. URL: https://arxiv.org/abs/2201.12944. doi:10.48550/ARXIV.2201.12944.

[20] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL: https://arxiv.org/abs/2205.11487.

[21] J. Johnson, A. Karpathy, L. Fei-Fei, Densecap: Fully convolutional localization networks for dense captioning, 2015. URL: https://arxiv.org/abs/1511.07571. doi:10.48550/ARXIV.1511.07571.

[22] J. Krause, J. Johnson, R. Krishna, L. Fei-Fei, A hierarchical approach for generating descriptive image paragraphs, in: Computer Vision and Patterm Recognition (CVPR), 2017.

[23] D. L. Fernandes., M. H. F. Ribeiro., F. R. Cerqueira., M. M. Silva., Describing image focused in cognitive and visual details for visually impaired people: An approach to generating inclusive paragraphs, in: Proceedings of the 17th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,, INSTICC, SciTePress, 2022, pp. 526–534. doi:10.5220/0010845700003124.

[24] Z. Bian, H. Liu, B. Wang, H. Huang, Y. Li, C. Wang, F. Cui, Y. You, Colossal-ai: A unified deep learning system for large-scale parallel training, 2021. URL: https://arxiv.org/abs/2110.14883. doi:10.48550/ARXIV.2110.14883.

[25] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, VQA: visual question answering, CoRR abs/1505.00468 (2015). URL: http://arxiv.org/abs/1505.00468. arXiv:1505.00468.

[26] S. Pezzelle, C. Greco, G. Gandolfi, E. Gualdoni, R. Bernardi, Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 2751–2767. URL: https://aclanthology.org/2020.findings-emnlp.248. doi:10.18653/v1/2020.findings-emnlp.248.

[27] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis, CoRR abs/2003.08934 (2020). URL: https://arxiv.org/abs/2003.08934. arXiv:2003.08934.

[28] S. Kar, S. Maharjan, A. P. López-Monroy, T. Solorio, MPST: A corpus of movie plot synopses with tags, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1274.

[29] R. Villegas, M. Babaeizadeh, P.-J. Kindermans, H. Moraldo, H. Zhang, M. T. Saffar, S. Castro, J. Kunze, D. Erhan, Phenaki: Variable length video generation from open domain textual descriptions, in: International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=vOEXS39nOF.

[30] T. Leitch, Film Adaptation and Its Discontents: From Gone with the Wind to The Passion of the Christ. 2007. „ Project MUSE, Johns Hopkins University Press, 2007. doi:10.1353/book.3302.

[31] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, E. Xing, Multimodal image synthesis and editing: The generative ai era, 2023. arXiv:2112.13592.

[32] K. Stenning, J. Oberlander, A cognitive theory of graphical and linguistic reasoning: Logic and implementation, Cognitive Science 19 (1995) 97–140. doi:https://doi.org/10.1207/s15516709cog1901_3.

[33] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2022. URL: https://arxiv.org/abs/2208.12242. doi:10.48550/ARXIV.2208.12242.

[34] N. Ljubešić, D. Fišer, A. Peti-Stantić, Predicting concreteness and imageability of words within and across languages via word embeddings, in: Proceedings of The Third Workshop on Representation Learning for NLP, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 217–222. URL: https://aclanthology.org/W18-3028. doi:10.18653/v1/W18-3028.

[35] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, Y. Taigman, Make-a-video: Text-to-video generation without text-video data, 2022. URL: https://arxiv.org/abs/2209.14792. doi:10.48550/ARXIV.2209.14792.

[36] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, D. Cohen-Or, An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL: https://arxiv.org/abs/2208.01618. doi:10.48550/ARXIV.2208.01618.

[37] N. Müller, A. Simonelli, L. Porzi, S. R. Bulò, M. Nießner, P. Kontschieder, Autorf: Learning 3d object radiance fields from single view observations, 2022. URL: https://arxiv.org/abs/2204.03593. doi:10.48550/ARXIV.2204.03593.

[38] J. Lawler, Studies in English Generics, University of Michigan Papers in Linguistics 1 (1973).

[39] G. Carlson, A unified analysis of the English bare plural, Linguistics and Philosophy 1 (1977) 413–457.

[40] L. McNally, Kinds, descriptions of kinds, concepts, and distributions, 2015. Submitted to a volume of papers from the BRIDGE-14 Workshop.

[41] A. Herbelot, E. M. Vecchi, Building a shared world: mapping distributional to model-theoretic semantic spaces, in: Proceedings of the 2015 Conference on Empirical Methods in

Natural Language Processing, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 22–32. URL: https://aclanthology.org/D15-1003. doi:10.18653/v1/D15-1003.
[42] R. Jackendoff, On beyond zebra, Cognition 26 (1987) 89–114.
[43] M. Vulchanova, V. Vulchanov, I. Fritz, E. A. Milburn, Language and perception: Introduction to the special issue "speakers and listeners in the visual world"., J. Cult Cogn Sci. 3 (2019) 103–112. doi:https://doi.org/10.1007/s41809-019-00047-z.
[44] J. Hankamer, I. Sag, Deep and surface anaphora, Linguistic Inquiry 7.3 (1976) 391–426.
[45] L. A. Friedman, Space, time, and person reference in american sign language., Language 51 (1975) 940–61. doi:https://doi.org/10.2307/412702.
[46] K. van Hoek, Conceptual spaces and pronominal reference in american sign language, Nordic Journal of Linguistics 15 (1992) 183–199. doi:10.1017/S0332586500002596.