Look, Read and Enrich - Learning from Scientific Figures and their Captions

Abstract

Compared to natural images, understanding scientific figures is particularly hard for machines. However, there is a valuable source of information in scientific literature that until now has remained untapped: the correspondence between a figure and its caption. In this paper we investigate what can be learnt by looking at a large number of figures and reading their captions, and introduce a figure-caption correspondence learning task that makes use of our observations. Training visual and language networks without supervision other than pairs of unconstrained figures and captions is shown to successfully solve this task. We also show that transferring lexical and semantic knowledge from a knowledge graph significantly enriches the resulting features. Finally, we demonstrate the positive impact of such features in other tasks involving scientific text and figures, like multi-modal classification and machine comprehension for question answering, outperforming supervised baselines and ad-hoc approaches.

Introduction

Scientific knowledge is heterogeneous and can present itself in many forms, including text, mathematical equations, figures and tables. Like many other manifestations of human thought, the scientific discourse usually adopts the form of a narrative, a scientific publication where related knowledge is presented in mutually supportive ways over different modalities. In the case of scientific figures, like charts, images and diagrams, these are usually accompanied by a text paragraph, a caption, that elaborates on the analysis otherwise visually represented.

In this paper, we make use of this observation and tap on the potential of learning from the enormous source of free supervision available in the scientific literature, with millions of figures and their captions.

We build models that learn from the scientific discourse both visually and textually by simply looking at the figures and reading their explanatory captions, inspired in how humans learn by reading a scientific publication. To this purpose, we explore how multi-modal scientific knowledge can be learnt from the correspondence between figures and captions.

The main contributions of this paper are the following:

An unsupervised Figure-Caption Correspondence task (FCC) that jointly learns text and visual features useful to address a range of tasks involving scientific text and figures.

A method to enrich such features with semantic knowledge transferred from structured knowledge graphs (KG).

A study of the complexity of figure-caption correspondence compared to classical image-sentence matching.

A qualitative and quantitative analysis of the learnt text and visual features through transfer learning tasks.

A corpus of scientific figures and captions extracted from SN SciGraph and AI2 Semantic Scholar.

We present the FCC task in section SECREF3, including the network architecture, training protocol, and how adding pre-trained word and semantic embeddings can enrich the resulting text and visual features. In section SECREF4, we first introduce our datasets and evaluate the performance of our method in the task it was trained to solve, the correspondence between scientific figures and captions. Then, we relate our work to the state of the art in image-sentence matching and evaluate our approach in two challenging

transfer learning tasks: caption and figure classification and multi-modal machine comprehension. In section SECREF5 we perform a qualitative study that illustrates how the FCC task leads to detailed textual and visual discrimination. Finally, in section SECREF6 we conclude the paper and advance future work.

Related work

Understanding natural images has been a major area of research in computer vision, with well established datasets like ImageNet BIBREF0, Flickr8K BIBREF1, Flickr30K BIBREF2 and COCO BIBREF3. However, reasoning with other visual representations like scientific figures and diagrams has not received the same attention yet and entails additional challenges: Scientific figures are more abstract and symbolic, their captions tend to be significantly longer and use specialized lexicon, and the relation between a scientific figure and its caption is unique, i.e. in a scientific publication there is only one caption that corresponds with one figure and vice versa.

The FCC task presented herein is a form of co-training BIBREF4 where there are two views of the data and each view provides complementary information. Similar two-branch neural architectures focus on image-sentence BIBREF5, BIBREF6 and audio-video BIBREF7 matching. Others like BIBREF8 learn common embeddings from images and text. However, in such cases one or both networks are typically pre-trained.

Focused on geometry, BIBREF9 maximize the agreement between text and visual data. In BIBREF10, the authors apply machine vision and natural language processing to extract data from figures and their associated text in bio-curation tasks. In BIBREF11, they parse diagram components and connectors as a Diagram Parse Graph (DPG), semantically interpret the DPG and use the model to answer diagram questions. While we rely on the correspondence between figures and captions, they train a specific

classifier for each component and connector type and yet another model to ground the semantics of the DPG in each domain, like food webs or water cycles.

Knowledge fusion approaches like BIBREF12 investigate the potential of complementing KG embeddings with text and natural images by integrating information across the three modalities in a single latent representation. They assume pre-trained entity representations exist in each individual modality, e.g. the visual features encoding the image of a ball, the word embeddings associated to the token "ball", and the KG embeddings related to the ball entity, which are then stitched together. In contrast, FCC co-trains text and visual features from figures and their captions and supports the enrichment of such features with lexical and semantic knowledge transferred from a KG during the training of the FCC task.

Figure-Caption Correspondence

The main idea of our approach is to learn a correspondence task between scientific figures and their captions as they appear in a scientific publication. The information captured in the caption explains the corresponding figure in natural language, providing guidance to identify the key features of the figure and vice versa. By seeing a figure and reading the textual description in its caption we ultimately aim to learn representations that capture e.g. what it means that two plots are similar or what gravity looks like.

We leverage this observation to learn a figure-caption correspondence task. In essence, FCC is a binary classification task that receives a figure and a caption and determines whether they correspond or not. For training, the positive pairs are actual figures and their captions from a collection of scientific publications. Negative pairs are extracted from combinations of figures and any other randomly selected captions. The network is then made to learn text and visual features from scratch, without additional labelled data.

Figure-Caption Correspondence ::: FCC Architecture and Model

We propose a 2-branch neural architecture (figure FIGREF7) that has three main parts: the vision and language subnetworks, respectively extracting visual and text features, and a fusion subnetwork that takes the resulting features from the visual and text blocks and uses them to evaluate figure-caption correspondence.

The vision subnetwork follows a VGG-style BIBREF13 design, with 3x3 convolutional filters, 2x2 max-pooling layers with stride 2 and no padding. It contains 4 blocks of conv+conv+pool layers, where inside each block the two convolutional layers have the same number of filters, while consecutive blocks have doubling number of filters (64, 128, 256, 512). The input layer receives 224x224x3 images. The final layer produces a 512-D vector after 28x28 max-pooling. Each convolutional layer is followed by batch normalization BIBREF14 and ReLU layers. Based on BIBREF15, the language subnetwork has 3 convolutional blocks, each with 512 filters and a 5-element window size with ReLU activation. Each convolutional layer is followed by a 5-max pooling layer, except for the final layer, which produces a 512-D vector after 35-max pooling. The language subnetwork has a 300-D embeddings layer at the input, with a maximum sequence length of 1,000 tokens. The fusion subnetwork calculates the element-wise product of the 512-D visual and text feature vectors into a single vector r to produce a 2-way classification output (correspond or not). It has two fully connected layers, with ReLU and an intermediate feature size of 128-D. The probability of each choice is the softmax of r, i.e. $\hat{y} = softmax(r) \in \mathbb {R}^{2}$. During training, we minimize the negative log probability of the correct choice.

This architecture enables the FCC task to learn visual and text features from scratch in a completely unsupervised manner, just by observing the correspondence of figures and captions. Next, we extend it to enable the transfer of additional pre-trained information. Here, we focus on adding pre-trained embeddings on the language branch, and then back-propagate to the visual features during FCC training.

Adding pre-trained visual features is also possible and indeed we also evaluate its impact in the FCC task in section SECREF14.

Let V be a vocabulary of words from a collection of documents D. Also, let L be their lemmas, i.e. base forms without morphological or conjugational variations, and C the concepts (or senses) in a KG. Each word w_k in V, e.g. made, has one lemma l_k (make) and may be linked to one or more concepts c_k in C (create or produce something).

For each word w_k, the FCC task learns a d-D embedding $\vec{w}_k$, which can be combined with pre-trained word ($\vec{w^{\prime }}_k$), lemma ($\vec{l}_k$) and concept ($\vec{c}_k$) embeddings to produce a single vector $\vec{t}_k$. If no pre-trained knowledge is transferred from an external source, then $\vec{t}_k=\vec{w}_k$. Note that we previously lemmatize and disambiguate D against the KG in order to select the right pre-trained lemma and concept embeddings for each particular occurrence of w_k. Equation DISPLAY_FORM8 shows the different combinations of learnt and pre-trained embeddings we consider: (a) learnt word embeddings only, (b) learnt and pre-trained word embeddings and (c) learnt word embeddings and pre-trained semantic embeddings, including both lemmas and concepts, in line with our recent findings presented in BIBREF16.

In our experiments, concatenation proved optimal to combine the embeddings learnt by the network and the pre-trained embeddings, compared to other methods like summation, multiplication, average or learning a task-specific weighting of the different representations as in BIBREF17. Since some words may not have associated pre-trained word, lemma or concept embeddings, we pad these sequences with $\varnothing _W$, $\varnothing _L$ and $\varnothing _C$, which are never included in the vocabulary. The dimensionality of $\vec{t}_k$ is fixed to 300, i.e. the size of each sub-vector in configurations (a), (b) and (c) is 300, 150 and 100, respectively. In doing so, we aimed at limiting the number of trainable parameters and balance the contribution of each information source.

In its most basic form, i.e. configuration (a), the FCC network has over 32M trainable parameters (28M in the language subnetwork, 4M in the vision subnetwork and 135K in the fusion subnetwork) and takes 12 hours to train on a single GPU Nvidia GeForce RTX 2080 Ti for a relatively small corpus (SN SciGraph, see section SECREF12). We used 10-fold cross validation, Adam optimization BIBREF18 with learning rate 10^{-4} and weight decay 10^{-5}. The network was implemented in Keras and TensorFlow, with batch size 32. The number of positive and negative cases is balanced within the batches.

Figure-Caption Correspondence ::: Semantic Embeddings

We use HolE BIBREF19 and Vecsigrafo BIBREF16 to learn semantic embeddings. The latter extends the Swivel algorithm BIBREF20 to jointly learn word, lemma and concept embeddings on a corpus disambiguated against the KG, outperforming the previous state of the art in word and word-sense embeddings by co-training word, lemma and concept embeddings as opposed to training each individually. In contrast to Vecsigrafo, which requires both a text corpus and a KG, HolE follows a graph-based approach where embeddings are learnt exclusively from the KG. As section SECREF14 will show, this gives Vecsigrafo a certain advantage in the FCC task. Following up with the work presented in BIBREF16, our experiments focus on Sensigrafo, the KG underlying Expert System's Cogito NLP proprietary platform. Similar to WordNet, on which Vecsigrafo has also been successfully trained, Sensigrafo is a general-purpose KG with lexical and semantic information that contains over 300K concepts, 400K lemmas and 80 types of relations rendering 3M links. We use Cogito to disambiguate the text corpora prior to training Vecsigrafo. All the semantic (lemma and concept) embeddings produced with HolE or Vecsigrafo are 100-D.

Results and Discussion

In this section, first we evaluate the actual FCC task against two supervised baselines. Then, we situate our work in the more general image-sentence matching problem, showing empirical evidence of the additional complexity associated to the scientific domain and the figure-caption case compared to natural images. Next, we test the visual and text features learnt in the FCC task in two different transfer learning settings: classification of scientific figures and captions and multi-modal machine comprehension for question answering given a context of text, figures and images.

Results and Discussion ::: Datasets

We have used the following datasets for training and evaluation:

The Semantic Scholar corpus BIBREF21 (SemScholar) is a large dataset of scientific publications made available by AI2. From its 39M articles, we downloaded 3,3M PDFs (the rest were behind paywalls, did not have a link or it was broken) and extracted 12.5M figures and captions through PDFFigures2 BIBREF22. We randomly selected 500K papers to train the FCC task on their figures and captions and another 500K to train Vecsigrafo on the text of their titles and abstracts.

Springer Nature's SciGraph contains 7M scientific publications organized in 22 scientific fields or categories. Since SciGraph does not provide a link to the PDF of the publication, we selected the intersection with SemScholar, producing a smaller corpus of 80K papers (in addition to the 1M papers from SemScholar mentioned above) and 82K figures that we used for training certain FCC configurations and supervised baselines (section SECREF14).

The Textbook Question Answering corpus BIBREF23 includes 1,076 lessons and 26,260 multi-modal test questions from middle school science curricula. Its complexity and scope make it a challenging textual and visual question answering dataset.

Wikipedia. We used the January 2018 English Wikipedia dataset as one of the corpora on which to train Vecsigrafo. As opposed to SciGraph or SemScholar, specific of the scientific domain, Wikipedia is a source of general-purpose information.

Flickr30K and COCO, as image-sentence matching benchmarks.

Results and Discussion ::: Figure-Caption Correspondence

We evaluate our method in the task it was trained to solve: determining whether a figure and a caption correspond. We also compare the performance of the FCC task against two supervised baselines, training them on a classification task against the SciGraph taxonomy. For such baselines we first train the vision and language networks independently and then combine them. The feature extraction parts of both networks are the same as described in section SECREF6. On top of them, we attach a fully connected layer with 128 neurons and ReLU activation and a softmax layer, with as many neurons as target classes.

The direct combination baseline computes the figure-caption correspondence through the scalar product between the softmax outputs of both networks. If it exceeds a threshold, which we heuristically fixed on 0.325, the result is positive. The supervised pre-training baseline freezes the weights of the feature extraction trunks from the two trained networks, assembles them in the FCC architecture as shown in section SECREF6, and trains the FCC task on the fully connected layers. While direct combination provides a notion of the agreement between the two branches, supervised pre-training is the most similar supervised approach to our method.

Table TABREF15 shows the results of the FCC task and the supervised baselines. FCC_k denotes the corpus and word representation used to train the FCC task. Acc$_{vgg}$ shows the accuracy after replacing our visual branch with pre-trained VGG16 features learnt on ImageNet. This provides an

estimate of how specific of the scientific domain scientific figures and therefore the resulting visual features can be, compared to natural images. As the table shows, the results obtained using pre-trained visual features are clearly worse in general (only slightly better in FCC_3), suggesting that the visual information contained in scientific figures indeed differs from natural images.

We trained the FCC network on two different scientific corpora: SciGraph (FCC_{1-5}) and SemScholar (FCC_{6-7}). Both FCC_1 and FCC_6 learnt their own word representations without transfer of any pre-trained knowledge. Even in its most basic form our approach substantially improves over the supervised baselines, confirming that the visual and language branches learn from each other and also that figure-caption correspondence is an effective source of free supervision.

Adding pre-trained knowledge at the input layer of the language subnetwork provides an additional boost, particularly with lemma and concept embeddings from Vecsigrafo (FCC_5). Vecsigrafo clearly outperformed HolE (FCC_3), which was also beaten by pre-trained fastText BIBREF24 word embeddings (FCC_2) trained on SemScholar.

Since graph-based KG embedding approaches like HolE only generate embeddings of the artifacts explicitly contained in the KG, this may indicate that Sensigrafo, the KG used in this task, provides a partial coverage of the scientific domain, as could be expected since we are using an off-the-shelf version. Deeper inspection shows that HolE only covers 20% of the lemmas in the SciGraph vocabulary. On the other hand, Vecsigrafo, trained on the same KG, also captures lexical information from the text corpora it is trained on, Wikipedia or SemScholar, raising lemma coverage to 42% and 47%, respectively.

Although the size of Wikipedia is almost triple of our SemScholar corpus, training Vecsigrafo on the latter resulted in better FCC accuracy (FCC_4 vs. FCC_5), suggesting that domain relevance is more significant than sheer volume, in line with our previous findings in BIBREF25. Training FCC on

SemScholar, much larger than SciGraph, further improves accuracy, as shown in FCC_6 and FCC_7.

Results and Discussion ::: Image-Sentence Matching

We put our FCC task in the context of the more general problem of image-sentence matching through a bidirectional retrieval task where images are sought given a text query and vice versa. While table TABREF20 focuses on natural images datasets (Flickr30K and COCO), table TABREF21 shows results on scientific datasets (SciGraph and SemScholar) rich in scientific figures and diagrams. The selected baselines (Embedding network, 2WayNet, VSE++ and DSVE-loc) report results obtained on the Flickr30K and COCO datasets, which we also include in table TABREF20. Performance is measured in recall at k (Rk), with k={1,5,10}. From the baselines, we successfully reproduced DSVE-loc, using the code made available by the authors, and trained it on SciGraph and SemScholar.

We trained the FCC task on all the datasets, both in a totally unsupervised way and with pre-trained semantic embeddings (indicated with subscript vec), and executed the bidirectional retrieval task using the resulting text and visual features. We also experimented with pre-trained VGG16 visual features extracted from ImageNet (subscript vgg), with more than 14 million hand-annotated images. Following common practice in image-sentence matching, our splits are 1,000 samples for test and the rest for training.

We can see a marked division between the results obtained on natural images datasets (table TABREF20) and those focused on scientific figures (table TABREF21). In the former case, VSE++ and DSVE-loc clearly beat all the other approaches. In contrast, our model performs poorly on such datasets although results are ameliorated when we use pre-trained visual features from ImageNet ("Oursvgg" and "Oursvgg-vec"). Interestingly, the situation reverts with the scientific datasets. While the recall of

DSVE-loc drops dramatically in SciGraph, and even more in SemScholar, our approach shows the opposite behavior in both figure and caption retrieval. Using visual features enriched with pre-trained semantic embeddings from Vecsigrafo during training of the FCC task further improves recall in the bidirectional retrieval task. Compared to natural images, the additional complexity of scientific figures and their caption texts, which in addition are considerably longer (see table TABREF19), seems to have a clear impact in this regard.

Unlike in Flickr30K and COCO, replacing the FCC visual features with pre-trained ones from ImageNet brings us little benefit in SciGraph and even less in SemScholar, where the combination of FCC and Vecsigrafo ("Oursvec") obtains the best results across the board. This and the extremely poor performance of the best image-sentence matching baseline (DSVE-loc) in the scientific datasets shows evidence that dealing with scientific figures is considerably more complex than natural images. Indeed, the best results in figure-caption correspondence ("Oursvec" in SemScholar) are still far from the SoA in image-sentence matching (DSVE-loc in COCO).

Results and Discussion ::: Caption and Figure Classification

We evaluate the language and visual representations emerging from FCC in the context of two classification tasks that aim to identify the scientific field an arbitrary text fragment (a caption) or a figure belong to, according to the SciGraph taxonomy. The latter is a particularly hard task due to the whimsical nature of the figures that appear in our corpus: figure and diagram layout is arbitrary; charts, e.g. bar and pie charts, are used to showcase data in any field from health to engineering; figures and natural images appear indistinctly, etc. Also, note that we only rely on the actual figure, not the text fragment where it is mentioned in the paper.

We pick the text and visual features that produced the best FCC results with and without pre-trained

semantic embeddings (table TABREF15, FCC_7 and FCC_6, respectively) and use the language and vision subnetworks presented in section SECREF6 to train our classifiers on SciGraph in two different scenarios. First, we only fine tune the fully connected and softmax layers, freezing the text and visual weights (non-trainable in the table). Second, we fine tune all the parameters in both networks (trainable). In both cases, we compare against a baseline using the same networks initialized with random weights, without FCC training. In doing so, through the first, non-trainable scenario, we seek to quantify the information contributed by the FCC features, while training from scratch on the target corpus should provide an upper bound for figure and caption classification. Additionally, for figure classification, we select a baseline of frozen VGG16 weights trained on ImageNet. We train using 10-fold cross validation and Adam. For the caption classification task, we select learning rate 10^{-3} and batch size 128. In figure classification, we use learning rate 10^{-4}, weight decay 10^{-5} and batch size 32.

The results in table TABREF23 show that our approach amply beats the baselines, including the upper bound (training from scratch on SciGraph). The delta is particularly noticeable in the non trainable case for both caption and figure classification and is considerably increased in "Ours FCC_7", which uses pre-trained semantic embeddings. This includes both the random and VGG baselines and illustrates again the additional complexity of analyzing scientific figures compared to natural images, even if the latter is trained on a considerably larger corpus like ImageNet. Fine tuning the whole networks on SciGraph further improves accuracies. In this case, "Ours FCC_6", which uses FCC features without additional pre-trained embeddings, slightly outperforms "Ours FCC_7", suggesting a larger margin to learn from the task-specific corpus. Note that both FCC_6 and FCC_7 were trained on SemScholar.

Results and Discussion ::: Textbook Question Answering (TQA) for Multi-Modal Machine Comprehension

We leverage the TQA dataset and the baselines in BIBREF23 to evaluate the features learnt by the FCC task in a multi-modal machine comprehension scenario. We study how our model, which was not

originally trained for this task, performs against state of the art models specifically trained for diagram question answering and textual reading comprehension in a very challenging dataset. We also study how pre-trained semantic embeddings impact in the TQA task: first, by enriching the visual features learnt in the FCC task as shown in section SECREF6 and then by using pre-trained semantic embeddings to enrich word representations in the TQA corpus.

We focus on multiple-choice questions, 73% of the dataset. Table TABREF24 shows the performance of our model against the results reported in BIBREF23 for five TQA baselines: random, BiDAF (focused on text machine comprehension), text only (TQA_1, based on MemoryNet), text+image (TQA_2, VQA), and text+diagrams (TQA_3, DSDP-NET). We successfully reproduced the TQA_1 and TQA_2 architectures and adapted the latter. Then, we replaced the visual features in TQA_2 with those learnt by the FCC visual subnetwork both in a completely unsupervised way (FCC_6 in table TABREF15) and with pre-trained semantic embeddings (FCC_7), resulting in TQA_4 and TQA_5, respectively.

While TQA_{1-5} used no pre-trained embeddings at all, TQA_{6-10} were trained including pre-trained Vecsigrafo semantic embeddings. Unlike FCC, where we used concatenation to combine pre-trained lemma and concept embeddings with the word embeddings learnt by the task, element-wise addition worked best in the case of TQA.

Following the recommendations in BIBREF23, we pre-processed the TQA corpus to i) consider knowledge from previous lessons in the textbook in addition to the lesson of the question at hand and ii) address challenges like long question contexts with a large lexicon. In both text and diagram MC, applying the Pareto principle to reduce the maximum token sequence length in the text of each question, their answers and context improved accuracy considerably. This optimization allowed reducing the amount of text to consider for each question, improving the signal to noise ratio. Finally, we obtained the most relevant paragraphs for each question through tf-idf and trained the models using 10-fold cross

validation, Adam, learning rate 10^{-2} and batch size 128. In text MC we also used 0.5 dropout and recurrent dropout in the LSTM layers.

Fitting multi-modal sources into a single memory, the use of visual FCC features clearly outperforms all the TQA baselines in diagram MC. Enhancing word representation with pre-trained semantic embeddings during training of the TQA task provides an additional boost that results in the highest accuracies for both text MC and diagram MC. These are significantly good results since, according to the TQA authors BIBREF23, most diagram questions in the TQA corpus would normally require a specific rich diagram parse, which we did not aim to provide.

Qualitative Analysis

We inspect the features learnt by our FCC task to gain a deeper understanding of the syntactic and semantic patterns captured for figure and caption representation. The findings reported herein are qualitatively consistent for all the FCC variations in table TABREF15.

Vision features. The analysis was carried out on an unconstrained variety of charts, diagrams and natural images from SciGraph, without filtering by figure type or scientific field. To obtain a representative sample of what the FCC network learns, we focus on the 512-D vector resulting from the last convolutional block before the fusion subnetwork. We pick the features with the most significant activation over the whole dataset and select the figures that activate them most. To this purpose, we prioritize those with higher maximum activation against the average activation.

Figure FIGREF27 shows a selection of 6 visual features with the 4 figures that activate each feature more significantly and their activation heatmaps. Only figures are used as input, no text. As can be seen, the vision subnetwork has automatically learnt, without explicit supervision, to recognize different types of

diagrams, charts and content, such as (from left to right) whisker plots, western blots (a technique used to identify proteins in a tissue sample), multi-image comparison diagrams, multi-modal data visualization charts (e.g. western plots vs. bar charts), line plots, and text within the figures. Furthermore, as shown by the heatmaps, our model discriminates the key elements associated to the figures that most activate each feature: the actual whiskers, the blots, the borders of each image under comparison, the blots and their complementary bar charts, as well as the line plots and the correspondence between them and the values in the x and y axes. Also, see (right-most column) how a feature discriminates text inserted in the figure, regardless of the remaining elements that may appear and the connections between them. This shows evidence of how the visual features learnt by the FCC task support the parsing of complex scientific diagrams.

We also estimated a notion of semantic specificity based on the concepts of a KG. For each visual feature, we aggregated the captions of the figures that most activate it and used Cogito to disambiguate the Sensigrafo concepts that appear in them. Then, we estimated how important each concept is to each feature by calculating its tf-idf. Finally, we averaged the resulting values to obtain a consolidated semantic specificity score per feature.

The scores of the features in figure FIGREF27 range between 0.42 and 0.65, which is consistently higher than average (0.4). This seems to indicate a correlation between activation and the semantic specificity of each visual feature. For example, the heatmaps of the figures related to the feature with the lowest tf-idf (left-most column) highlights a particular visual pattern, i.e. the whiskers, that may spread over many, possibly unrelated domains. On the other hand, the feature with the highest score (second column) focuses on a type of diagrams, western blots, almost exclusive of protein and genetic studies. Others, like the feature illustrated by the figures in the fifth column, capture the semantics of a specific type of 2D charts relating two magnitudes x and y. Analyzing their captions with Cogito, we see that concepts like e.g. isochronal and exponential functions are mentioned. If we look at the second and four top-most

figures in the column, we can see that such concepts are also visually depicted in the figures, suggesting that the FCC task has learnt to recognize them both from the text and visually.

Text features. Similar to the visual case, we selected the features from the last block of the language subnetwork with the highest activation. For visualization purposes, we picked the figures corresponding to the captions in SciGraph that most activate such features (figure FIGREF28). No visual information is used.

Several distinct patterns emerge from the text. The text feature in the first column seems to focus on genetics and histochemistry, including terms like western blots or immunostaining and variations like immunoblot-s/ted/ting. Interestingly, it also seems to have learnt some type of is-a relations (western blot is a type of immunoblot). The second feature focuses on variations of the term radiograph, e.g. radiograph-y/s. The third feature specializes in text related to curve plots involving several statistic analysis, e.g. Real-time PCR, one-way ANOVA or Gaussian distribution. Sometimes (fourth figure from top) the caption does not mention the plot directly, but focuses on the analysis instead, e.g. "the data presented here are mean values of duplicate experiments", indicating transfer of knowledge from the visual part during training. The fourth feature extracts citations and models named after prominent scientists, e.g. Evans function (first and fourth figure), Manley (1992) (second), and Aliev-Panfilov model (third). The fifth feature extracts chromatography terminology, e.g. 3D surface plot, photomicrograph or color map and, finally, the right-most feature focuses on different types of named diagrams, like flow charts and state diagrams, e.g. phylogenetic trees.

All the captions show a strong semantic correspondence with their associated figures. Figure FIGREF29 shows the activation heatmaps for two sample captions, calculated on the embeddings layer of the language subnetwork. The upper one corresponds to the fourth column left-right and third figure top-down in figure FIGREF28. Its caption reads: "The Aliev-Panfilov model with $\alpha =0.01$...The phase portrait

depicts trajectories for distinct initial values $\varphi _0$ and r_0...". Below, (first column, fourth figure in figure FIGREF28): "Relative protein levels of ubiquitin-protein conjugates in M. quadriceps...A representative immunoblot specific to ubiquitin...". Consistently with our analysis, activation focuses on the most relevant tokens for each text feature: "Aliev-Panfilov model" and "immunoblot", respectively.

Conclusions

There is a wealth of knowledge in scientific literature and only a fraction of it is text. However, understanding scientific figures is a challenging task for machines, which is beyond their ability to process natural images. In this paper, we provide empirical evidence of this and show that co-training text and visual features from a large corpus of scientific figures and their captions in a correspondence task (FCC) is an effective, flexible and elegant unsupervised means towards overcoming such complexity. We show how such features can be significantly improved by enriching them with additional knowledge sources and, particularly, structured KGs. We prove the benefits of our approach against supervised baselines and in different transfer learning tasks, including text and visual classification and multi-modal machine comprehension applied to question answering, with results generally beyond the state of the art. In the future, it will be interesting to further the study of the interplay between the semantic concepts explicitly represented in different KGs, contextualized embeddings e.g. from SciBERT BIBREF31, and the text and visual features learnt in the FCC task. We also plan to continue to charter the knowledge captured in such features and to pursue the optimization and practical application of our approach.

Acknowledgments

The research reported in this paper is supported by the EU Horizon 2020 programme, under grants European Language Grid-825627 and Co-inform-770302.