DISCRIMINATIVE HALLUCINATION FOR MULTI-MODAL FEW-SHOT LEARNING

Frederik Pahde[1,2], *Moin Nabi*[1], *Tassilo Klein*[1], *Patrick Jähnichen*[2]

[1] Machine Learning Research, SAP SE, Berlin, Germany
[2] Humboldt-Universität zu Berlin
{frederik.pahde, m.nabi, tassilo.klein}@sap.com
patrick.jaehnichen@hu-berlin.de

ABSTRACT

State-of-the-art deep learning algorithms yield remarkable results in many visual recognition tasks. However, they still catastrophically struggle in low data scenarios. To a certain extent, the lack of visual data can be compensated by multimodal information. Information missing in one modality (e.g. image) due to the limited data can be included in the other modality (e.g. text). In this paper, we propose a benchmark for few-shot learning with multi-modal data which can be used by other researchers. We also introduced a method for few-shot fine-grained recognition, utilizing textual descriptions of the visual data. We developed a two-stage framework built upon the idea of cross-modal data hallucination. For each visual category, we first generate a set of images by conditioning on the textual description of the category using StackGANs. Next, we rank the generated images based on their class-discriminativeness and only pick the most discriminative images to extend the dataset. Lastly, a classifier invariant to our framework can be trained using an extended training set. We show the results of our proposed discriminative hallucinated method for 1-, 2-, and 5-shot learning on the CUB dataset, where the accuracy is improved by employing the multi-modal data.

Keywords: Few-Shot Learning, Multi-Modal, Meta-Learning, Fine-grained Recognition

1. INTRODUCTION

In recent years, deep learning techniques have achieved exceptional results in many domains such as computer vision and NLP. These advances can be explained by improvements to algorithms and model architecture along with increasing computational power, and in particular growing availability of big data. However, the big data assumption, which is key for deep learning applications, is at the same time the limiting factor. For many applications, it is often too expensive or even impossible to acquire enough training samples in order to learn a model at sufficient accuracy. Furthermore, the requirement for large amounts of training data is in stark contrast to human learning, which can quickly learn from few instances. This is what makes alternative learning approaches that require less training data an attractive research topic. Thus research in the domain of few-shot learning, i.e. learning and generalizing from few training samples, has gained more and more interest (e.g. [1, 2, 3, 4, 5, 6, 7]). However, research conducted has mainly focused on approaches with data coming from only one modality, especially image data. Overcoming this limitation and even including data from other modalities, e.g. textual descriptions in addition, can further improve the model. For example, in a bird classification task, additionally to the few images for a novel bird category, a textual description about the most distinguishing features of the bird can be provided. The assumption of our approach is that having fine-grained descriptions provided from multi-modal data can force the model to focus on the more discriminative features of novel classes in order to achieve improve performance in the few-shot learning setting [8, 9]. This assumption leads to the proposed study of few-shot learning with multi-modal data, more precisely images with fine-grained textual descriptions. To approach this problem we propose a two-stage framework that learns how to generate images given textual descriptions using conditional generative adversarial networks (GAN) [10], [11]. This is followed by a strategy to pick the images with the most class-discriminative information. With this strategy serving as a quality control, we can train a classifier invariant to our proposed framework using the few existing samples plus the generated images as training data. Thus, few-shot learning on multi-modal data is tackled by guided image hallucination conditioned on the textual description.

The most closely related work is by Hariharan et al. [2] and Wang et al. [12], who also use hallucinated data for few-shot learning with the difference of the restriction to a mono-modal image context. Similarly, Zhang et al. [13] proposed StackGANs to generate high-resolution images from textual descriptions and tested this framework in a zero-shot scenario.

Our work has multiple contributions: **First**, we extend the few-shot learning benchmark of [2] to work with multi-modal data. **Second**, we propose a novel approach that employs a multi-stage framework based on StackGAN that facilitates few-shot learning by hallucinating images conditioned on textual descriptions. **Third**, our framework includes a new strategy for quality assessment of the hallucinator to pick the best generated images.

2. RELATED WORK

2.1. Few-Shot Learning

For learning with limited amounts of data, [14] proposed a metric learning approach for which siamese convolutional networks were used in a one-shot learning scenario to rank the similarity of inputs. Other work seeks to avoid overfitting by modifications to the loss function or the regularization term. [15] proposed a clustering of neurons on each layer of the network and calculated a single gradient for all members of a cluster during the training to prevent overfitting. A more intuitive strategy is to approach few-shot learning on data-level, meaning that the performance of the model can be improved by finding strategies to enlarge the training data. For example, [16] proposed a semi-supervised approach in which a large unlabeled dataset containing similar images was included in addition to the original training set. [2] combined both strategies (data-level and algorithm-level) by defining the squared gradient magnitude (SGM) loss on the one hand and generating new images by hallucinating features on the other hand. Other recent approaches to few-shot learning have leveraged meta-learning strategies. [1] trained a long short-term memory (LSTM) network as meta-learner that learns the exact optimization algorithm to train a learner neural network that performs the classification in a few-shot learning setting. [17] introduced matching networks for one-shot learning tasks. This network is able to apply an attention mechanism over embeddings of labeled samples in order to classify unlabeled samples. [3] proposed prototypical networks which can be interpreted as generalization for matching networks. Prototypical networks search for a non-linear embedding space in which classes can be represented as the mean of all corresponding samples, called a prototype and classification is performed by finding the closest prototype in the embedding.

2.2. Multi-modal Learning

By defining a encoder-decoder pipeline, [18] proposed a method to align visual and semantic information in a joint embedding space. [19] were able to improve this mixed representation by incorporated a triplet ranking loss. The work of [20] aims to generate image descriptions. Their model is able to infer latent alignments between regions of the image and segments of the sentences for the image description. [21] put their focus on fine-grained visual descriptions. They collected two datasets containing fine-grained visual descriptions and proposed a deep structured joint embedding that is end-to-end trainable.

2.3. Conditional GANs

After the introduction of GANs in [10], the conditional generation of data was investigated in [22]. [23] further studied image synthesis based on textual information. Using StackGANs, [13] pushed the quality of the generated images to photo-realistic high-resolution level by stacking multiple GANs. Following, the same authors presented an end-to-end trainable version of StackGAN [13].

3. METHOD

3.1. Preliminaries

The core of our framework is built upon the idea to generate images based on textual descriptions. Thus, the functionality of StackGAN is key for our work. The idea behind StackGAN is to use multiple GANs with different levels of granularity. In a StackGAN with s stacked GANs, there are the generators G_1, ..., G_s and discriminators D_1, ..., D_s. G_1 is conditioned on a text embedding φ_t for text t and generates a low-resolution image I_1. D_1 gets the generated image I_1 and φ_t as input and predicts whether the image is real or fake given the textual description. Based on the decision of D_1 the generated image will be passed as input to the next level GAN. Having this pipeline, the resolution is increased at every stage of the StackGAN, eventually turning into a high-resolution image at the last stage. See [13] for further details. If not noted differently, we use G and D to refer to the last stage generator and discriminator, respectively.

3.2. Task Description

We extend the few-shot learning benchmark defined by [2] to work with multi-modal data. Our goal is to model a few-shot framework, which consists of multiple stages, with the overall idea of learning a representation on a large training data corpus followed by few-shot finetuning for novel categories with few training samples. As proposed by [2], the classes C are split into base classes C_{Base} for which many samples exist and novel classes C_{Novel} with just a few samples. In this scenario, data from base classes can be used to learn meaningful representations to perform few-shot learning on novel classes. Our proposed approach is multi-modal in training, such that training samples are tuples $x^j = \left(I^j, T^j\right)$ consisting of an image $I^j \in \mathcal{I}$ and a textual description $T^j \in \mathcal{T}$, where $\mathcal{I}$ and $\mathcal{T}$ are the image space and text space respectively. The testing phase is mono-modal on image data of C_{Novel}.

3.3. Two-Stage Framework for Image Hallucination

The overall pipeline of the proposed method (see Fig. 1) can be split into two phases: 1) representation learning in which a generative model is trained to hallucinate images given a textual description and 2) class-discriminative finetuning in

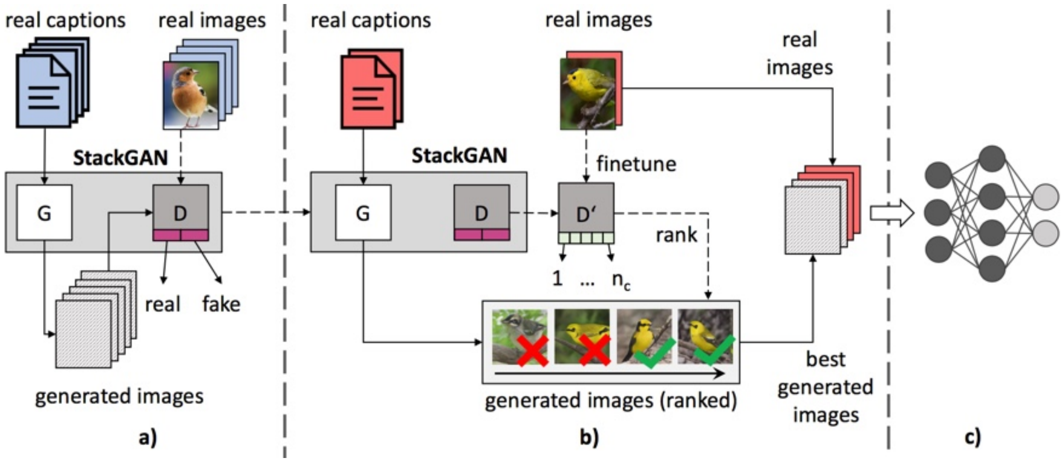

Fig. 1. Two-stage Framework for proposed method for multi-modal few-shot learning followed by a classifier: a) representation learning phase in which the StackGAN is trained, b) class-discriminative finetuning to learn how to pick the best generated images, c) training of a classifier on extended training set.

which we learn to pick the most discriminative images out of the generated data. Finally, we train a generic classifier.

3.3.1. Train image hallucinator

During representation learning the large amount of data available in C_{Base} is used to learn a meaningful representation of the data. Doing so in multi-modal fashion yields a mapping: $\mathcal{T} \rightarrow \mathcal{I}$. This in turn allows cross-modal generation, facilitating the compensation of the lack of data in C_{Novel}.

3.3.2. Class-discriminative finetuning of hallucinator

In the class-discriminative finetuning phase the (sub-)set of n training samples that are available within C_{Novel} are used to improve the output of the StackGAN. Next, the cross-modal image synthesis is employed, obtained from the representation phase. This allows for the creation of a potentially infinite amount of samples given textual descriptions. However, the challenge is to pick adequate samples out of the pool of generated samples that allow for building a better classifier within the few-shot scenario. Here we use the score of the discriminator D as a measure for assessing the quality of the generated images. However, it should be noted that D does not possess any class-discriminative information. In order to add this property, we take the D trained in the representation learning phase and replace the last layer (classification layer) by a linear layer containing $n_c = |C_{Novel}|$ output neurons. This network is then finetuned using the few samples of C_{Novel}. We refer to this network as D'. Next, per category generated images are ranked based on the score of the output neuron corresponding to the given class. We use $I_{generated}$ to refer to the set containing the best generated images based on their ranking using the score of D'. In a last step, an image classifier is built using the training set $D_{train} = I_{real} \cup I_{generated}$, where I_{real} is a set containing the n real samples.

4. EXPERIMENTS

4.1. Data

For our experiments we use the CUB bird dataset ([24]), which contains 11,788 images of 200 different bird species, with $I \in \mathbb{R}^{256x256}$. The data is split equally in training and test data, meaning that there are roughly 30 training and 30 test images per category. 10 short textual descriptions per image are provided by [21]. Similar to [13], we use the text-encoder pre-trained by [21], yielding a text embedding $T \in \mathbb{R}^{1024}$. Following [13], we split the data such that $|C_{Base}| = 150$ and $|C_{Novel}| = 50$. To perform few-shot learning $n = \{1, 2, 5, 10, 20\}$ images of C_{Novel} are used for training, as proposed by [2].

4.2. Model

In the representation learning phase we train a StackGAN in the setup suggested by [13] for 600 epochs, yielding D and G. For the sake of simplicity, a basic convolutional neural network (CNN) architecture is employed for classification. It should be noted that in theory any other classifier could be used. The CNN consists of two convolutional layers and two max-pooling layers, followed by two linear layers that are connected with dropout and finally a softmaxlayer with $|C_{Novel}|$ units. For training, SGD is used for 800 epochs with a learning rate of 0.01 and momentum of 0.5.

The experiments are composed of: 1) A baseline (R) in which we train the classifier only on real data, i.e. n images per category. 2) We generate images using G conditioned on one caption randomly chosen (out of 10) for the missing $30 - n$ images of C_{Novel}. Following this, the classifier is

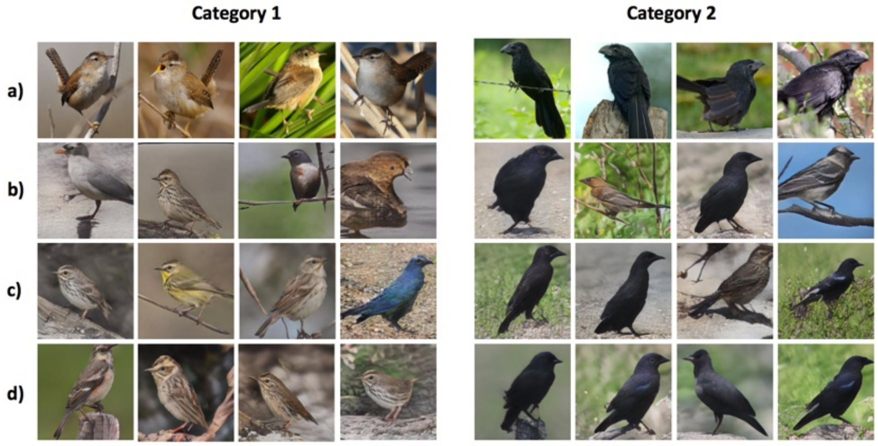

Fig. 2. Images for two different bird categories: a) real image data, b) data generated with StackGAN conditioned on random captions, c) data generated with StackGAN conditioned on best captions based on TFIDF scores and ranked with D, d) data generated with StackGAN conditioned on best captions based on TFIDF scores and ranked with D'.

trained on an extended dataset that contains the real data and the generated data (StG). 3) We pick the $30 - n$ captions containing the most discriminative information by making use of term frequency - inverse document frequency (TF-IDF). For this we sum up the TF-IDF scores for every word in a caption, excluding stopwords. We then generate 10 images per chosen caption with G, and rank the images by the score of D (real vs. fake discriminative) (StGD), of which again $30-n$ are retained. 4) Similar to 3), with difference of employing the class-discriminative D' for ranking generated images (StGD'). Doing that we do not pick images based on their realistic appearance only, but based on how class-discriminative they are. The top-5 accuracy of the classifier for our different experiments is reported in table 1.

Table 1. Top-5 accuracy in percent of our classifier on only real data (R), generated data on random captions (StG), generated data assessed with D (StGD) and generated data assessed with D' (StGD'). Best results are bold.

			n		
Method	1	2	5	10	20
R	19.9	24.8	36.8	49.2	**70.6**
StG [13]	27.3	30.7	37.4	45.3	69.0
Our (StGD)	25.7	30.1	36.8	50.5	68.6
Our (StGD')	**28.5**	**31.6**	**41.7**	**52.2**	68.5

5. RESULTS

We observe that the proposed approach outperforms the baseline by a large margin in the particular challenging few-shot scenarios with $n = 1, 2, 5$ by 4.9 to 8.6 percentage points, respectively. In addition to commonly reported top-5 accuracy, we evaluated our experiments with top-1 and top-3 accuracy and observed similar performance results. Using the score of D as measure to rank the hallucinated images has shown to be harmful compared to image generation without further post-processing. However, forcing the discriminator to pick images based on a class-discriminative score by means of D' leads to higher accuracies. Further, qualitative analysis by means of visualization of generated data (see figure 2) in our experiments confirms that images ranked high by D' contain the most class-discriminative features (d). In contrast to that, only picking random descriptions as input for the hallucinator leads to an undesirable large variety of birds because many descriptions do not include sufficient class-discriminative information (b). Further, ranking on D produces realistic looking images, however, mixing categories (c).

6. CONCLUSION AND FUTURE WORK

Our experiments confirm that multi-modality allows to close the information gap in few-shot scenarios, yielding more robust classifiers. First results on additional datasets suggest similar trends and confirm our findings. For future work we plan to investigate the use of D' as the final classifier, since it already contains class-discriminative abilities. Furthermore, we seek to incorporate class-discriminativeness into the representation learning phase to further improve the performance. This would force the generator G to output more discriminative images. For example, this can be achieved by adding a pairwise ranking loss for D during the training of the StackGAN. Last, future research will be conducted on optimizing the text embedding in context of multi-modality.

7. REFERENCES

[1] Sachin Ravi and Hugo Larochelle, "Optimization as a model for few-shot learning," in *InternationalConference on Learning Representations*, 2017.

[2] Bharath Hariharan and Ross Girshick, "Low-shot Visual Recognition by Shrinking and Hallucinating Features," in *ICCV*, 2017.

[3] Jake Snell, Kevin Swersky, and Richard Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems 30*, pp. 4080–4090. Curran Associates, Inc., 2017.

[4] Luca Bertinetto, Joo F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi, "Learning feed-forward one-shot learners," in *Advances in Neural Information Processing Systems*, 2016, pp. 523–531.

[5] Yu-Xiong Wang and Martial Hebert, "Learning from Small Sample Sets by Combining Unsupervised Meta-Training with CNNs," 2016, pp. 244–252.

[6] Yao-Hung Hubert Tsai and Ruslan Salakhutdinov, "Improving One-Shot Learning through Fusing Side Information," *arXiv:1710.08347 [cs]*, Oct. 2017, arXiv: 1710.08347.

[7] Eleni Triantafillou, Richard S. Zemel, and Raquel Urtasun, "Few-shot learning through an information retrieval lens," *CoRR*, vol. abs/1707.02610, 2017.

[8] Jonathan Krause, Benjamin Sapp, Andrew Howard, Howard Zhou, Alexander Toshev, Tom Duerig, James Philbin, and Li Fei-Fei, "The Unreasonable Effectiveness of Noisy Data for Fine-Grained Recognition," *arXiv:1511.06789 [cs]*, Nov. 2015, arXiv: 1511.06789.

[9] Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed Elgammal, "Link the head to the "beak": Zero shot learning from noisy text description at part precision," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen, "Improved techniques for training gans," 2016, pp. 2234–2242.

[12] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan, "Low-Shot Learning from Imaginary Data," *arXiv:1801.05401 [cs]*, Jan. 2018, arXiv: 1801.05401.

[13] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *ICCV*, 2017.

[14] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *ICML Deep Learning Workshop*, 2015, vol. 2.

[15] "Efficient K-Shot Learning with Regularized Deep Networks," in *Association for the Advancement of Artificial Intelligence (AAAI) 2018*.

[16] Matthijs Douze, Arthur Szlam, Bharath Hariharan, and Hervé Jégou, "Low-shot learning with large-scale diffusion," *CoRR*, 2017.

[17] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al., "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3630–3638.

[18] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models," *arXiv:1411.2539 [cs]*, Nov. 2014, arXiv: 1411.2539.

[19] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler, "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives," *arXiv:1707.05612 [cs]*, July 2017, arXiv: 1707.05612.

[20] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[21] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele, "Learning deep representations of fine-grained visual descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 49–58.

[22] Mehdi Mirza and Simon Osindero, "Conditional Generative Adversarial Nets," in *Deep Learning Workshop NIPS 2014*, 2014.

[23] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee, "Generative adversarial text to image synthesis," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 1060–1069.

[24] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Tech. Rep., 2011.