

Dataset Creation for Visual Entailment using Generative AI

Rob Reijtenbach

Leiden University

rob.reijtenbach@gmail.com

Suzan Verberne

Leiden University

(s.verberne|g.wijnholds)@liacs.leidenuniv.nl

Gijs Wijnholds

Leiden University

Abstract

In this paper we present and validate a new synthetic dataset for training visual entailment models. Existing datasets for visual entailment are small and sparse compared to datasets for textual entailment. Manually creating datasets is labor-intensive. We base our synthetic dataset on the SNLI dataset for textual entailment. We take the premise text from SNLI as input prompts in a generative image model, Stable Diffusion, creating an image to replace each textual premise. We evaluate our dataset both intrinsically and extrinsically. For extrinsic evaluation, we evaluate the validity of the generated images by using them as training data for a visual entailment classifier based on CLIP feature vectors. We find that synthetic training data only leads to a slight drop in quality on SNLI-VE, with an F-score 0.686 compared to 0.703 when trained on real data. We also compare the quality of our generated training data to original training data on another dataset: SICK-VTE. Again, there is only a slight drop in F-score: from 0.400 to 0.384. These results indicate that in settings with data sparsity, synthetic data can be a promising solution for training visual entailment models.

1 Introduction

Natural language inference (NLI) is a classification problem for pairs of two texts, a premise and a hypothesis. The pair is labeled as *entailment* (the premise entails the hypothesis), *neutral* or *contradiction* (the hypothesis contradicts the premise). In visual entailment (VE) tasks (Xie et al., 2019), the premise is substituted by an image, while the hypothesis is still in text form.

In order to create and train effective models for VE, large datasets are needed. While datasets of images combined with hypotheses and labels do exist, they are relatively small and sparse compared to datasets for textual entailment. Existing datasets are SNLI-VE (Xie et al., 2019) and SICK-

VTE (Iokawa et al., 2024) which are both based on NLI datasets and which were created by manual labor leveraging Amazon Mechanical Turk workers. In this paper we evaluate the use of generative AI for VE dataset creation which would allow cheaper and easier dataset creation. This is done by first generating a synthetic dataset, of which we then verify the validity. We introduce a synthetic version of the SNLI-VE dataset called Synthetic-NLI-VE and show how models trained on this dataset have similar performance when tested on real data compared to models trained on real data.

In summary, the contributions of this paper are threefold: (1) we present the new dataset Synthetic-NLI-VE¹; (2) we find that the performance of models trained on the generated dataset have similar performance compared to models trained on real data; (3) A cross-data evaluation shows that generalizability of visual entailment models to a different dataset is poor, whether or not the training set was generated or original.

2 Related work

Visual entailment and dataset creation The idea of visual entailment was first proposed by Xie et al. (2019). For this task they introduce the Explainable Visual Entailment (EVE) model, based on Attention Visualization. In the same paper the authors introduce the SNLI-VE dataset (Section 3). Antol et al. (2015) introduced a dataset for visual question answering (QA). They used the Microsoft Common Objects in Context (MS COCO) dataset (Lin et al., 2014) as a starting point: ~200k images of real-world scenes with 5 captions per image. They added 50k images of abstract scenes for which they also collected 5 captions per image.

Marelli et al. (2014) created the SICK dataset. SICK (*sentences involving compositional knowl-*

¹<https://huggingface.co/datasets/robreijtenbach/Synthetic-NLI-VE>

edge) contains sentence pairs with both relatedness scores and entailment labels. This dataset was created by pairing the Flickr8K dataset (Hodosh et al., 2013) and the SemEval-2012 STS data (Agirre et al., 2012) and having Amazon Mechanical Turk workers annotate them with both similarity scores and entailment labels. Wijnholds and Moortgat (2021) created the Dutch version of SICK using a semi-automatic translation. Bowman et al. (2015) introduced the SNLI dataset on which the aforementioned SNLI-VE was based, with as motivation that the SICK dataset is too small and not balanced enough. For SNLI they created a balanced dataset of around $\sim 500k$ sentence pairs compared to the $\sim 10k$ in the SICK dataset.

There are also efforts made to improve existing datasets. This was already the case with Goyal et al. (2017), who improved and extended the VQA dataset resulting in the VQA-v2 dataset. The dataset was improved by, among other things, reducing bias and extended it by adding more images. This has also been done for the SNLI-VE dataset by Do et al. (2021) who created the e-SNLI-VE-2.0.

Synthetic data Unlike the largely human made datasets that were previously discussed, the CLEVR dataset (Johnson et al., 2016) is automatically generated. This dataset contains images of abstract shapes combined with automatically generated questions. The images were created by randomly sampling a scene graph and rendering it using the open-source 3D rendering software Blender.

Yuan et al. (2024) proposed an evaluation framework for assessing synthetic data generated by large language models (LLMs). This framework includes measures for fidelity, utility and privacy. In this work, we only focus on the fidelity and utility of the generated data.

Some research suggests that using synthetic datasets for model training could have a negative effect on performance in the future, if generated datasets are used for training computer vision models (Hataya et al., 2023). As opposed to synthetic datasets used to train generative models, the images that we generate are used to train classification models. Furthermore, these classification models are evaluated on original data, ensuring good real world generalizability.

3 Data

In this work we use two datasets which we briefly describe in this section.

SNLI-VE This was introduced by (Xie et al., 2019), by combining the SNLI dataset (Bowman et al., 2015) with the Flickr30k dataset (Young et al., 2014). The Flickr30k dataset was created by taking 31,783 photos of everyday activities which were harvested from Flickr. Each image receives 5 different captions resulting in 158,915 captions in total. Figure 3 in the appendix shows an example of an image and its captions.

The SNLI dataset (Bowman et al., 2015) is a well known dataset specifically created for natural language inference. In short, it was constructed by having Amazon Mechanical Turk workers generate 3 hypotheses per caption, where captions came from the Flickr30k dataset. From this, Xie et al. (2019) could therefore create the SNLI-VE dataset by replacing each premise by the original corresponding image. The dataset contains a total of 31,783 images, 157,567 premises and 565,286 hypotheses.

SICK-VTE Along the lines of the creation of SNLI-VE, Iokawa et al. (2024) introduces SICK-VTE, a visual entailment version of (a subset of) the SICK dataset (Marelli et al., 2014), but with an additional multilingual component, including also the Dutch (Wijnholds and Moortgat, 2021) and Japanese (Yanaka and Mineshima, 2022) translations of the SICK dataset. The construction of the original SICK dataset was based on sentence transformation rules over image captions instead of human-generated hypothesis. By construction the dataset contains only cases of Entailment and Contradiction: for 488 unique images there are 2,899 sentence pairs, with 1,930 examples of Entailment and 969 examples of Contradiction.

4 Methods

We generate a synthetic dataset as described in §4.1. We then report on the intrinsic evaluation of image quality by comparing the generated images directly with the original images based on a similarity analysis in §4.2. Finally, we perform extrinsic evaluation of synthetic data, comparing it to original data for visual entailment model training in §4.3.

4.1 Image Generation

Our approach for creating the generated dataset is to use the premise text from SNLI as input prompts in a generative model, creating an image for every premise caption. This results in a dataset similar to SNLI-VE, however, instead of multiple premises

referencing the same image, here the resulting dataset has a unique image for every premise. We refer to the generated images as *child images* to express the fact that they were indirectly derived from an original *parent image*. Examples of generated child images are shown in Figure 1.

Our choice of generative model is Stability AI’s Stable Diffusion². The ability to run the model locally as opposed to the cloud based solutions from OpenAI and Midjourney was essential for generating the large amount of images necessary for our work.

The chosen resolution was square images of 512x512 pixels as this is the image size Stable Diffusion was trained on and it is close to the average image size of the original SNLI-VE dataset.³ The checkpoint chosen for this research is Realistic Vision v51⁴ which was finetuned for generating photorealistic images.

4.2 Intrinsic evaluation

To assess intrinsic image quality we rely on two measures. As an initial verification we compute pairwise cosine similarity between the CLIP feature vectors of original and generated images and assess the distribution of these values, expecting to see a normal distribution.

Secondly, we use ranked similarity scores over the full dataset to inspect whether, for a given original image, the 5 generated images for it will appear as highly similar or not. We specifically use recall@k and precision@k for evaluation:

In the ranking problem in this work, we take the query to be an original image, and the ranked list of documents to be the 100 most similar generated images as determined by cosine similarity. The relevance function is now binary, returning 1 for an image that was indeed generated from one of the captions of the original image, and 0 otherwise.

For precision@k, we divide the true positives by the number of retrieved images.

4.3 Extrinsic evaluation

We test the validity of the generated images by using them as training data for a classifier to learn the visual entailment classification problem. The

approach for this experiment is based on Song et al. (2022) who proposed using CLIP for visual entailment. Their method includes taking the CLIP feature vector of both the premise image and the hypothesis text, fusing these according to Equation 1 and training an MLP on this fused vector representation to output the correct entailment label.

$$\text{fuse}(v_1, v_2) = [v_1, v_2, v_1 + v_2, v_1 - v_2, v_1 \cdot v_2] \quad (1)$$

The input dimension for this perceptron is 2560 which is a direct result of the output size of the fuse function. The fuse function concatenates the feature vector of the image, the feature vector of the hypothesis, the sum of these two vectors as well as the difference between these vectors and finally the product of these vectors. This results in a total of five vectors that are concatenated and with each vector having a size of 512 numbers, the result has a length of $5 * 512 = 2560$.

The resulting vector is used as an input for the MLP which has one hidden layer of size 250. After experimenting with different layer sizes, the size of this hidden layer did not seem to affect the accuracy of the classifier but had an impact on the computational performance. After this one hidden layer the network only has one more layer which is the output layer. This output layer has a size of 3 corresponding to the three possible labels: entailment, neutral, contradiction.

We use this method to train classifiers on both the original images and the generated images of the SNLI-VE dataset. These classifiers are then tested on the original as well as on the generated test sets, after which their performance is compared. Note that absolute performance of the classifier is not the primary goal. Rather, we are interested in the relative performance of a classifier trained on generated images compared to a classifier trained on real images. We, however, aim for good performance of both as this yields the most accurate data to compare between these two.

5 Experiments and Results

In this section we first report on the results for the intrinsic evaluation (§5.1), after which we discuss the downstream performance in the Visual Entailment task (§5.2), and finally we discuss the results of transferring the Visual Entailment model to the SICK-VTE dataset (§5.3).

²<https://github.com/Stability-AI/generative-models>

³The mean width and height were 459 and 395 respectively, and the standard deviations were 67 for width and 74 for height with both having a maximal value of exactly 500.

⁴<https://huggingface.co/stablediffusionapi/realistic-vision-v51>



A wedding party walks out of a building.



The group of people are assembling for a wedding.



A man and woman dressed for a wedding function.

Figure 1: Three examples of generated images based on three of the captions in Figure 3.

5.1 Intrinsic evaluation

The starting point of our intrinsic comparison is the cosine similarity distribution for images in the development and test set of the SNLI-VE dataset and its generated child images. Each original image is compared to all the generated images and the similarity scores are saved. We found that the similarity values follow a normal distribution for both the development and test set. The mean for both sets is 0.465 with a standard deviation of ~ 0.085 . This is also illustrated in Figure 4 in the appendix.

Ranked similarity After assessing the similarity distribution between original and generated images, we report on the recall@k and precision@k curves. Initially, we computed average recall@k and precision@k values for $k = 100$, which reveals that on average only 1.6 of the 100 most similar synthetic images to the real images were based actually generated based on one of the premises accompanying that real image. These results stem from the fact that finding the 100 most similar out of $\sim 160k$ generated images will likely not result in finding all of the 5 images that are relevant. This is illustrated in Figure 2 where an image is shown together with the most cosine similar generated image which is not one of its child images. These two images could be considered rather similar by a human. It is likely that there are more images in the collection that are similar than only the child images, making the recall@k measure an underestimation of the real quality of the generated images. The recall@k and precision@k curves for this setting are in Figures 5a and 5b in the appendix.

To get a fairer picture of the similarity evaluation, we recalculate recall@k and precision@k curves for a sampled version of the data which is needed as the train set is large very large compared to the dev and test set, which are only 1000 original images

Train set	Original	Generated
Original	70.3% / 0.703	71.1% / 0.710
Generated	68.9% / 0.686	73.2% / 0.732

Table 1: Accuracies/F1 scores of both models on both test sets of SNLI-VE.

each. We randomly sample 1000 examples from the train set of SNLI-VE, and consequently calculate recall@k and precision@k values for train, development, and test sets separately, each time considering 1000 original images and its ~ 5000 generated child images. The resulting plots for the recall@k and precision@k of the samples are in Figure 6b and Figure 6a in the appendix. We find that the average success rate is between 3.5 and 4 out of the five possible relevant images, indicating that most of the relevant real images are found within the first 100 most similar generated images.

For completeness, we include the variance of the recall and precision curves of the samples in Figure 7 in the appendix where one standard deviation above and below each curve is marked.

5.2 Extrinsic Evaluation: Classification

We train both a model on the dataset of original images, and a model on the dataset of generated images, using the same train/dev/test split as suggested for the SNLI-VE dataset. We trained the model for 100 epochs and selecting the epoch for which the model performs highest on the development set, which was saved for evaluating on the test set. The accuracy and loss on the training set and dev set are shown in the appendix in Figure 8a and 8b and Figure 9a and 9b respectively.

We report accuracies and F1 scores in Table 1. We observe the best overall performance when using the model trained on generated data evaluated on the generated data as well. This suggests that



(a) Original



(b) Generated

Figure 2: An example of an image and a generated image which looks similar but is not considered relevant as the generated image is not a child of the original image in this evaluation. The original image (a) had 5 captions in the dataset written by 5 different workers. Image (b) was generated for the caption “A group of young men have finished their drinks while sitting at a table in a restaurant .”

the generated images and their classification has less variability compared to the original data. We also see that the model trained on original images performs better on the generated test set than it does on the original test set. This could suggest that the generated test set is “easier” to classify. Lastly, and most importantly, we do see that the model trained on generated data and tested on original data has a somewhat lower performance in this experiment, but the difference is small. It suggests that synthetic training data results in slightly worse performance in real world tasks.

5.3 Cross-data generalizability

The final part of the experiments evaluate the performance of the trained models when they are tested on another dataset, in this case the SICK-VTE dataset. As discussed in Section 3, SICK-VTE and its synthetic counterpart do not contain any neutral examples. To train visual entailment models, having neutral examples would be essential however for the purpose of testing the generalizability pretrained models, a dataset with neutral examples is preferred.

The experimental setup is similar to that of the classification experiment in Section 5.2, except that we now reuse the trained models from the prior experiment as we assess transfer capabilities. Both of the trained models were tested on the original SICK-VTE dataset and, for completeness, also on the generated version of SICK-VTE. Similar to the previous experiment, we report both accuracy and F1 scores in Table 2. Note that, in contrast to the results on the SNLI-VE dataset, accuracy and F1 scores diverge, due to label imbalance in SICK-VTE.

Train set	Original	Generated
Original	50.7% / 0.400	51.4% / 0.391
Generated	47.2% / 0.384	47.6% / 0.384

Table 2: Accuracies/F1 scores of both models on the SICK-VTE datasets.

We find that performance is relatively poor, given a majority baseline of 0.6657 for a model only predicting Entailment. This result is in line with the findings of [Talman and Chatzikyriakidis \(2019\)](#), who found similar issues when transferring models trained on the SNLI dataset to the SICK dataset. Secondly, we can conclude that the model trained on generated data performs slightly worse compared to the model trained on original data. This is in line with the findings in the previous experiment (§5.2).

6 Conclusion

In this paper we introduced a synthetic VE dataset Synthetic-NLI-VE. The dataset proved to have similar utility compared to the dataset it was based on while being far less costly to create. This also proves the viability of using generative AI to create datasets for the VE task, whereby we pave the way for future research into using synthetic data for VE dataset creation. As future work we propose changing the single set of parameters for the generation model to a variety of different values. Secondly, generating more than one image per caption could result in better training data compared to the one image per caption dataset we generated. Lastly, evaluating different classification algorithms could further strengthen the findings.

Limitations

Our experiments are limited evaluation for the CLIP model, and the findings might be different for other visual entailment models.

We investigated cross-data generalizability in synthetic VTE datasets. One limitation of our experiments is that both SNLI-VE and SICK-VTE are created based on Flickr30K, which makes them relatively more similar to each other than datasets based on other sources, such as NLVR and NLVR2.⁵ We leave this cross-domain evaluation for future work.

References

- Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. [SemEval-2012 task 6: A pilot on semantic textual similarity](#). In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [VQA: visual question answering](#). *CoRR*, abs/1505.00468.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2021. [e-snli-ve-2.0: Corrected visual-textual entailment with natural language explanations](#). *CoRR*, abs/2004.03744.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. [Making the v in vqa matter: Elevating the role of image understanding in visual question answering](#). *CoRR*, abs/1612.00837:6325–6334.
- Ryuichi Hataya, Han Bao, and Hiromi Arai. 2023. [Will large-scale generative models corrupt future datasets?](#) In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20498–20508. IEEE.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. [Framing image description as a ranking task: data, models and evaluation metrics](#). *Journal of Artificial Intelligence Research*, 47(1):853–899.
- Nobuyuki Iokawa, Gijs Wijnholds, and Hitomi Yanaka. 2024. [Multilingual visual-textual entailment benchmark with diverse linguistic phenomena](#). *Proceedings of the Annual Conference of JSAI*, JSAI2024:4C3GS1104–4C3GS1104.
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2016. [CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning](#). *CoRR*, abs/1612.06890.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft COCO: common objects in context](#). *CoRR*, abs/1405.0312.
- Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. [A SICK cure for the evaluation of compositional distributional semantic models](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. 2022. [CLIP models are few-shot learners: Empirical studies on VQA and visual entailment](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6088–6100, Dublin, Ireland. Association for Computational Linguistics.
- Aarne Talman and Stergios Chatzikyriakidis. 2019. [Testing the generalization power of neural network models across NLI benchmarks](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 85–94, Florence, Italy. Association for Computational Linguistics.
- Gijs Wijnholds and Michael Moortgat. 2021. [SICK-NL: A dataset for Dutch natural language inference](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1474–1479, Online. Association for Computational Linguistics.
- Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. [Visual entailment: A novel task for fine-grained image understanding](#). *CoRR*, abs/1901.06706.
- Hitomi Yanaka and Koji Mineshima. 2022. [Compositional evaluation on Japanese textual entailment and similarity](#). *Transactions of the Association for Computational Linguistics*, 10:1266–1284.
- Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. [From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions](#). *Transactions of the Association for Computational Linguistics*, 2:67–78.

⁵<https://lil.nlp.cornell.edu/nlvr/>

Yefeng Yuan, Yuhong Liu, and Liang Cheng. 2024.
[A multi-faceted evaluation framework for assessing
synthetic data generated by large language models.](#)
Preprint, arXiv:2404.14445.

Appendix

Additional figures are on the following pages.



- A bearded man, and a girl in a red dress are getting married.
- A wedding party walks out of a building.
- The group of people are assembling for a wedding.
- A man and woman dressed for a wedding function.
- A woman holds a man's arm at a formal event.

Figure 3: One of the $\sim 30k$ photos and its 5 accompanying captions from the SNLI dataset.

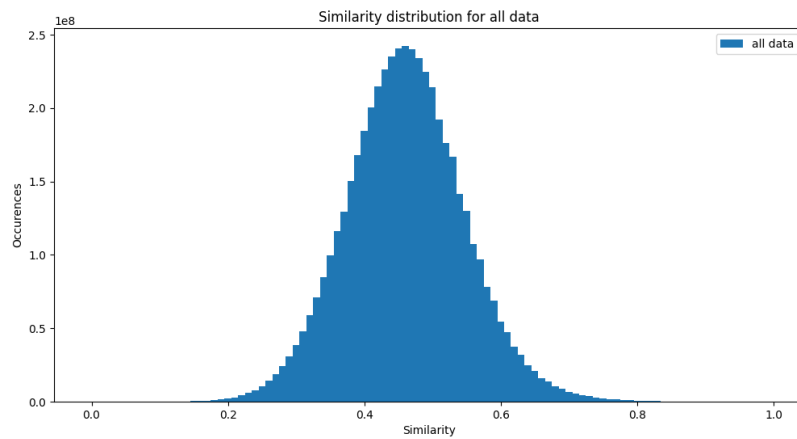


Figure 4: Cosine similarity values for the dataset, showing the expected normal distribution.

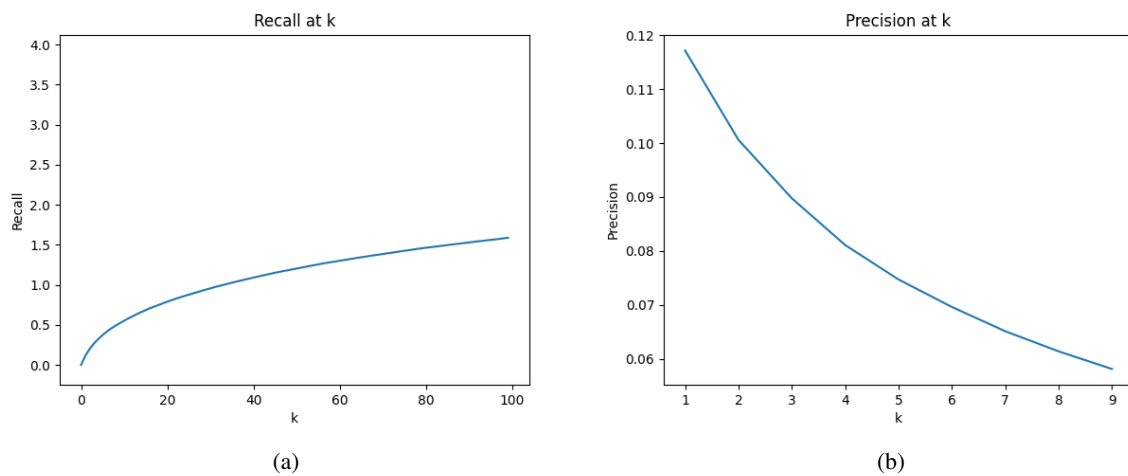


Figure 5: Recall (a) and precision (b) curves, calculated as averaged over the full dataset of images.

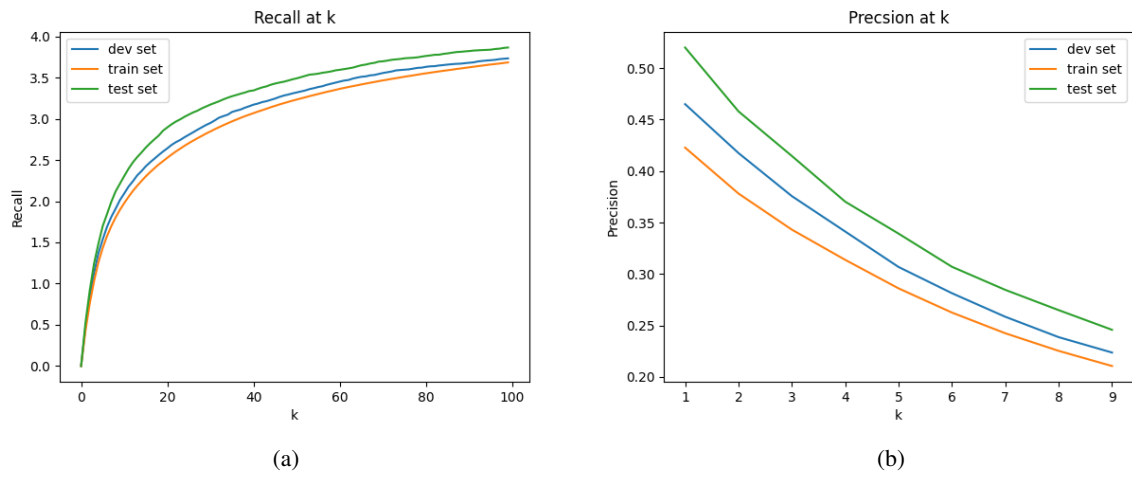


Figure 6: Precision and recall curves where the train set is sampled in samples of 1000 images.

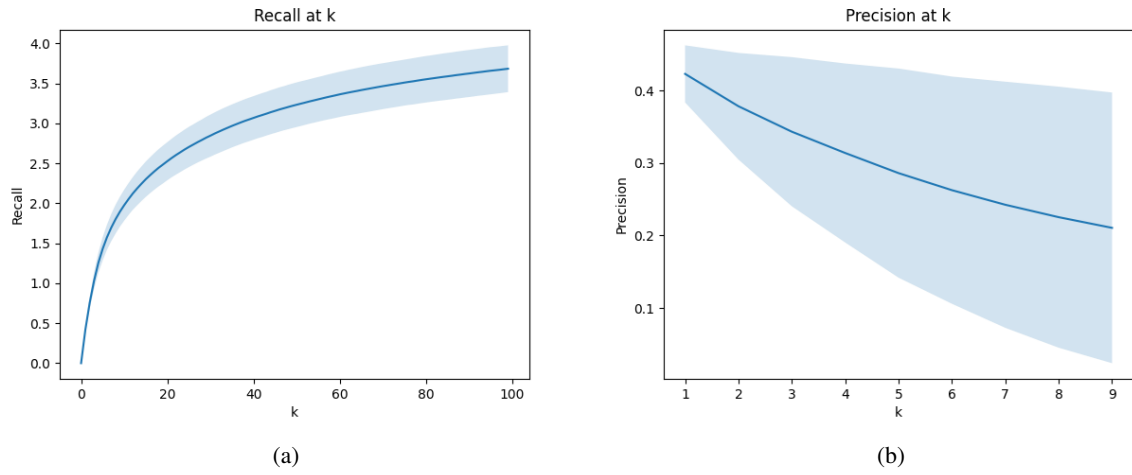


Figure 7: Average precision and recall curves with one standard deviation.

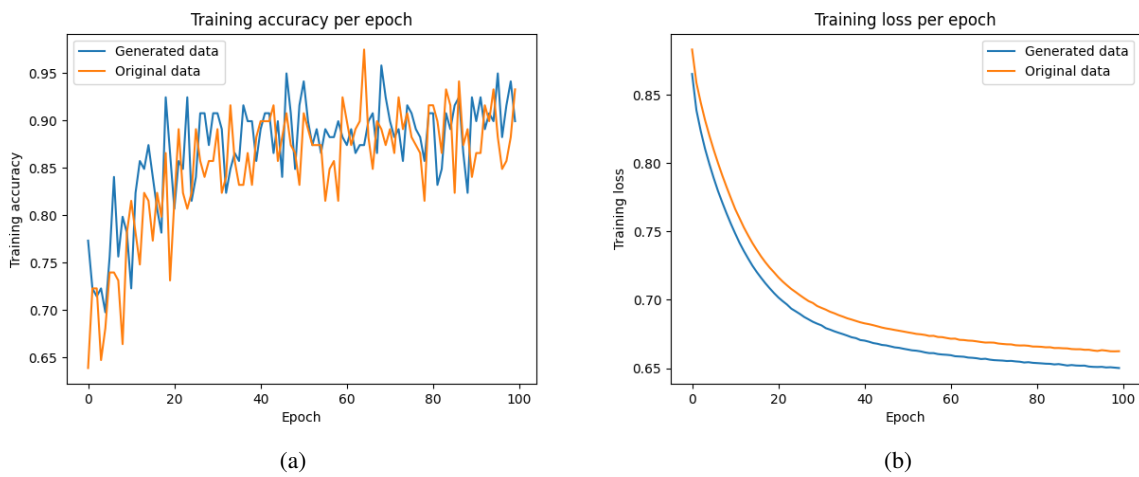
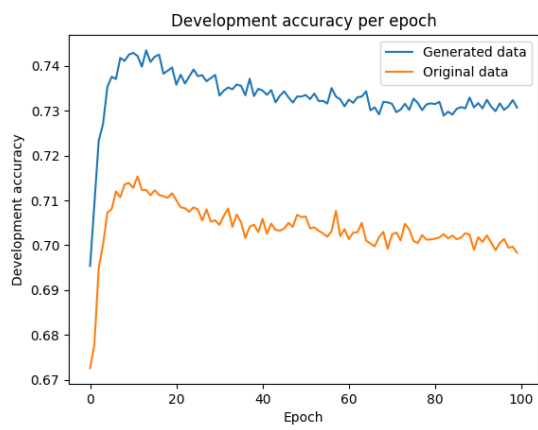
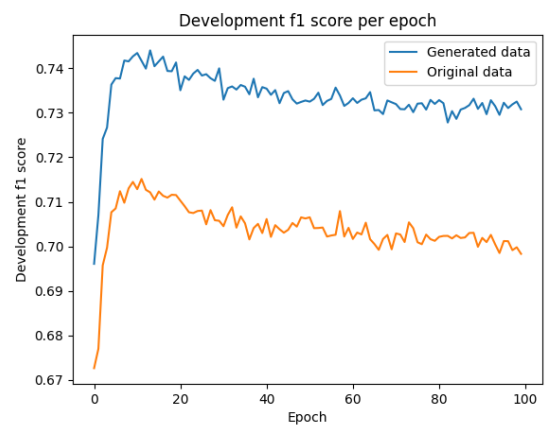


Figure 8: Performance on the training set during training.



(a)



(b)

Figure 9: Performance on the development set after each epoch.