Module: C5 – Final Presentation

Project: **Image Captioning**

Coordinator: E. Valveny

Team 8: G. Grigoryan,V. Heuer, P. Zetterberg

Table of Contents

Content Overview

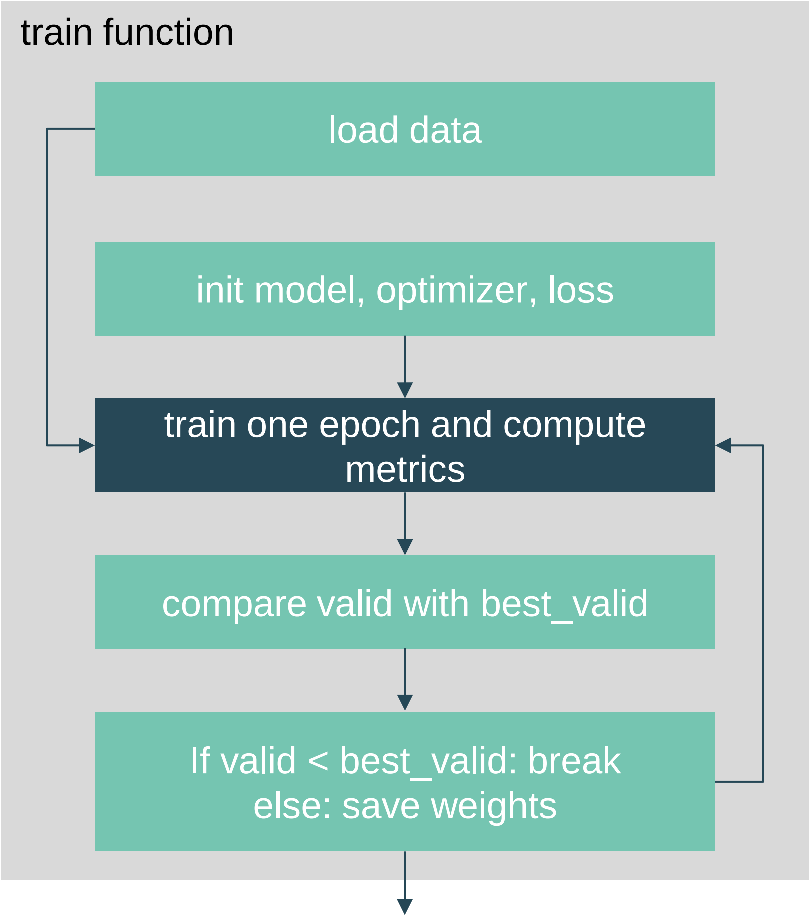
- Explored image captioning improvements over three weeks.
- **Week 1:** Modified baseline by changing encoder, decoder, and text representation.
- **Week 2:** Evaluated ViT-GPT2 and LLM-based models with fine-tuning strategies.
- **Week 3:** Generated synthetic images with Stable Diffusion to improve model performance.
- **Focus:** Analyze the impact of architecture and data augmentation on captioning quality.

[1] https://www.soccer-net.org/challenges/2025

Week 1: Train and Evaluate the Baseline Model – Experimental Setup.

Experimental Setup

- For training the model we decided to split the functionality into train and into train_one_epoch.
- This allows us to controll the data loading, logging, validation and early stopping seperate from the actual training.
- We used wandb for logging the metrics.
- We decided to use following hyperparameters:
- Optimizer: Adam
- Lr: 1e-3
- Batch Size: 8
- Loss: Cross Entropy Loss
- Epochs: 10
- Also we implemented a teacher forcing method that with a value of 0.5.



Week 1: Train and Evaluate the Baseline Model – Results.

Two very different dishes with the same prediction: „<SOS>Coase "

Results of Baseline Model		
metric	before training	10 epochs
BLEU-1	0.0001	0.0012
BLEU-2	0.0000	0.0000
ROUGE-L	0.0000	0.0024
METEOR	0.0012	0.0012
training time		~ 8 hrs.

Analysis

- **Training limitations:** Despite using an RTX3090, long training times limited extensive hyperparameter searches.
- **Performance overview:** Slight BLEU-1 and ROUGE-L gains after 10 epochs; BLEU-2 and METEOR remained unchanged, indicating issues with coherent phrasing and semantic accuracy.
- **Model analysis:** Improvements suggest basic word pattern learning, but the model still struggles with meaningful word combinations and subsequence generation.
- **Next steps:** Likely underfitting; better results may require more training, hyperparameter tuning, and advanced architectures like attention mechanisms.

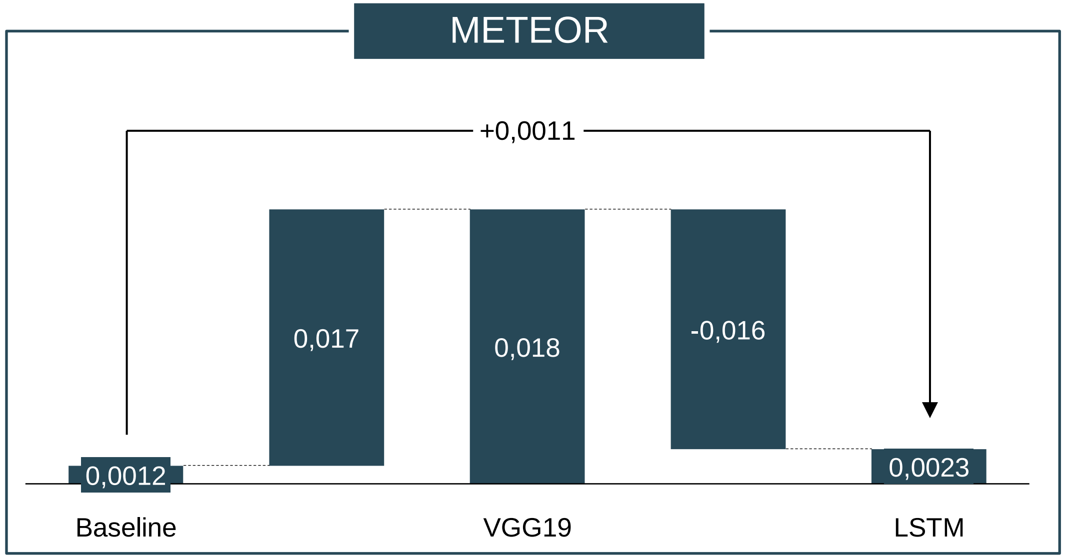
Week 1: Changing Encoder to VGG19 and Decoder to LSTM

Results of VGG19 Model on the test set			Results of LSTM Model on the test set		
metric	before training	after 10 epochs	metric	before training	10 epochs
BLEU-1	0.0017	0.0007	BLEU-1	0.0004	0.0019
BLEU-2	0.0007	0.0000	BLEU-2	0.0000	0.0003
ROUGE-L	0.0011	0.0007	ROUGE-L	0.0010	0.0042
METEOR	0.0026	0.0004	METEOR	0.0011	0.0023
training time		~ 8 hrs.	training time		~ 8 hrs.

Analysis

- **Encoder change (ResNet to VGG19):** Switching to VGG19 improved performance over ResNet but training remained slow initially.
- **Training issues with VGG19:** Despite improvements, the model showed very low metric values, suggesting it was effectively learning from scratch with few epochs.
- **Trainer optimization:** Introducing a faster trainer reduced training time significantly but did not immediately solve the low performance issue.
- **Decoder change (GRU to LSTM):** LSTM outperformed GRU slightly and led to better BLEU-1 and ROUGE-L scores, indicating improved word and subsequence modeling.
- **Overall decoder impact:** LSTM helped with longer-range dependencies and context but BLEU-2 and METEOR scores stayed low, pointing to a need for further tuning and training.

Summary Week 1: Despite extensively training our models and experimenting with various techniques, we were unable to generate meaningful captions, highlighting the challenges in achieving coherent and accurate text generation.



Two very different dishes with the same prediction: „<SOS>Coase "

Baseline

'<SOS>Coase '

'<SOS>Corled Soufffins with Sauce<EOS>'

'<SOS>Coailled Breen '

'<SOS><SOS>…Egu2<SOS><SOS><SOS><SOS> '

VGG19

'<SOS>Saasted a'

'<SOS>Shice'

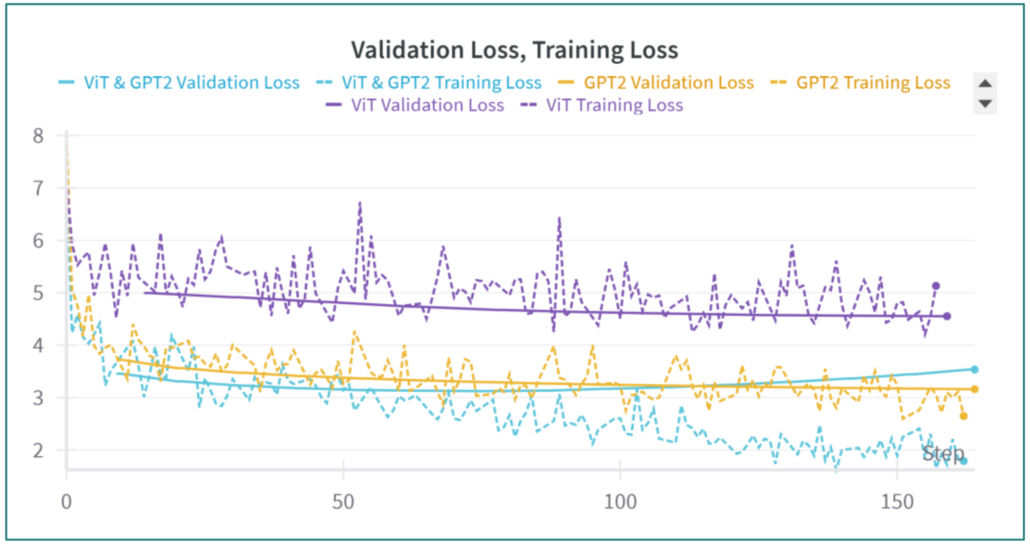
'<SOS>Saae'

'<SOS >Shesteeand Shorted Potato and '

Performance of Different Models

metric	Baseline	VGG19	LSTM
BLEU-1	0.0012	0.0376	0.0019
BLEU-2	0.0000	0.0000	0.0003
ROUGE-L	0.0024	0.0414	0.0042
METEOR	0.0012	0.0182	0.0023

Week 2: Fine Tuning the Model with Different Freezing Strategies



Method

- We finetuned the model in order to increase ist performance metrics.
- We used optuna and wandb to search for optimal hyperparameters and optimal experiment tracking.
- We fine tuned using three different set ups: ViT and GPT2 unfrozen, ViT frozen, GPT frozen.
- Fine tuning ViT & GPT2 at the same time yoields the best performance

- **Training loss** consistently decreases and ends lower than the others.
- **Validation loss** follows the training curve quite closely, with a mild upward trend at the end.
- **Takeaway:** This setup performs best in terms of generalization. The low gap between train and val loss suggests good learning without strong overfitting.

- **Training loss** is reasonably low and stable.
- **Validation loss** plateaus and even slightly increases, though not drastically.
- **Takeaway:** GPT2 has some flexibility to adapt, but it's constrained by the frozen ViT embeddings. Still performs decently — better than fine-tuning only ViT.

- **Training loss** is unstable and remains quite high.
- **Validation loss** is the worst of the three and trends downward only slightly.
- **Takeaway:** Freezing GPT2 seems to limit performance significantly. ViT alone can’t compensate, likely because the language generation component isn’t adapting at all. Poor generalization, and the model struggles to minimize loss.

Week 2: Analysis of the Different Freezing Strategies while Finetuning.

Overall Results of Last Week and Current Week						
metric	Last week‘s baseline 10 epochs	VGG19 10 epochs	LSTM 10 epochs	ViT & GPT2 11 Epochs	ViT 11 epochs	GPT2 11 epochs
BLEU-1	0.0012	0.0376	0.0019	0.1019	0.0654	0.1009
BLEU-2	0.0000	0.0000	0.0003	0.0362	0.0067	0.0301
ROUGE-L	0.0024	0.0414	0.0042	0.1261	0.1055	0.0986
METEOR	0.0012	0.0182	0.0023	0.1079	0.0840	0.0775

Comparison

- **Transformer models:** ViT & GPT-2 outperformed all previous models across all metrics.
- **Partial fine-tuning:** Fine-tuning only ViT or GPT-2 showed gains but was less effective than full model tuning.
- **Baseline comparison:** VGG19 and LSTM baselines performed poorly with minimal BLEU and ROUGE scores.
- **Data quality issue:** Overall results are limited, likely due to noisy and inconsistent dataset labels and images.

Week 2: Qualitative analysis of Fine Tuning VIT-GPT2

- GT: Matzo Ball Soup
- VIT-GPT2 Pretrained: "a bowl of **soup** and a glass of wine"
- VIT-GPT2 Fine-Tuned: "Bibimbap at Home (Baba Ghanoush) with Garlic-Lime **Matzo Balls** and Leeks (Baba Bar Bihonada)"
- VIT Fine-Tuned: "a pot roast with garlic bread and herbs"
- GPT2 Fine-Tuned: "Beef Stew with Potatoes and Meat**balls** (Beef Stew in Guisado) and Parsley **Soup** (Beiguza) Bisque Blanco) Bisque Blanco (Beiguza Bisque Verde) Bisque Bl"

- GT: Smoked Salmon with Black Pepper Potato Chips
- VIT-GPT2 Pretrained: "two pieces of bread with cheese and tomatoes"
- VIT-GPT2 Fine-Tuned: "**Smoked** Trout with Horseradish Cream and Citrus Caviar Filling (Figs au romaine) Grapefruit Fritters with Grapefruit Compote and Horseradish Cream Toasts Grapefruit Compote au rom"
- VIT Fine-Tuned: "a pizza with basil and cheese on a plate"
- GPT2 Fine-Tuned: "**Smoked Salmon** Smørrebrøders with Horseradish Cream Sauce and Pickled Pear Tarts with Feta Purée and Pickled Pear Tartar Sauce. Caviar and Pickled Pear Tarts with Pickled Pear T"

- GT: Bacon Swiss Burgers with Tomato and Avocado
- VIT-GPT2 Pretrained: "a sandwich with lettuce, **tomato**, and cheese"
- VIT-GPT2 Fine-Tuned: "Mushroom **Burger**s with Bologna and Cheddar Sliders on Polenta ""Tacos"" with Curry Barbecue Sauce and Green Onions ""Onions ""Onions ""Onions ""Onions ""Onions ""Onions """
- VIT Fine-Tuned: "a ham**burger** with cheese and onions on a bun"
- GPT2 Fine-Tuned: "Bison **Burgers** with Beet and Chipotle Slaw (Ensalada de Burgers en salchicha) and Chipotle-Cheddar Red Sauce (Ensalada de Burgers en salchicha) and Chipotle-C"

Week 2: Qualitative analysis of Fine Tuning Llama

- GT: Parisian
- VIT-Llama 1B Fine-Tuned: "ovioviovi;oviovioviovioviovioviovioviovioviovioviovioviovioviovioviovioviovioviovio-vioviovioviovioviovioviovioviovioviovioviovioviovioviovi"
- VIT-Llama 3B Fine-Tuned: "project and and the C, for the same &"

- GT: Spicy Curry Noodle Soup with Chicken and Sweet Potato
- VIT-Llama 1B Fine-Tuned: "settled TOovioviovioviovioviovioviovioviovioviwoviovissnsov ioviovioviovioviovioviovioviovivoviovioviovioviovioviovioviovioviovioviowsoviovioviovii"
- VIT-Llama 3B Fine-Tuned: "nuclear"

Conclusion

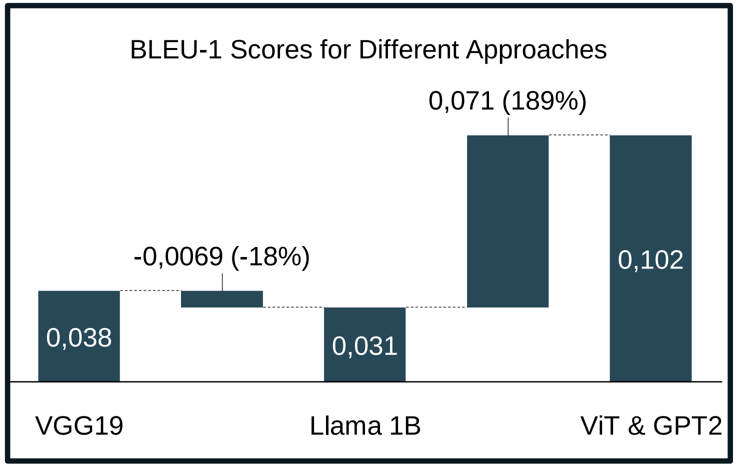
✓ Both models generate mostly incoherent, repetitive, or non-linguistic outputs — no meaningful captions are produced.
✓ Results vary between runs making outputs unstable and unreliable for evaluation.
✓ The models were not trained with chat-style prompts, which may be causing mismatches when used in captioning pipelines.
✓ Visual embeddings from ViT likely don't align well with LLaMA's language space, leading to unconditioned generation.
✓ Metrics are meaningless here; qualitative analysis confirms near-complete caption failure.

Week 2 Summary: Modern Architectures Yield More Coherent Captions, But Strong Results Remain Elusive.

Overall Results of Last Week and Current Week									
metric	Last week's baseline fine tuned	VGG19 fine tuned	LSTM fine tuned	ViT & GPT2 fine tuned	ViT fine tuned	GPT2 fine tuned	Gemma 12B	ViT & Llama 1B fine tuned	ViT & Llama 3B fine tuned
BLEU-1	0.0012	0.0376	0.0019	**0.1019**	0.0654	0.1009	0.0119	0.0307	0.0189
BLEU-2	0.0000	0.0000	0.0003	**0.0362**	0.0067	0.0301	0.0030	0.0057	0.0049
ROUGE-L	0.0024	0.0414	0.0042	**0.1261**	0.1055	0.0986	0.0662	0.0430	0.0315
METEOR	0.0012	0.0182	0.0023	**0.1079**	0.0840	0.0775	0.0381	0.0414	0.0324

- GT: Smoked Salmon with Black Pepper Potato Chips
- VIT-GPT2 Pretrained: "two pieces of bread with cheese and tomatoes"
- VIT-GPT2 Fine-Tuned: "**Smoked** Trout with Horseradish Cream and Citrus Caviar Filling (Figs au romaine) Grapefruit Fritters with Grapefruit Compote and Horseradish Cream Toasts Grapefruit Compote au rom"

- GT: Parisian
- VIT-Llama 1B Fine-Tuned: "ovioviovi;oviovioviovioviovioviovioviovioviovioviovioviovioviovioviovio vioviovioviovio-vioviovioviovioviovioviovioviovioviovioviovioviovioviovi"
- VIT-Llama 3B Fine-Tuned: "project and and the C, for the same &"



Week 3: First tests

Method

- After the first installation and testing, we thought about how our prompts should look like so the models are able to generate suitable images which are similar to the original dataset – the prompts were generated with Chat-GPT 4o
- Therefore we tried out **three** different styles of description:
  - ***Clear visual description of the dish***
  - ***Cuisine-Style Title and short description of the dish***
  - ***Cuisine-Style Title***
- Good (left) and bad (right) examples with their prompts are shown below; images were generated with the Stable-diffusion XL (default settings)

Clear description

grilled chicken in a creamy spiced tomato sauce, photorealistic, food styling, high detail

marinated raw fish with citrus juice, onions, and cilantro, photorealistic, food styling, high detail

Cuisine + Description

Ink pasta midnight. Black squid ink pasta with garlic and shrimp, photorealistic, food styling, high detail

Banana fire on pancake dune. Pancakes with caramelized bananas, photorealistic, food styling, high detail

Cuisine Title

Chashu Ramen, photorealistic, food styling, high detail

Duck à l'Orange, photorealistic, food styling, high detail

Conclusion

- The first bigger experiments showed that the model can generate very realistic images of food, if the prompt is precise enough or if the model knows the cuisine-style words
- **But** it was clear for us that if we want to be sure that all generated images are realistic we need some descriptions rather than just cosine titles of the dishes

Week 3: Exploring Parameters of Stable Diffusion Models

Method

Models tested:
Stable-diffusion 2.1, Stable-diffusion 2.1 turbo, Stable-diffusion XL, Stable-diffusion XL turbo

Parameters tested (loop):
- DDPM and DDIM,
- with and without negative prompting,
- CFG: [7, 9, 12, 17, 22, 27, 32],
- Steps: [25, 50, 75, 100, 125, 150, 175]

Positive Prompt:
"Chocolate lava cake in a minimalist Scandinavian setting, photorealistic, food styling, ultra-detailed"

Negative Prompt:
"blurry, deformed, text, watermark, cartoon"

Best Model

Model:
Stable-diffusion XL turbo

Parameters:
- DDPM
- with negative prompting
- CFG: 9
- Steps: 50

Inference time:
- 4.5 seconds per image

Observation

- Some outputs had an unnatural lighting effect — the image appeared to be overly shiny or reflective
- This issue was partially mitigated by using a well-crafted negative prompt and by adding more steps
- Compared to sd-turbo, this model delivers noticeably better image quality overall while being still very fast

Week 3: Exploring Parameters of Stable Diffusion Models

without negative prompt | **with negative prompt**

Stable-diffusion XL

without negative prompt

cfg / steps	20	60
5		
13		

with negative prompt

cfg / steps	20	60
5		
13		

Stable-diffusion XL turbo

without negative prompt

cfg / steps	20	60
5		
13		

with negative prompt

cfg / steps	20	60
5		
13		

Week 3: Problem Identification, Research Question, Method and Pipeline

Problem Identification

Model Performance: Fine-tuned ViT and GPT-2 performed best, but generated captions are often inaccurate or overly generic.

Dataset Style Issues: Captions in the dataset are short, abstract, and cuisine-styled, making it hard for the model to learn detailed visual-text relations

Data Quality Problems: The dataset contains noisy samples like recipe books or screenshots

Signs of Overfitting: Strange outputs like repeated punctuation or meaningless tokens suggest overfitting or confusion from noisy data.

GT: Harissa-Crusted Tri-Tip Roast

VIT-GPT2 Fine-Tuned: "Tri-Tip with Chimichurri Sauce and **Roast**ed Peperonata Salad (Maiale Adleru) with Israeli Salsa Verde (Maiale Adleru) and Tomato-Yogurt Sauce (Maiale""

Research question

1. Can the addition of synthetic food images improve the performance of our captioning model ?
2. To what extent can synthetic data help reduce overfitting and increase the quality of the predictions ?
3. And what is the optimal amount of synthetic data to include in the training set for the best results ?

Method

Synthetic Data Generation: Used Chat-GPT 4o to create 5.4k (50% of training set) prompts for food images, aligning with the cuisine and descriptive style of the original dataset

Image Creation: Generated 5,400 synthetic images based on these prompts using Stable Diffusion, then cleaned the prompts to create matching titles

Model Training: Fine-tuned the best previous model (ViT + GPT-2) by adding synthetic data in increments of 10%-50% alongside the original training set

Additional Experiments:
1. Also trained the model purely on synthetic data then evaluating on the original test set
2. Trained validated and evaluated the model purely on synthetic data

Pipeline

Request → Chat-GPT 4o → Prompt

Prompt → Cleaned Titles

Prompt → Stable-diffusion XL turbo

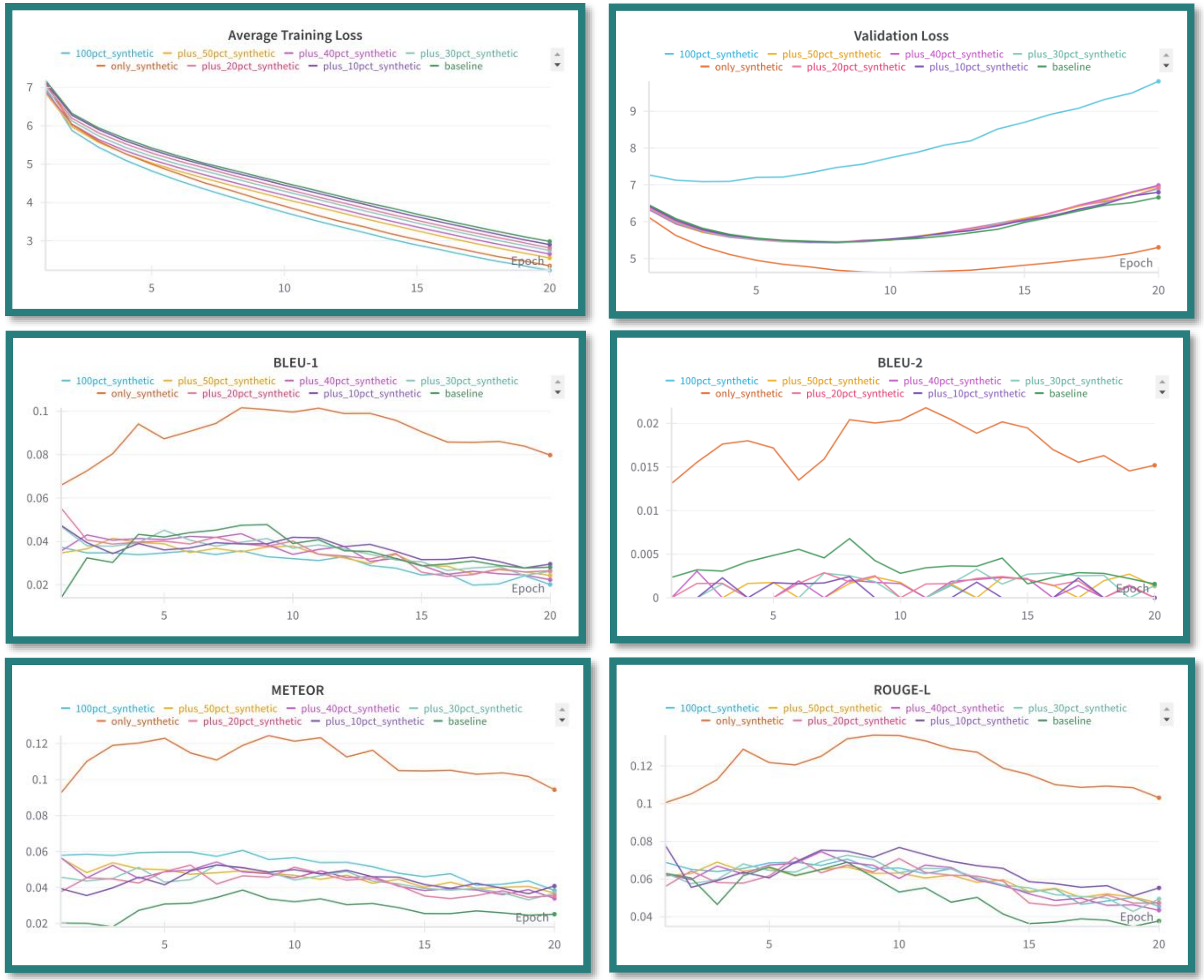
Stable-diffusion XL turbo → images

Cleaned Titles → Captioning Model

images → Captioning Model

Week 3: Quantitative Results and Analysis

Quantitative Results - Training



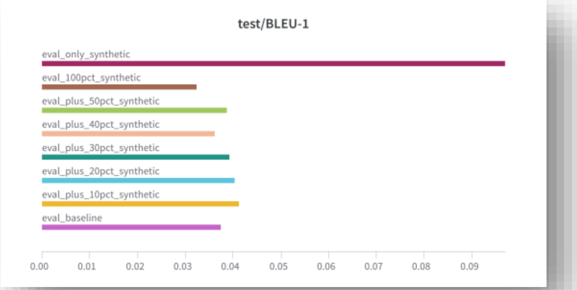
Quantitative Results - Testing

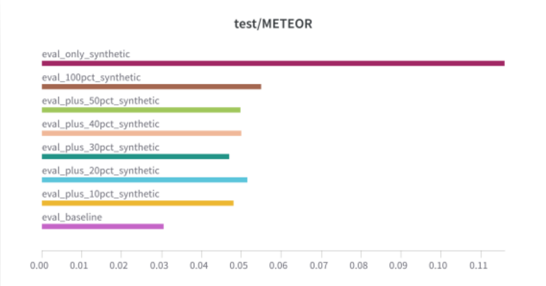
Quantitative An

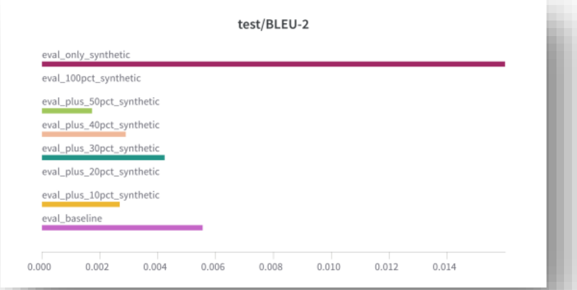
- The results show a noticeable improvement across all evaluation met
- The model trained and tested on only synthetic data achieved the hig
- These models also show the lowest training and validation loss, sugg
- This supports the idea that the original dataset's labels may be a limit provides cleaner and more learnable supervision
- Training with synthetic data resulted in small but consistent improvem
- The "best" came from models trained with 20–40% synthetic data, wh
- BLEU scores showed no significant gains, and the 100% synthetic mo signs of overfitting
- Overall, synthetic data was beneficial when used in moderation

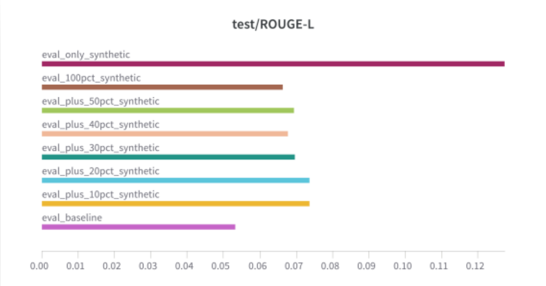
Week 3: Task e: Training the Captioning Model by only using synthetic data for the training, validation and test set – Quantitative Results (2) and Quantitative Analysis

Quantitative Results - Testing









Quantitative Analysis

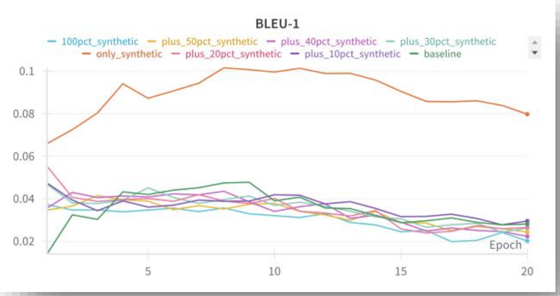
- The results show a noticeable improvement across all evaluation metrics compared to the mixed or original datasets
- The model trained and tested on only synthetic data achieved the highest scores in BLEU-1, BLEU-2, ROUGE-L, and METEOR
- These models also show the lowest training and validation loss, suggesting better learning stability and less overfitting
- This supports the idea that the original dataset's labels may be a limiting factor in model performance, and that this synthetic data provides cleaner and more learnable supervision

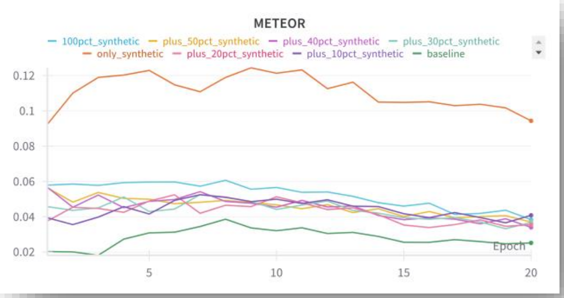
Week 3: Quantitative Results

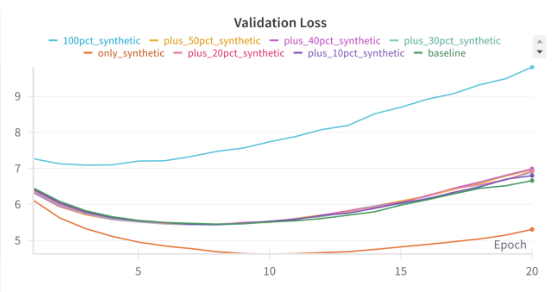
Idea & Method

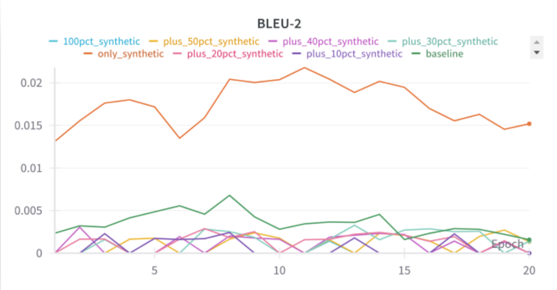
- Since we did not improve our model in comparison to last week, even with 5.400 realistic looking synthetic images, we thought of showing that neither the model nor the synthetic data is the problem but possibly the dataset or more specifically the Titles of the dishes or the ground truth
- Therefore we decided to not only train but also validate and test with only synthetic data
- The reason for that is, that the synthetic data or more specific the generated prompts or Titles are way more descriptive than the original ground truth
- Given this idea we created a whole new data set (split 80/10/10) only consisting of synthetic data consisting of 5.400 generated images and used the same model as well as the same parameters for the training and testing

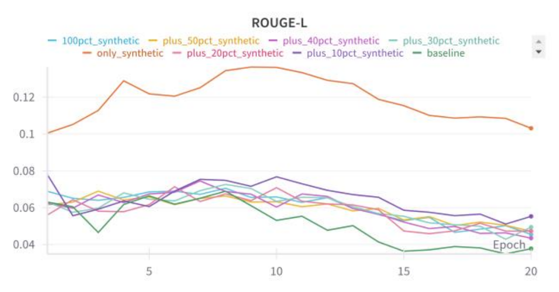
Quantitative Results - Training









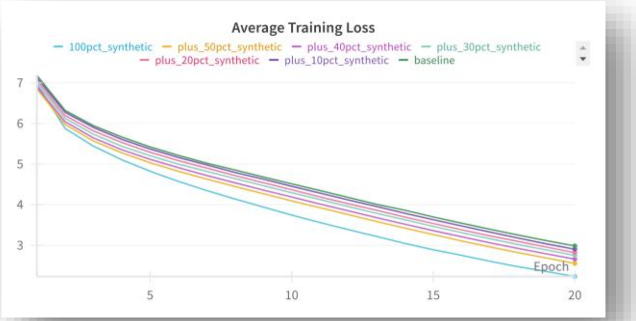


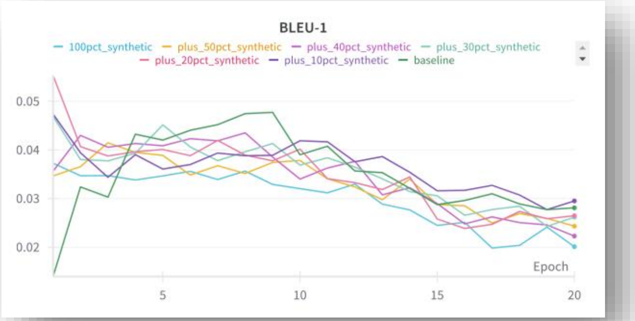
Week 3: Task e: Training the Captioning Model by adding synthetic data to the training set – Method and Quantitative Results (1)

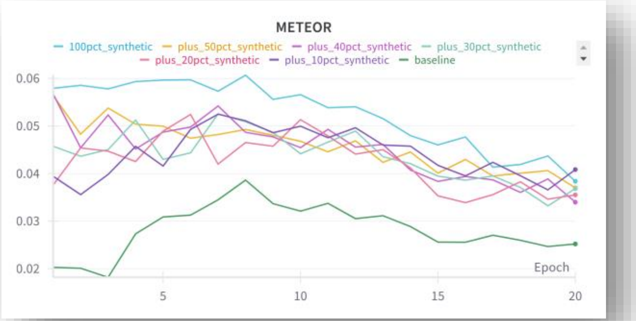
Method

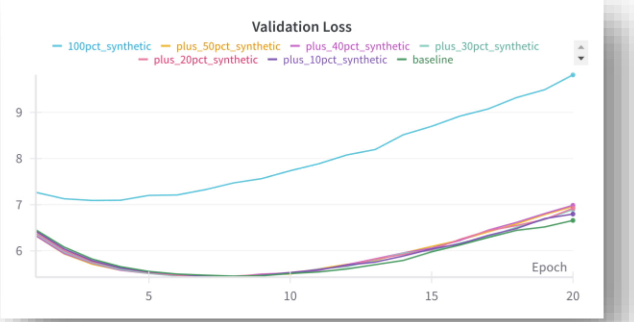
- After generating 5,400 synthetic samples, we focused on analyzing how different amounts of this synthetic data affect model performance by fine-tuning our best-performing ViT-GPT2 model from earlier weeks using the following setups:
- The original training set was extended with synthetic samples in steps of:
  - 10% (1,080 synthetic samples), 20%, 30%, 40%, 50% (5,400 synthetic samples) and only training with synthetic data (100%)

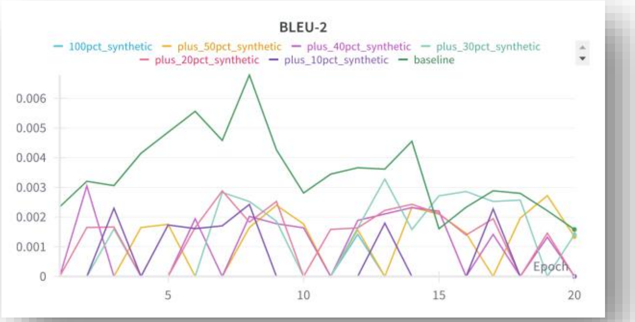
Quantitative Results - Training

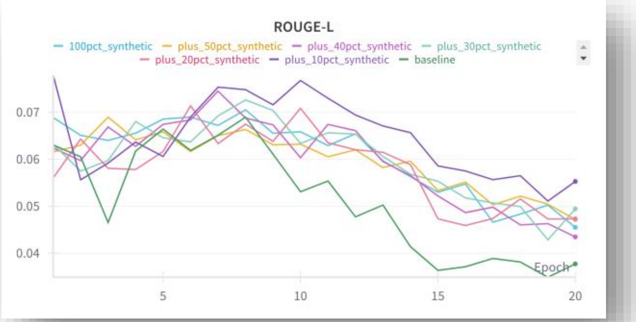






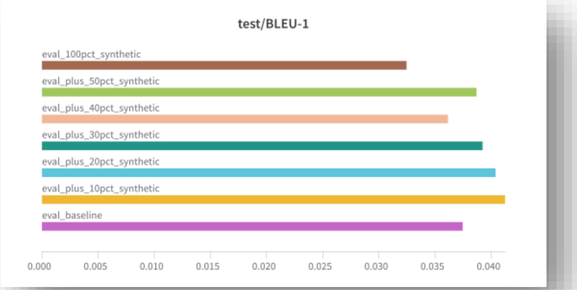


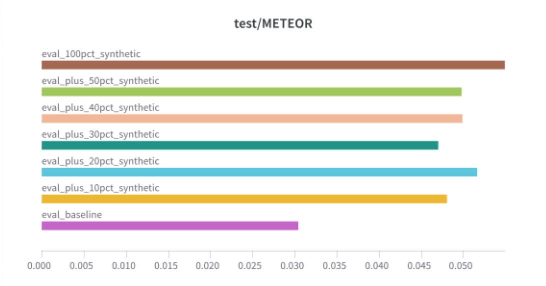


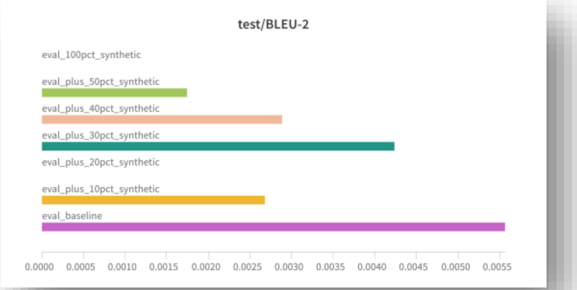


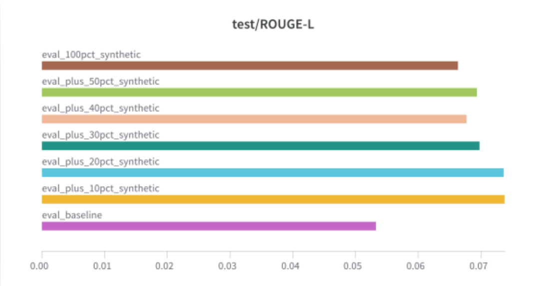
Week 3: Task e: Training the Captioning Model by adding synthetic data to the training set – Quantitative Results (2) and Quantitative Analysis

Quantitative Results - Testing









Quantitative Analysis

- Training with synthetic data resulted in small but consistent improvements, especially in ROUGE-L and METEOR.
- The "best" came from models trained with 20–40% synthetic data, which outperformed the baseline without overfitting to much
- BLEU scores showed no significant gains, and the 100% synthetic model performed worst in terms of generalization, showing clear signs of overfitting
- Overall, synthetic data was beneficial when used in moderation

Week 3: Task e: Training the Captioning Model by adding synthetic data to the training set - Qualitative Results and Analysis

- GT: Instant Pot Braised Lamb with White Beans and Spinach
- Baseline: - Beans and Beans with Beans Yog and Beans Yog Sauce Yog and Beans Yog Sauce Yogi Yogi
- Plus 10% synthetic: - Beans Lent with and Beans Pe and Beans Yog Sauce Yog Sauce Yog Sauce Yog Sauce Yog and Beans
- Plus 20% synthetic: - Beans Lent with and Beans Pe and Beans Yog Sauce Yog Sauce Yog Sauce Yog Sauce
- Plus 30% synthetic: - Meataf with Beans Goat and Beans Yog Sauce Yog Sauce Yog Sauce Yog Sauce Yog
- Plus 40% synthetic: -ed with and andils beef with andumin andumin paste garlico goatis greens herbs herbs
- Plus 50% synthetic: - Beefender with and Beans Lent and Beans Yog Sauce Yog Sauce Yog Sauce Yog Sauce
- 100% synthetic: ris fried with and sauce lemonah lemonah lemonah vinegar olive oil lemonah oil garlic herbs olive

- GT: Hibiscus Tea Sorbet
- Baseline: - Ice Ice with Ice,,,,,,,,,,,,,,,
- Plus 10% synthetic: -rozen Ice with Ice and Ice Creamustaze Cream Ice Cream Ice with Ice Cream Ice Cream Ice
- Plus 20% synthetic: -ime ice with and- ice creambet dries ice with ice and ice cream drizzle ice
- Plus 30% synthetic: - Ice Ice with Ice,,,,,,,,,,,,,,,
- Plus 40% synthetic: - Ice Ice with Ice,,,,,,,,,,,,,,,
- Plus 50% synthetic: --------------------
- 100% synthetic: and ice with and creamzzle ice cream strawberryousse ice creamzzle creamzzle creamzzle creamzzle cream

Qualitative Analysis

- The baseline model struggles with repetition and incoherent phrases like “Yog Sauce Yog Sauce Yog Sauce”
- As synthetic data is added, predictions become slightly more relevant, with food-related terms appearing more often
- By 30–50% synthetic, the model starts generating ingredients like “beef,” “beans,” and “herbs,” which loosely match the ground truth
- For the second image, the 100% synthetic model outputs terms like “ice cream” and “strawberry mousse,” which, while not accurate, are closer in meaning to “sorbet"
- In short, synthetic data helps reduce repetition and brings predictions closer to the right concept, but full accuracy and fluency remain a challenge

Week 3: Task e: Training the Captioning Model by only using synthetic data for the training, validation and test set - Qualitative Analysis

- GT: Muesli with apple slices and almond butter
- Only synthtetic: asted and apple with and oats oats oats cinnamon cinnamonins apples cinnamonins oats cinnamonins apples cinnamonaze

- GT: Tofu eggless sandwich with tomato and spinach
- Only synthtetic: illed sandwich tomato with and cheesezz cheesezz cheeseioli cheeseolazz breadzz cheeseioli

Qualitative Analysis

- While the predictions from the model trained on only synthetic data are not perfect, they show a clear shift towards structured and food-related phrases.
- In the first example, terms like "oats", "apples", and "cinnamon" appear frequently, showing the model's focus on key ingredients even if the phrasing is repetitive
- In the second case, the prediction includes "sandwich", "tomato", and multiple variations of "cheese", indicating the model's understanding of food components despite some unnatural endings like "cheesezz" or "cheeseolazz"
- These outputs suggest that while the model may over-repeat or invent tokens, the core food concept is preserved more consistentl - highlighting that synthetic captions help the model focus on relevant ingredients, though at the cost of good linguitics

Week 3: Task e: Training the Captioning Model by using synthetic data - Conclusion

Conclusion

- We tried two strategies using synthetic data to improve our captioning model
- One was to add synthetic samples to the original training set in different amounts
- The other was to train and evaluate the model using only synthetic data

- The experiments showed that adding synthetic data can help reduce overfitting and sometimes improve performance. Small and medium amounts seemed to give the most balanced results

- Interestingly, the model trained only on synthetic data (with no original samples) outperformed all other models across all key evaluation metrics
- This was surprising, especially because the total number of synthetic samples was still lower than the full original dataset
- One possible reason is that the synthetic titles are more descriptive and easier to learn from compared to the original dataset

- These results suggest that the quality and clarity of the captions - whether prompts or ground truth – can play a big role in how well the model learns

- This means that improving or rewriting the original titles could also be a promising direction

Week 3 Summary: Synthetic Data Shows Potential, with Caption Quality Likely Playing a Key Role.

Model and Parameter Tuning

Models tested:

Stable-diffusion 2.1
Stable-diffusion 2.1 turbo
Stable-diffusion XL
Stable-diffusion XL turbo

Parameters tested (loop):

DDPM and DDIM
With and without negative prompting
CFG: [7, 12, 17, 22, 27, 32]
Steps: [25, 50, 75, 100, 125, 150, 175]

cfg / steps	20	60
5		
13		

sdxl-turbo

DDPM

with negative prompt

steps=50

cfg=9

Pipeline

Request → Chat-GPT 4o → Prompt

Prompt → Cleaned Titles

Prompt → Stable-diffusion XL turbo

Cleaned Titles → Captioning Model

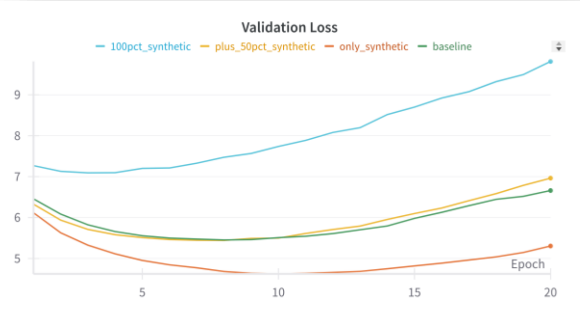
Captioning Model

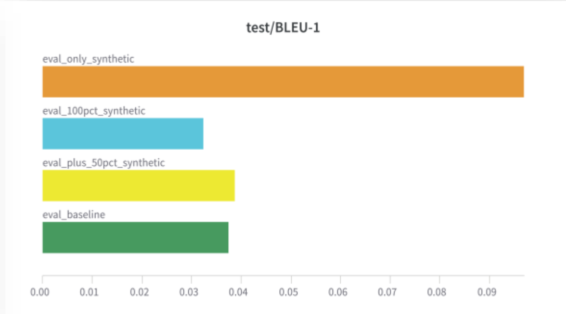
Method

We generated 5.4k synthetic samples to match our 10.8k training set and fine-tuned ViT-GPT2 with 10–50% added synthetic data.

We also trained two models on synthetic data only for comparison, then testing on original and synthetic data

Results





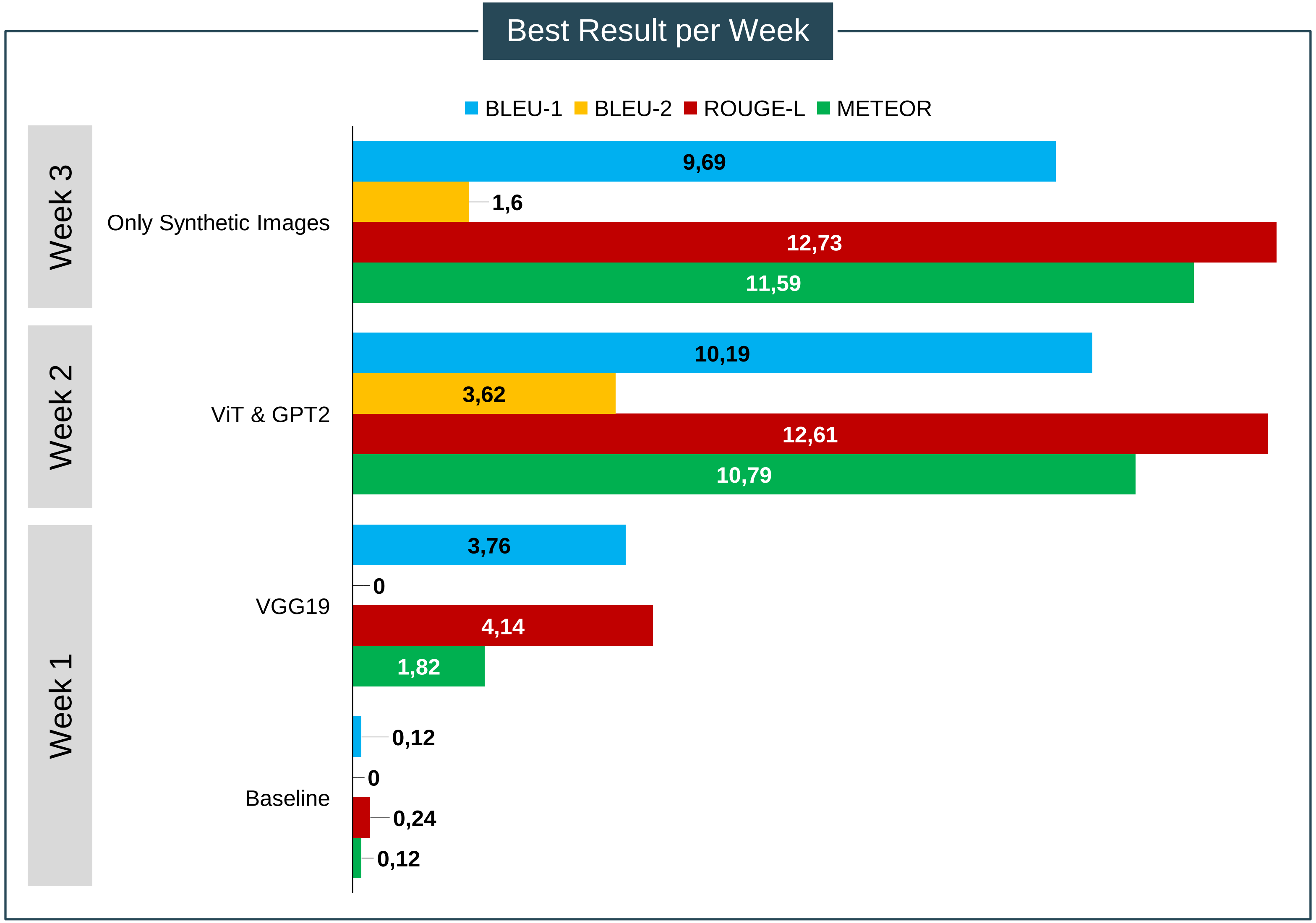
- Like in previous weeks overfitting started around the 10. Epoch and was still pretty strong
- Only minor improvements by adding synthetic data
- Metrics-wise the model trrained and tested only on synthetic data performed by far the best – this may be because the captions are more descriptive and the data is cleaner

Predictions

Plus 50% synthetic:
- Beefender with and Beans Lent and Beans Yog Sauce Yog Sauce Yog Sauce Yog Sauce

Train/test on sythetic:
asted and apple with and oats oats oats cinnamon cinnamonins apples cinnamonins oats cinnamonins apples cinnamonaze

Comparison of Results.



- **Baseline:** Almost no performance.
- **VGG19 encoder:** Slight overall improvement, no gain on BLEU-2.
- **ViT + GPT-2:** Drastic performance boost.
- **Diffusion (only synthetic data):** Best results by full replacement of training data.

- **VGG19 as encoder:** Better visual features than baseline, slight gain.
- **ViT + GPT-2:** Stronger global image understanding and language generation, major gain.
- **Synthetic images:** Increased data diversity, slight gain.