Master in
Computer Vision
Barcelona

Module:	C5 – Task 4
Project:	**Image Captioning II**
Coordinator:	E. Valveny
Team 8:	G. Grigoryan,V. Heuer, P. Zetterberg

Task 1.1: Direct Evaluation Using Pretrained ViT and GPT2 from Hugging Face

[1]

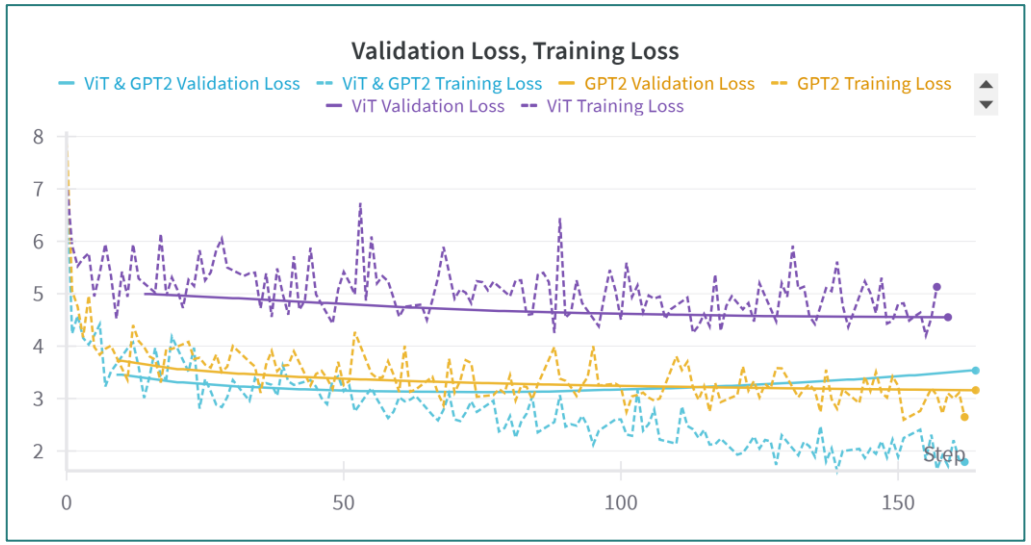
```python
import torch
from transformers import VisionEncoderDecoderModel, ViTImageProcessor, AutoTokenizer

# Load the Model
model = VisionEncoderDecoderModel.from_pretrained('nlpconnect/vit-gpt2-image-captioning')
feature_extractor = ViTImageProcessor.from_pretrained('nlpconnect/vit-gpt2-image-captioning')
tokenizer = AutoTokenizer.from_pretrained('nlpconnect/vit-gpt2-image-captioning')
model.config.pad_token_id = tokenizer.eos_token_id
model.to(torch.device("cuda"))
```

Method

- This week‘s first challenge is to use hugginface‘s ViT and GPT2 implemetation to infer image captions on the same dataset (Food Ingredients and Recipes Dataset [2]) as in the last week.
- We used last week‘s code as baseline. The implementation of the new model is shown in the image.
- We used the same custom data loader class with the split (0.8, 0.1, 0.1)
- We infered the captions on the MCV server‘s GTX 3090 and computed BLEU-1, BLEU-2, Rouge-L and METEOR.

[1]: https://www.linux-magazin.de/news/hugging-face-baut-open-source-repo-von-deepseek-r1/
[2]: https://www.kaggle.com/datasets/pes12017000148/food-ingredients-and-recipe-dataset-with-images

Task 1.2: 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.

Task 1.2: Analysis of Key Metrics

BLEU-1 & -2:

- ViT&GPT2 and GPT2-only fine-tuning achieved comparable BLEU-1 & -2 scores, indicating that updating the language model is key to performance.
- ViT-only fine-tuning underperformed significantly, suggesting limited contribution of vision-only updates to token-level accuracy.
- Most performance gains occurred within the first 3-5 epochs, with minor improvements beyond that, indicating early convergence for text generation quality.

METEOR:

- ViT-only and GPT2-only fine-tuning showed similar performance, with a steady but gradual improvement across all epochs.
- ViT&GPT2 fine-tuning outperformed both single-component fine-tunings, highlighting the benefit of joint optimization.
- The most significant gains for ViT&GPT2 occurred in the early epochs, with the curve flattening out after initial rapid improvement.
- Peak METEOR score was achieved at epoch 11 during the ViT&GPT2 run, indicating an optimal balance between vision and language learning.

ROUGE-L:

- ViT-only and GPT2-only fine-tuning resulted in comparable performance, with consistent, gradual improvements across epochs.
- oint fine-tuning (ViT&GPT2) clearly outperformed the individual fine-tuning approaches.
- The majority of gains in ViT&GPT2 occurred in the first few epochs, followed by a plateau, indicating early convergence.
- Best ROUGE-L score was observed at epoch 11 during ViT&GPT2 fine-tuning, aligning with METEOR trends.

Task 1.2: Analysis of Fine Tuning GPT2.

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

- The ViT & GPT2 model significantly outperformed all previous models across all metrics.
- Fine-tuning only GPT2 achieved comparable BLEU-1 scores, but lagged slightly on ROUGE-L and METEOR.
- Fine-tuning only ViT also showed solid gains but was less effective than full model tuning.
- The baseline models, including VGG19 and LSTM, performed poorly in comparison, with minimal BLEU and ROUGE scores.
- Overall, the transformer-based architectures proved substantially more effective than the traditional CNN-RNN setups from last week.
- The results are still not impressive. We assume the dataset is not clean, as recipe naming is highly subjective and some datapoints include images of recipe books instead of actual food.

Task 1.2: Qualitative analysis of Fine Tuning VIT-GP2 (1)

- 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"

Task 1.2: Qualitative analysis of Fine Tuning VIT-GP2 (2)

- GT: Harissa-Crusted Tri-Tip Roast
- VIT-GPT2 Pretrained: "a plate of food with meat and vegetables "
- 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"
- VIT Fine-Tuned: "a meat and vegetable pizza with sauce and cheese"
- GPT2 Fine-Tuned: "Pork Shoulder Steak with Sesame-Chipotle Mayonnaise and Asian-Style Mustard Sauce (Pork Tenderloin) with Asian-Caper Relish (Pork **Roast** With Sesame-Caper Relish"

Conclusion

- ✓ Despite low quantitative metrics (e.g., BLEU, METEOR), many model outputs are visually grounded and semantically plausible.
  This is because the captions in the dataset are often abstract, culinary-specific, or overly concise, which penalizes surface-level mismatch in string-based evaluation.
- ✓ The models demonstrate a solid understanding of visual concepts such as shape, layout, and high-level food categories (e.g., “a bowl of soup with vegetables and meat” for Matzo Ball Soup).
  This supports the hypothesis that the vision encoders have learned robust visual representations.

- ✓ Several predictions align well with how a human might describe a dish visually, even if the names differ from the reference titles (e.g., calling “Caramelized Clementines” “a plate of food with some fruit on it” is semantically acceptable given no visible label in the image).
- ✓ Long hallucinated captions, particularly from the fine-tuned GPT2 variants, reflect exposure to over-annotated training data.
  This model tends to overgenerate gourmet-style ingredient lists, which often deviate from the actual food depicted.

- ✓ The pretrained model often generates shorter, more visually descriptive captions that better match what’s observable in the image — suggesting that pretraining on broader, generic datasets may provide stronger generalization for visual grounding.

Task 2.1: Direct Evaluation Using Pretrained Multimodal Image to Text Model from Hugging Face

[1]

Method

- This week‘s second challenge is to use one of hugginface‘s multimodal models implemetations to infer image captions on the same dataset (Food Ingredients and Recipes Dataset [2]) as in the first week.
- We chose Google‘s 12B gemma version. Implementation in existing architecture was straight forward.
- We infered the captions on the MCV server‘s GTX 3090 and computed BLEU-1, BLEU-2, Rouge-L and METEOR.
- Even though we had capable GPUs, the inference on the test set computed for over 11 hours.

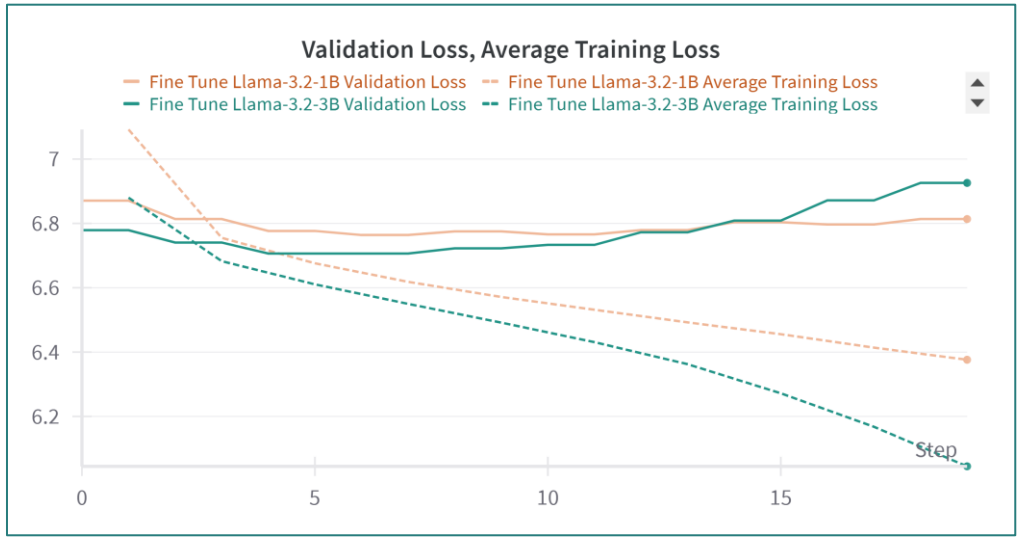
Results on Inference with Gemma

metric	Gemma 12B	ViT & GPT2 11 Epochs	Delta
BLEU-1	0.0119	0.1019	-0.0900 (-88%)
BLEU-2	0.0030	0.0362	-0.0332 (-92%)
ROUGE-L	0.0662	0.1261	-0.0599 (-48%)
METEOR	0.0381	0.1079	-0.0698 (-65%)
Runtime	~ 11 h		

The fine-tuned ViT & GPT2 model outperformed the pretrained Gemma 12B across all evaluation metrics. BLEU-1 and BLEU-2 scores from ViT & GPT2 were an order of magnitude higher, indicating much better n-gram precision.
This suggests that despite Gemma 12B’s size and general capabilities, domain-specific fine-tuning on the target dataset yields far superior results. It highlights the importance of task-specific adaptation over relying solely on large-scale pretrained multimodal models.

[1]: https://mashable.com/article/google-announcement-open-source-ai-model-gemma
[2]: https://www.kaggle.com/datasets/pes12017000148/food-ingredients-and-recipe-dataset-with-images

Task 2.2 – Fine Tuning of Llama 3.2-1B/3B Decoders Using LoRA.



Method

- We finetuned the Llama decoders using LoRA in order to increase ist performance metrics.
- We used optuna and wandb to search for optimal hyperparameters and optimal experiment tracking.
- We used following LoRA configs: r = 16, α = 32, dropout = 0.1
- Runtime was multiple hours per run.

- Initial Performance: The 3B model starts with a slightly lower validation loss than the 1B model, indicating better initial capacity for fitting the task.
- Stability: The 1B model shows a more stable curve with gradual improvement and plateauing, while the 3B model initially improves but then exhibits a sharp rise in validation loss after step 13.
- Overfitting in 3B: The steep increase in validation loss for the 3B model suggests overfitting, likely due to its higher parameter count and possibly insufficient data or suboptimal regularization.
- Best Model: The 1B model maintained a lower and more stable validation loss by the end of training, suggesting better generalization under current training conditions.
- Takeaway: Bigger isn't always better—especially with limited or noisy data. In this case, LLaMA-3.2-1B fine-tuned with LoRA generalizes better than the larger 3B model

Task 2.2: Analysis of Key Metrics

BLEU-1 & -2:

- Overall performance is poor. 1B starts for both at the highest value for both models and quickly declines to 0 or slightly above 0.
- The 3B model starts below the 1B model and alternates performance increase and decrease throughout the epochs with no clear trend.

METEOR:

- Both 1B and 3B models show poor overall performance on the METEOR metric, indicating challenges in producing semantically or lexically aligned captions.
- The performance trend of both models mirrors that of BLEU-1 and BLEU-2, with 1B starting higher but degrading over time, while 3B improves gradually.1B's early advantage disappears quickly, suggesting that it fails to retain improvements related to synonymy, word order, or recall-based matching—areas where METEOR is more sensitive.

ROUGE-L:

- Overall performance is poor, indicating difficulty in generating captions with strong lexical overlap and sequence alignment with reference texts.
- The 1B model follows a similar degradation trend as seen in BLEU and METEOR, starting decently but declining rapidly and stabilizing at a low point.3B starts slightly higher than 1B, but steadily declines over the training epochs, without any noticeable recovery or learning gains.
- Both models struggle with maintaining or improving the longest common subsequence quality, which ROUGE-L emphasizes.

Task 1.2: Qualitative analysis of Fine Tuning Llama

- GT: Parisian
- VIT-Llama 1B Fine-Tuned: "ovioviovioviovioviovioviovioviovioviovioviovioviovioviovioviovioviovioviovioviovioviovioviovio-vioviovioviovioviovioviovioviovioviovioviovioviovioviovioviovi"
- 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.

Team 8 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-vioviovioviovioviovioviovioviovioviovioviovioviovioviovioviovi"
- VIT-Llama 3B Fine-Tuned: "project and and the C, for the same &"