Reversing the Text-to-Image Relationship

Aesha Gandhi, Gaurav Law and Pranshul Bhatnagar

Duke PRATT SCHOOL OF Engineering

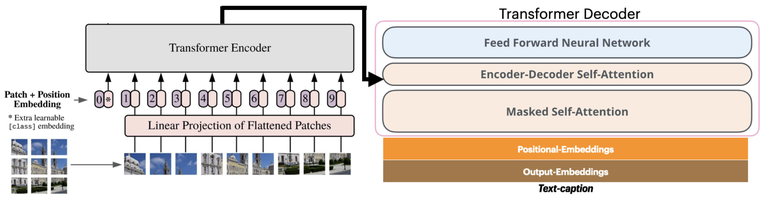
Introduction & Motivation

While text-to-image generation is widely explored, the reverse task, i.e., generating rich, context-aware text from images remains a significant challenge. This project investigates whether Vision-Language Pretraining (VLP) models can effectively invert this relationship to produce coherent narratives rather than simple labels.

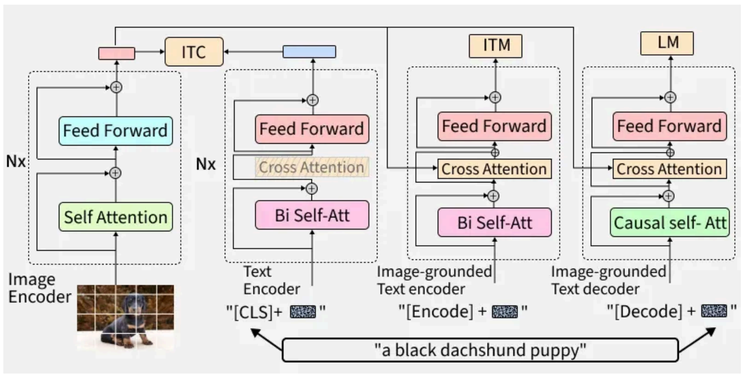
- **The Models:** We compare ViT (pure vision encoder) and BLIP (multimodal framework) to assess their ability to bridge visual embeddings with textual decoding.
- **The Method:** By leveraging transfer learning on transformer architectures, we adapt powerful pre-trained features to our specific dataset, maximizing efficiency and accuracy.
- **Evaluation**: Success is measured using BLEU and BERTScore to quantify the improvement in linguistic quality and semantic alignment.

Model Architecture

Visual Transformer Encoder (ViT) with GPT-2 Decoder



Bootstrapping Language-Image Pre-Training (BLIP) Model



Evaluation Metric

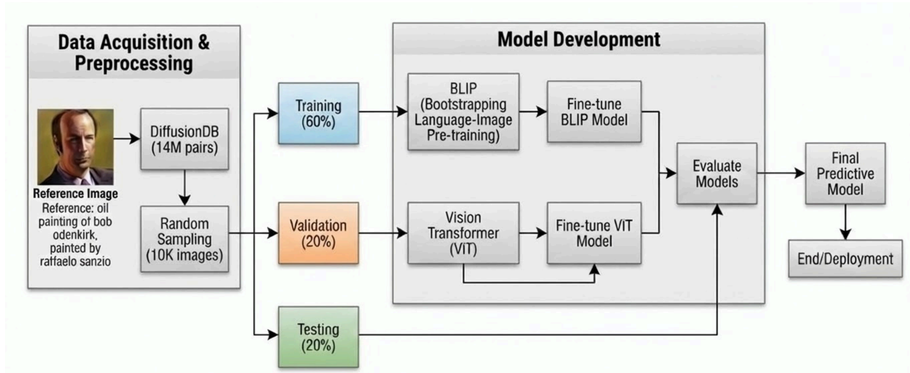
Combined Score =
0.65 × BERTScore
+ 0.35 × BLEU

65% 35%

BERTScore (65%)
Semantic alignment via contextual embeddings

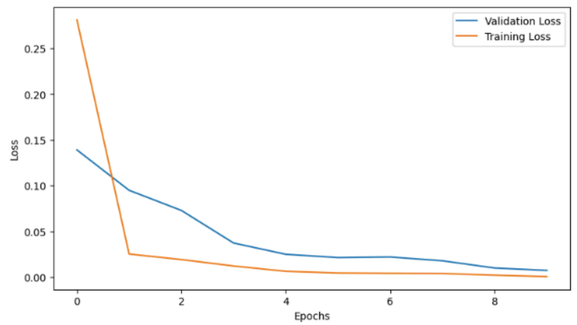
BLEU (35%)
Lexical precision via n-gram overlap

Methodology

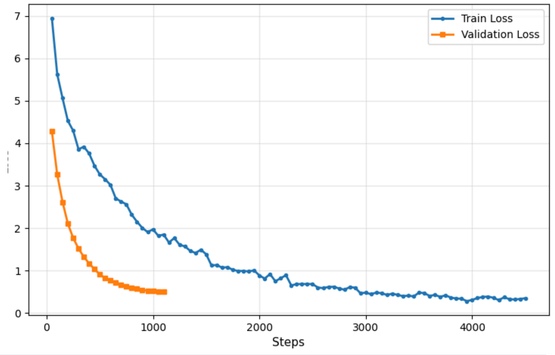


Training Vs Validation Loss Curves

BLIP Model



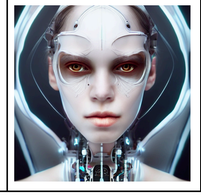
Visual Transformer Encoder (ViT) with GPT-2 Decoder



Results

Model	Metric	Pretrained	Fine-tuned	Absolute Change	Improvement
BLIP	BERTScore	0.4408	0.7632	0.3224	73.15%
	BLEU	0.271	0.7838	0.5128	189.20%
	Combined Score	0.3813	0.7704	0.3891	102%
ViT	BERTScore	0.4177	0.4285	0.0108	2.60%
	BLEU	0.0017	0.0045	0.0028	163.15%
	Combined Score	0.2721	0.2801	0.008	2.95%

Qualitative Results

	Image	Prompt	Caption	Score
ViT		a close up photo of a glass duck on a table, visible background, professional photography	close photo a duck a of glass on table visible, background professional	0.4814
ViT		A portrait of a beautiful anime girl wearing a long coat, shoulder length straight black hair	cute girl a anime wearing stylish coat shoulder straight, length bob hair straight hair	0.4333
BLIP		top-heavy 20 year old with messy black hair and big beard eats mayonnaise straight out of the jar with his bare hands	obese man with messy black hair and big beard drinking mayonnaise from the jar	0.8168
BLIP		A portrait of a beautiful female cyborg with a cracked porcelain face by wlop, exposed inner structure, brightly glowing white eyes	a portrait of a beautiful female cyborg wearing an intricate venetian mask by Wlop	0.7310

Key Findings

- BLIP shows substantial improvement through fine-tuning, with over 100% gain in combined score
- BLIP excels with clear subjects/distinct art styles, fine details/themes, hallucinates details in ambiguous images, often names people incorrect
- ViT shows positive learning direction but requires further refinement
- Model Comparison: BLIP outperforms ViT significantly, multimodal pre-training may provide stronger foundation for image-to-text generation than pure vision encoding
- Transfer learning significantly enhances caption quality

References

- Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.
- A. Kumar, “The Illustrated Image Captioning using transformers,” Ankur | NLP Enthusiast, Nov. 20,2022.https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transfo rmers/ (accessed Mar. 08, 2023)
- Hugging Face. (2022). Diffuse the rest - a hugging face space by Huggingface-Projects. Diffuse The Rest - a Hugging Face Space by huggingface-projects. Retrieved April 11, 2023