Character-Centric Story Visualization via Visual Planning and Token Alignment

Chen et al., 2022

Paper presentation for Natural Language Processing course by

Auriane Mahfouz and Elena Zoppellari

Università degli Studi di Padova

Introduction

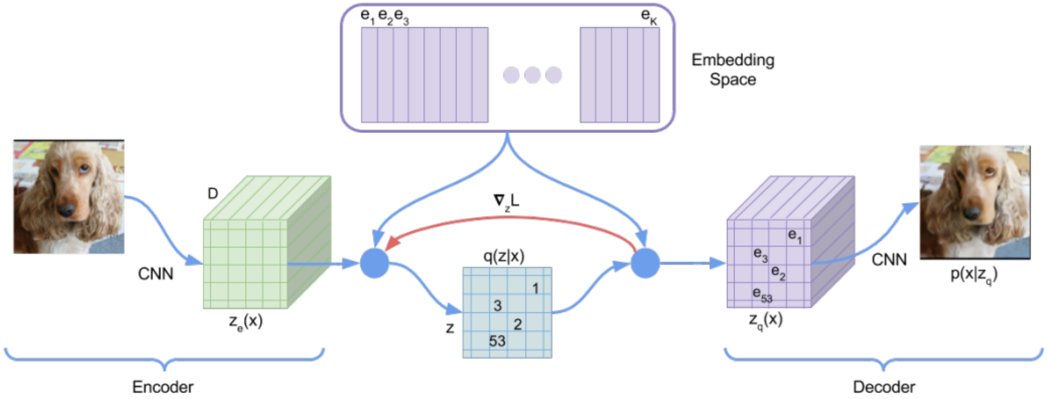
This paper proposes Character-Centric Story Visualization via Visual Planning (VP-CSV), a model that improves story visualization.

Previous works involve text-to-image generation and include:

- GAN-based (Generative Adversarial Nets) models
- DALL-E (Ramesh et al., 2021);
- VQ-VAE-LM that combines VQ-VAE (Van Den Oord et al., 2017) with a text-to-visual-token transformer (Ramesh et al., 2021; Yan et al., 2021)

Vector Quantised Variational AutoEncoder

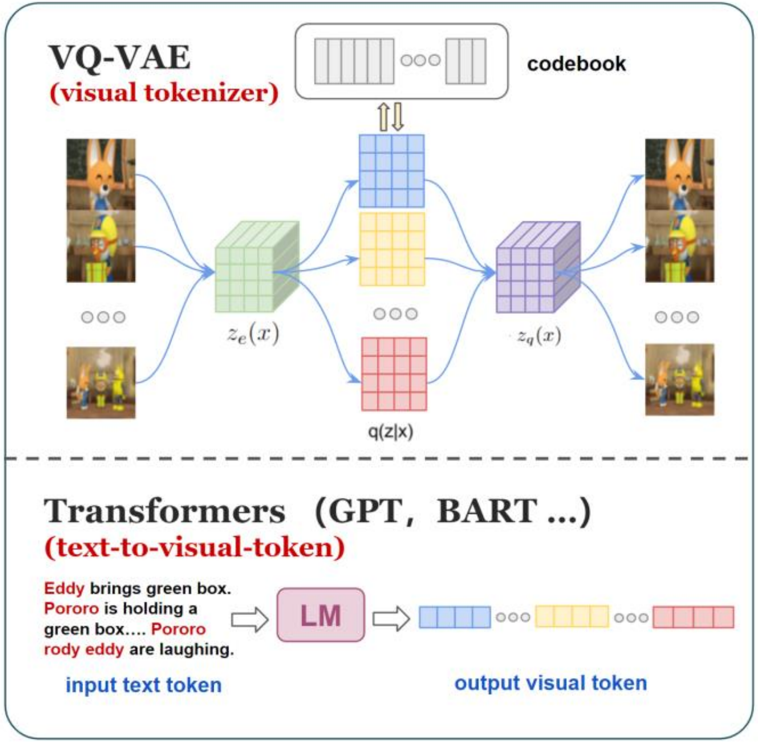
- VQ-VAE encodes images into discrete latent representations.
- Firstly, the images are encoded into continuous latent representations, then each of them is replaced with the closest embedding from the codebook.
- In this way the discrete representations are built and then fed into the decoder.



$$\mathcal{L} = \underbrace{\|x - z_q(x)\|_2^2}_{\mathcal{L}_{\text{recon}}} + \underbrace{\|sg[z_e(x)] - e\|_2^2}_{\mathcal{L}_{\text{codebook}}} + \underbrace{\beta\|sg[e] - z_e(x)\|_2^2}_{\mathcal{L}_{\text{commit}}}$$

1 image from Neural Discrete Representation Learning, van den Oord et al., 2018

VQ-VAE integrated with Language Model: VQ-VAE-LM

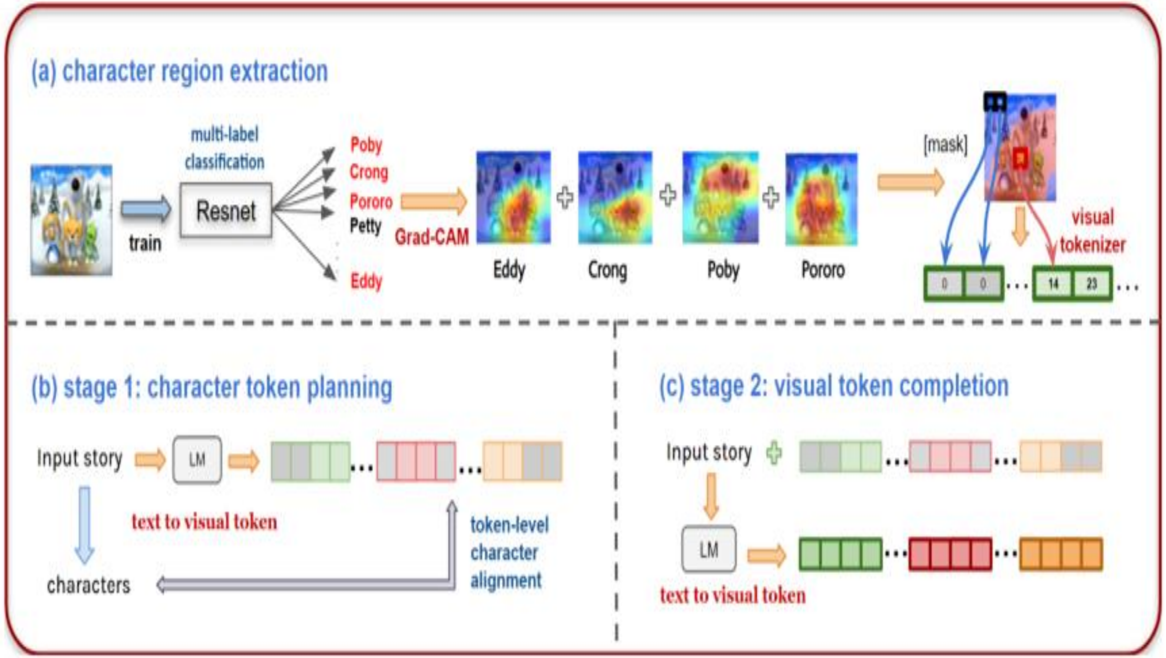


- As the first step, VQ-VAE is trained separately
- LM takes input text sentences
- LM is trained using MLE with the visual tokens from the VQ-VAE encoder as targets.
- Visual tokens from LM are fed into the VQ-VAE decoder.
- Images are reconstructed from the decoder.
- The two models are trained from scratch by the authors.

VP-CSV: Visual Planning based Character-centric Story Visualization

Università degli Studi di Padova

VP-CSV enhances VQ-VAE-LM with a two-stage module.



(a) Character region extraction:
Train a multi-label classifier to identify the character regions

(b) Plan module:
Train GPT-2 to generate the planned character token prepared by the previous stage with a training loss of: $L\theta = -\log p(r|s, \theta)$

(c) Completion module:
The model is trained to generate the background visual tokens z having a loss of $L\theta = -\log p(z|s, r, \theta)$

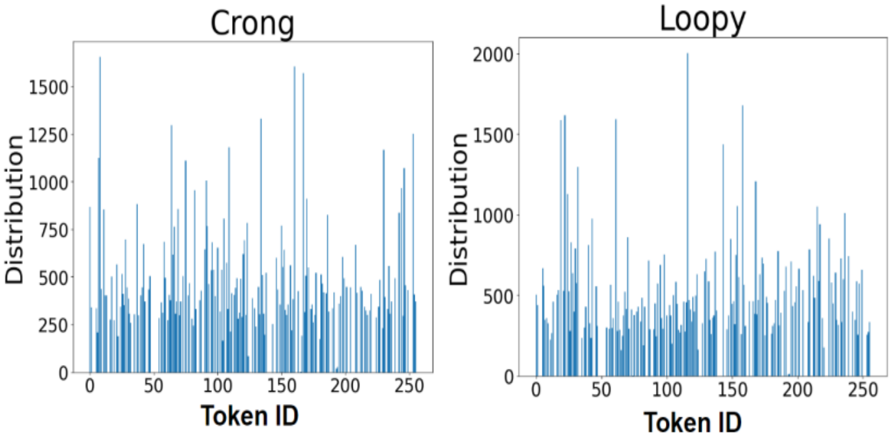
Token-level character alignment (TA)

The alignment process aims to match each character in the text with the corresponding visual token in the visual representation

- compute the visual token distribution for each character
- use a semantic loss to encourage the character-to-visual token alignment.
- calculate the semantic loss as:

$$\mathcal{L}^s(Q,\boldsymbol{p}) = -\log \sum_{\boldsymbol{z}\models Q} \prod_{\boldsymbol{z}^j\in P} p_j \prod_{\boldsymbol{z}^j\in N} (1-p_j)$$

The intuition of this objective is that if all characters' top visual tokens show up in the predicted images z (i.e. z |= Q), we increase the probability of tokens in P.



Experimental Setup

- The story-visualization dataset is Pororo-SV
- Each story is composed to 5 paragraphs
- Each paragraph is associated with an image.

Evaluation

Character preservation

- Character F1 score
- Frame Accuracy (Exact Match)

Image quality

- Frechet Inception Distance (FID)

Semantic alignment

- BLEU score
- R-precision

Human evaluation

- Visual Quality
- Character Preservation

Results (I)

Università degli Studi di Padova

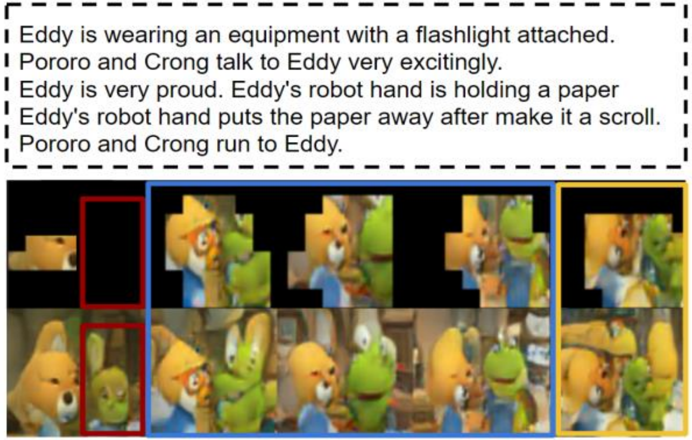
Method	Character F1	Frame Accuracy	FID↓	BLEU2/3	R-Precision
StoryGAN	18.59	9.34	158.06	3.24/1.22	1.51 ± 0.15
CP-CSV	21.78	10.03	149.29	3.25/1.22	1.76 ± 0.04
DUCO-StoryGAN	38.01	13.97	96.51	3.68/1.34	3.56 ± 0.04
VLC-StoryGAN	43.02	17.36	84.96	3.80/1.44	3.28 ± 0.00
VQ-VAE-LM	49.90	19.42	66.56	4.04/1.65	5.72± 0.02
+ Visual Planning	52.97	23.00	69.54	4.32/1.76	6.39 ± 0.00
+ Token Alignment	53.34	22.92	**63.34**	4.40/1.77	6.37 ± 0.00
VP-CSV	**56.84**	**25.87**	65.51	**4.45/1.80**	**6.95 ± 0.00**

- GAN-based models are outperformed by VQ-VAE-LM-based models

Metrics	VLC. vs VQ-VAE-LM		
	VLC.	VQ-VAE-LM	Tie
Visual	27.45	**62.75***	9.80
Character	37.25	**41.18**	21.57
	VQ-VAE-LM vs + TA		
	VQ-VAE-LM	+ TA	Tie
Visual	33.33	**42.10**	24.56
Character	38.59	**40.35**	21.06
	VQ-VAE-LM vs VP-CSV		
	VQ-VAE-LM	VP-CSV	Tie
Visual	34.51	**44.25**	21.24
Character	33.17	**52.21***	14.62

- Adding Token Alignment to VQ-VAE-LM produces better Visual Quality than VQ-VAE-LM alone according both automatic metric and human evaluation
- Overall VP-CSV model produces the best scores

Results (II)

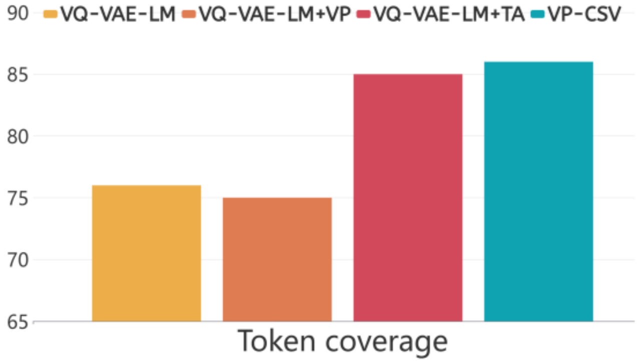


Analysis on character alignment

- Models with TA outperform the others

Analysis on visual planning

- Blue square represents correct character identification with background masked
- Red squares highlight errors of the completion module in generating non-existing characters
- Orange square underlines poor image quality



Conclusion

- In this paper, visual planning and character token alignment was proposed to improve character preservation and visual quality.
- Results show that the VP-CSV model outperforms all other models.
- Future research can aim at integrating actions and relationships among characters.

LIMITATIONS

- It is hard to generate every individual in the image.
- The image quality is still low.
- It is still hard to see the clear action performed by each character.

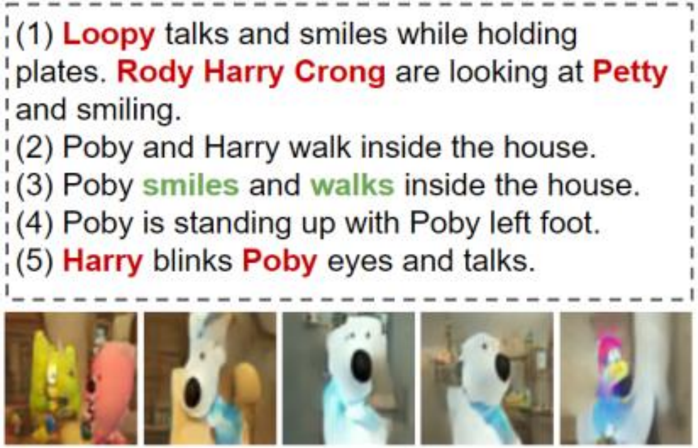


Figure 8: Limitation of generation image sequence.