A Neural Algorithm of Artistic Style

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge

Presenters:

Siba Smarak Panigrahi & Sohan Patnaik

Reading Session I
Kharagpur Data Analytics Group
IIT Kharagpur

February 14, 2021

Important Notes

- We expect you all to have certain queries regarding the presentation, certain intricate doubts maybe... Please put them in the **chat of this meeting**. We will address them all at the end!
- To ensure smooth flow of the presentation, we need you all to **keep your microphones and video turned off**.
- At the end, a **Google Form will be released** for feedback -BUT- most importantly, we ask you to put up name of **any AI-related paper** to be presented in the upcoming sessions!

Remember, if we can convert it into a presentation, we are indeed gonna present it!

Outline

Some Output Images!

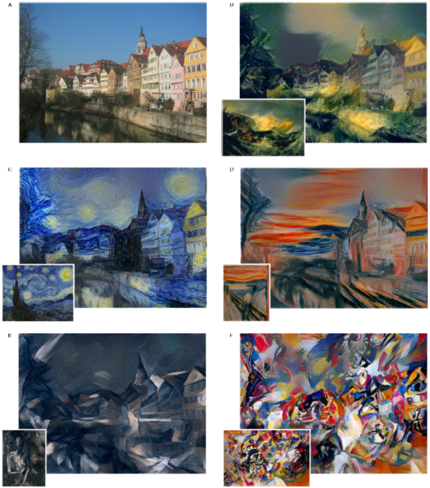

Figure: Have a look at this picture!

Images that combine the content of a photograph with the style of several well-known artworks. The images were created by finding an image that simultaneously matches the content representation of the photograph and the style representation of the artwork.

The original photograph depicting the **Neckarfront in Tubingen, Germany.**

Some Output Images!

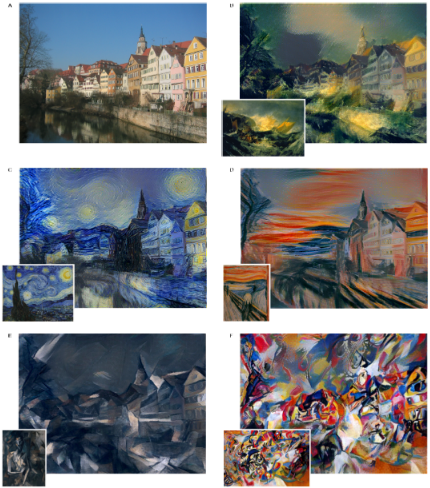

Figure: Have a look at this picture!

Images that combine the content of a photograph with the style of several well-known artworks. The images were created by finding an image that simultaneously matches the content representation of the photograph and the style representation of the artwork.

The original photograph depicting the **Neckarfront in Tubingen, Germany.**

Brief Introductory Ideas

Introduction

- Convolutional Neural Networks trained on Object Recognition develop representations of images that capture information explicitly along the processing hierarchy.
- Actual content of the image is captured as we go deeper in the network
- Content refers to the high level arrangement of objects, rather than the exact pixel values. In contrast to this, the initial layers in the processing hierarchy reflect the exact pixel values rather than high level content.
- And also in the initial layers the exact texture information is captured.

Makes sense, right!

Introduction

- Convolutional Neural Networks trained on Object Recognition develop representations of images that capture information explicitly along the processing hierarchy.
- Actual content of the image is captured as we go deeper in the network
- Content refers to the high level arrangement of objects, rather than the exact pixel values. In contrast to this, the initial layers in the processing hierarchy reflect the exact pixel values rather than high level content.
- And also in the initial layers the exact texture information is captured.

Makes sense, right!

Introduction

- Convolutional Neural Networks trained on Object Recognition develop representations of images that capture information explicitly along the processing hierarchy.
- Actual content of the image is captured as we go deeper in the network
- Content refers to the high level arrangement of objects, rather than the exact pixel values. In contrast to this, the initial layers in the processing hierarchy reflect the exact pixel values rather than high level content.
- And also in the initial layers the exact texture information is captured.

Makes sense, right!

Feature Spaces

- A **Feature** is an individual measurable property or characteristic of a phenomenon being observed. Quite intuitive!
- Well, what are the features here?
- We know that, in CNNs, the response that the image representations have to filters in a particular layer propagates forward so as to perform the desired tasks.
- These **"filter responses"** are nothing but constitute the feature spaces, more intuitively vector feature spaces.
- **Content Feature Space** carries information regarding the content of the image, whereas the **Style Feature Space** carries information with respect to the style of the image.

Feature Spaces

- A **Feature** is an individual measurable property or characteristic of a phenomenon being observed. Quite intuitive!
- Well, what are the features here?
- We know that, in CNNs, the response that the image representations have to filters in a particular layer propagates forward so as to perform the desired tasks.
- These **“filter responses“** are nothing but constitute the feature spaces, more intuitively vector feature spaces.
- **Content Feature Space** carries information regarding the content of the image, whereas the **Style Feature Space** carries information with respect to the style of the image.

Feature Spaces

- A **Feature** is an individual measurable property or characteristic of a phenomenon being observed. Quite intuitive!
- Well, what are the features here?
- We know that, in CNNs, the response that the image representations have to filters in a particular layer propagates forward so as to perform the desired tasks.
- These **“filter responses“** are nothing but constitute the feature spaces, more intuitively vector feature spaces.
- **Content Feature Space** carries information regarding the content of the image, whereas the **Style Feature Space** carries information with respect to the style of the image.

Content and Style Representation

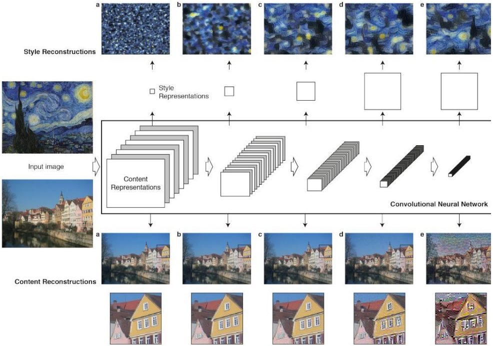

Figure: Content & Style Reconstruction

Content Representations: Feature responses in higher layers of the network.In a Convolutional Neural Network, the output of each given layer consists of "feature-maps". These feature maps also called "filter-responses", constitute the content feature space.

Content and Style Representation

Style Representations: Simply compute the correlations between different types of neurons in the network.

- Hey wait, what do you mean?
- Ok. See. We have the style feature space, right?
- Yes!
- We have the individual space feature maps. We find correlation between them. Simple.
- What is correlation between them?
- Ummm... Ok. See,
  - In case of vectors, we simply use cosine similarity or even simpler, dot product between them to comment how similar they are.
  - (Jargon Alert!) Here, we will use Gram Matrix for this purpose!
- **Note**: Have patience! We will cover how exactly Gram Matrix is evaluated when we talk about Style Loss!

Content and Style Representation

Style Representations: Simply compute the correlations between different types of neurons in the network.

- Hey wait, what do you mean?
- Ok. See. We have the style feature space, right?
- Yes!
- We have the individual space feature maps. We find correlation between them. Simple.
- What is correlation between them?
- Ummm... Ok. See,
  - In case of vectors, we simply use cosine similarity or even simpler, dot product between them to comment how similar they are.
  - (JARGON ALERT!) Here, we will use Gram Matrix for this purpose!
- **Note**: Have patience! We will cover how exactly Gram Matrix is evaluated when we talk about Style Loss!

Content and Style Representation

Style Representations: Simply compute the correlations between different types of neurons in the network.

- Hey wait, what do you mean?
- Ok. See. We have the style feature space, right?
- Yes!
- We have the individual space feature maps. We find correlation between them. Simple.
- What is correlation between them?
- Ummm... Ok. See,
  - In case of vectors, we simply use cosine similarity or even simpler, dot product between them to comment how similar they are.
  - (Jargon Alert!) Here, we will use Gram Matrix for this purpose!
- **Note**: Have patience! We will cover how exactly Gram Matrix is evaluated when we talk about Style Loss!

Content and Style Representation

Style Representations: Simply compute the correlations between different types of neurons in the network.

- Hey wait, what do you mean?
- Ok. See. We have the style feature space, right?
- Yes!
- We have the individual space feature maps. We find correlation between them. Simple.
- What is correlation between them?
- Ummm... Ok. See,
  - In case of vectors, we simply use cosine similarity or even simpler, dot product between them to comment how similar they are.
  - (JARGON ALERT!) Here, we will use Gram Matrix for this purpose!
- **Note**: Have patience! We will cover how exactly Gram Matrix is evaluated when we talk about Style Loss!

Let's mix style and content

A **key finding** in the paper was that representations of style and content in the Convolutional Neural Network are separable.

- It is obvious that we cannot completely disentangle the content and style of an image.
- Well, this is the reason we cannot reconstruct an image that matches the style of one and content of another image completely
- Intuitively, we can manipulate our learning model in way that gives priority to one over the other.
- Here comes the role of loss function. A strong emphasis on style will result in images that match the appearance of the artwork, effectively giving a texturised version of it whereas, a strong emphasis on content will make one identify the photograph clearly
- For a specific pair of source images one can adjust the trade-off between content and style to create visually appealing images.

Let's mix style and content

A **key finding** in the paper was that representations of style and content in the Convolutional Neural Network are separable.

- It is obvious that we cannot completely disentangle the content and style of an image.
- Well, this is the reason we cannot reconstruct an image that matches the style of one and content of another image completely
- Intuitively, we can manipulate our learning model in way that gives priority to one over the other.
- Here comes the role of loss function. A strong emphasis on style will result in images that match the appearance of the artwork, effectively giving a texturised version of it whereas, a strong emphasis on content will make one identify the photograph clearly
- For a specific pair of source images one can adjust the trade-off between content and style to create visually appealing images.

Let's mix style and content

A **key finding** in the paper was that representations of style and content in the Convolutional Neural Network are separable.

- It is obvious that we cannot completely disentangle the content and style of an image.
- Well, this is the reason we cannot reconstruct an image that matches the style of one and content of another image completely
- Intuitively, we can manipulate our learning model in way that gives priority to one over the other.
- Here comes the role of loss function. A strong emphasis on style will result in images that match the appearance of the artwork, effectively giving a texturised version of it whereas, a strong emphasis on content will make one identify the photograph clearly
- For a specific pair of source images one can adjust the trade-off between content and style to create visually appealing images.

Let's mix style and content

A **key finding** in the paper was that representations of style and content in the Convolutional Neural Network are separable.

- It is obvious that we cannot completely disentangle the content and style of an image.
- Well, this is the reason we cannot reconstruct an image that matches the style of one and content of another image completely
- Intuitively, we can manipulate our learning model in way that gives priority to one over the other.
- Here comes the role of loss function. A strong emphasis on style will result in images that match the appearance of the artwork, effectively giving a texturised version of it whereas, a strong emphasis on content will make one identify the photograph clearly
- For a specific pair of source images one can adjust the trade-off between content and style to create visually appealing images.

Let's mix style and content

A **key finding** in the paper was that representations of style and content in the Convolutional Neural Network are separable.

- It is obvious that we cannot completely disentangle the content and style of an image.
- Well, this is the reason we cannot reconstruct an image that matches the style of one and content of another image completely
- Intuitively, we can manipulate our learning model in way that gives priority to one over the other.
- Here comes the role of loss function. A strong emphasis on style will result in images that match the appearance of the artwork, effectively giving a texturised version of it whereas, a strong emphasis on content will make one identify the photograph clearly
- For a specific pair of source images one can adjust the trade-off between content and style to create visually appealing images.

Let's mix style and content

A **key finding** in the paper was that representations of style and content in the Convolutional Neural Network are separable.

- It is obvious that we cannot completely disentangle the content and style of an image.
- Well, this is the reason we cannot reconstruct an image that matches the style of one and content of another image completely
- Intuitively, we can manipulate our learning model in way that gives priority to one over the other.
- Here comes the role of loss function. A strong emphasis on style will result in images that match the appearance of the artwork, effectively giving a texturised version of it whereas, a strong emphasis on content will make one identify the photograph clearly
- For a specific pair of source images one can adjust the trade-off between content and style to create visually appealing images.

Figure: Content & Style Ratio Analysis

Loss Overview

Just a fact: The convolutional network architecture used is VGG-19, a network that rivals human performance on visual recognition benchmark task.

Before going to notations, you should be clear that the filter responses are flattened for further mathematical manipulations, i.e. responses of each filter is flattened and then stacked over one another to form a 2D matrix.

Assume that we have N_l distinct filters in layer l, which maps to features each of size M_l. M_l is height times the width of the feature map.Therefore the responses in layer l can be stored in a matrix $F^l \in \mathcal{R}^{N_l x M_l}$, where F^l_{ij} is the activation of i^{th} filter at position j in layer l.

In order to reconstruct image with amalgamation of style and content, we perform gradient descent on a white noise image that matches feature responses of the original image.

Loss Overview

Just a fact: The convolutional network architecture used is VGG-19, a network that rivals human performance on visual recognition benchmark task.

Before going to notations, you should be clear that the filter responses are flattened for further mathematical manipulations, i.e. responses of each filter is flattened and then stacked over one another to form a 2D matrix.

Assume that we have N_l distinct filters in layer l, which maps to features each of size M_l. M_l is height times the width of the feature map. Therefore the responses in layer l can be stored in a matrix $F^l \in \mathcal{R}^{N_l \times M_l}$, where F^l_{ij} is the activation of i^{th} filter at position j in layer l.

In order to reconstruct image with amalgamation of style and content, we perform gradient descent on a white noise image that matches feature responses of the original image.

Loss Overview

Just a fact: The convolutional network architecture used is VGG-19, a network that rivals human performance on visual recognition benchmark task.

Before going to notations, you should be clear that the filter responses are flattened for further mathematical manipulations, i.e. responses of each filter is flattened and then stacked over one another to form a 2D matrix.

Assume that we have N_l distinct filters in layer l, which maps to features each of size M_l. M_l is height times the width of the feature map.Therefore the responses in layer l can be stored in a matrix $F^l \in \mathcal{R}^{N_l x M_l}$, where F^l_{ij} is the activation of i^{th} filter at position j in layer l.

In order to reconstruct image with amalgamation of style and content, we perform gradient descent on a white noise image that matches feature responses of the original image.

Loss Overview

Just a fact: The convolutional network architecture used is VGG-19, a network that rivals human performance on visual recognition benchmark task.

Before going to notations, you should be clear that the filter responses are flattened for further mathematical manipulations, i.e. responses of each filter is flattened and then stacked over one another to form a 2D matrix.

Assume that we have N_l distinct filters in layer l, which maps to features each of size M_l. M_l is height times the width of the feature map.Therefore the responses in layer l can be stored in a matrix $F^l \in \mathcal{R}^{N_l x M_l}$, where F^l_{ij} is the activation of i^{th} filter at position j in layer l.

In order to reconstruct image with amalgamation of style and content, we perform gradient descent on a white noise image that matches feature responses of the original image.

Content Loss

We have input image $\overrightarrow{x}$, which is nothing but a random noise. Consider, $\overrightarrow{p}$ to be the original image. Let F^l and P^l be their respective feature representation in layer l. The squared-error loss between the two feature representation:

$$\mathcal{L}_{content}(\overrightarrow{p}, \overrightarrow{x}, l) = \frac{1}{2}\sum_{i,j}(F^l_{ij} - P^l_{ij})^2 \tag{1}$$

The derivative with respect to F^l_{ij}

$$\frac{\partial \mathcal{L}_{content}}{\partial F^l_{ij}} = \begin{cases} (F^l - P^l)_{ij} & \text{if} \quad F^l_{ij} > 0 \\ 0 & \text{if} \quad F^l_{ij} < 0 \end{cases} \tag{2}$$

Content Loss

We have input image $\overrightarrow{x}$, which is nothing but a random noise. Consider, $\overrightarrow{p}$ to be the original image. Let F^l and P^l be their respective feature representation in layer l. The squared-error loss between the two feature representation:

$$\mathcal{L}_{content}(\overrightarrow{p}, \overrightarrow{x}, l) = \frac{1}{2}\sum_{i,j}(F^l_{ij} - P^l_{ij})^2 \tag{1}$$

The derivative with respect to F^l_{ij}

$$\frac{\partial \mathcal{L}_{content}}{\partial F^l_{ij}} = \begin{cases} (F^l - P^l)_{ij} & \text{if} \quad F^l_{ij} > 0 \\ 0 & \text{if} \quad F^l_{ij} < 0 \end{cases} \tag{2}$$

Style Loss

Style representation is captured by the correlations of different filters, where the expectation is taken over the spatial extent of the input image.

We compute **Gram Matrix** which gives a representation of the correlation between different filters.

Let $\overrightarrow{a}$ be the original style image, A^l and G^l represent the gram matrices of the style image and the noise.

Gram Matrix, $G^l \in \mathcal{R}^{N_l x N_l}$, is the inner product of vectorised feature maps i and j in layer l.

Style Loss

Style representation is captured by the correlations of different filters, where the expectation is taken over the spatial extent of the input image.

We compute **Gram Matrix** which gives a representation of the correlation between different filters.

Let $\overrightarrow{a}$ be the original style image, A^l and G^l represent the gram matrices of the style image and the noise.

Gram Matrix, $G^l \in \mathcal{R}^{N_l x N_l}$, is the inner product of vectorised feature maps i and j in layer l.

Style Loss

Style representation is captured by the correlations of different filters, where the expectation is taken over the spatial extent of the input image.

We compute **Gram Matrix** which gives a representation of the correlation between different filters.

Let $\overrightarrow{a}$ be the original style image, A^l and G^l represent the gram matrices of the style image and the noise.

Gram Matrix, $G^l \in \mathcal{R}^{N_l x N_l}$, is the inner product of vectorised feature maps i and j in layer l.

Style Loss

Style representation is captured by the correlations of different filters, where the expectation is taken over the spatial extent of the input image.

We compute **Gram Matrix** which gives a representation of the correlation between different filters.

Let $\overrightarrow{a}$ be the original style image, A^l and G^l represent the gram matrices of the style image and the noise.

Gram Matrix, $G^l \in \mathcal{R}^{N_l x N_l}$, is the inner product of vectorised feature maps i and j in layer l.

Style Loss Equations

Each element of Gram Matrix G^l:

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk} \tag{3}$$

The contribution of the layer l to total style loss is

$$E_l = \frac{1}{4N_l^2 M_l^2} \cdot \sum_{i,j} (G^l_{ij} - A^l_{ij})^2 \tag{4}$$

The total style loss? Sum all E_l.

$$\mathcal{L}_{style}(\overrightarrow{a}, \overrightarrow{x}) = \sum_{l=0}^{L} w_l E_l \tag{5}$$

Derivative of E_l as per 4

$$\frac{\partial E_l}{\partial F^l_{ij}} = \begin{cases} \frac{1}{N_l^2 M_l^2}((F^l)^T(G^l - A^l))_{ji} & \text{if} \quad F^l_{ij} > 0 \\ 0 & \text{if} \quad F^l_{ij} < 0 \end{cases} \tag{6}$$

Style Loss Equations

Each element of Gram Matrix G^l:

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l \tag{3}$$

The contribution of the layer l to total style loss is

$$E_l = \frac{1}{4N_l^2 M_l^2} . \sum_{i,j} (G_{ij}^l - A_{ij}^l)^2 \tag{4}$$

The total style loss? Sum all E_l.

$$\mathcal{L}_{style}(\overrightarrow{a}, \overrightarrow{x}) = \sum_{l=0}^{L} w_l E_l \tag{5}$$

Derivative of E_l as per 4

$$\frac{\partial E_l}{\partial F_{ij}^l} = \begin{cases} \frac{1}{N_l^2 M_l^2}((F^l)^T(G^l - A^l))_{ji} & \text{if} \quad F_{ij}^l > 0 \\ 0 & \text{if} \quad F_{ij}^l < 0 \end{cases} \tag{6}$$

Style Loss Equations

Each element of Gram Matrix G^l:

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l \tag{3}$$

The contribution of the layer l to total style loss is

$$E_l = \frac{1}{4N_l^2 M_l^2} . \sum_{i,j} (G_{ij}^l - A_{ij}^l)^2 \tag{4}$$

The total style loss? Sum all E_l.

$$\mathcal{L}_{style}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l \tag{5}$$

Derivative of E_l as per 4

$$\frac{\partial E_l}{\partial F_{ij}^l} = \begin{cases} \frac{1}{N_l^2 M_l^2}((F^l)^T (G^l - A^l))_{ji} & \text{if} \quad F_{ij}^l > 0 \\ 0 & \text{if} \quad F_{ij}^l < 0 \end{cases} \tag{6}$$

Style Loss Equations

Each element of Gram Matrix G^l:

$$G_{ij}^l = \sum\nolimits_k F_{ik}^l F_{jk}^l \tag{3}$$

The contribution of the layer l to total style loss is

$$E_l = \frac{1}{4N_l^2 M_l^2} . \sum\nolimits_{i,j} (G_{ij}^l - A_{ij}^l)^2 \tag{4}$$

The total style loss? Sum all E_l.

$$\mathcal{L}_{style}(\vec{a}, \vec{x}) = \sum\nolimits_{l=0}^{L} w_l E_l \tag{5}$$

Derivative of E_l as per 4

$$\frac{\partial E_l}{\partial F_{ij}^l} = \begin{cases} \frac{1}{N_l^2 M_l^2}((F^l)^T (G^l - A^l))_{ji} & \text{if} \quad F_{ij}^l > 0 \\ 0 & \text{if} \quad F_{ij}^l < 0 \end{cases} \tag{6}$$

Put Both Losses Together

Note: In eq (5), w_l are weighting factors of the contribution of each layer to the total style loss. In the paper, they have assumed each to be $\frac{1}{5}$ for each of the active layers and 0 for rest of the layers.

The total loss function we minimise is

$$\mathcal{L}_{total}(\overrightarrow{p}, \overrightarrow{a}, \overrightarrow{x}) = \alpha\mathcal{L}_{content}(\overrightarrow{p}, \overrightarrow{x}) + \beta\mathcal{L}_{style}(\overrightarrow{a}, \overrightarrow{x}) \tag{7}$$

The variables α and β are nothing but weights given to the content and style loss respectively. More the value of one, more emphasis is given to the corresponding loss, and that part is more pronounced in the final output image ($\overrightarrow{x}$). This resulted in the 20 different output images we saw earlier depending on the ratio of α and β

Put Both Losses Together

Note: In eq (5), w_l are weighting factors of the contribution of each layer to the total style loss. In the paper, they have assumed each to be $\frac{1}{5}$ for each of the active layers and 0 for rest of the layers.

The total loss function we minimise is

$$\mathcal{L}_{total}(\overrightarrow{p}, \overrightarrow{a}, \overrightarrow{x}) = \alpha\mathcal{L}_{content}(\overrightarrow{p}, \overrightarrow{x}) + \beta\mathcal{L}_{style}(\overrightarrow{a}, \overrightarrow{x}) \tag{7}$$

The variables α and β are nothing but weights given to the content and style loss respectively. More the value of one, more emphasis is given to the corresponding loss, and that part is more pronounced in the final output image ($\overrightarrow{x}$). This resulted in the 20 different output images we saw earlier depending on the ratio of α and β

Put Both Losses Together

Note: In eq (5), w_l are weighting factors of the contribution of each layer to the total style loss. In the paper, they have assumed each to be $\frac{1}{5}$ for each of the active layers and 0 for rest of the layers.

The total loss function we minimise is

$$\mathcal{L}_{total}(\overrightarrow{p}, \overrightarrow{a}, \overrightarrow{x}) = \alpha\mathcal{L}_{content}(\overrightarrow{p}, \overrightarrow{x}) + \beta\mathcal{L}_{style}(\overrightarrow{a}, \overrightarrow{x}) \tag{7}$$

The variables α and β are nothing but weights given to the content and style loss respectively. More the value of one, more emphasis is given to the corresponding loss, and that part is more pronounced in the final output image ($\overrightarrow{x}$). This resulted in the 20 different output images we saw earlier depending on the ratio of α and β

Conclusion

- This work offers how neural representation can independently capture the content of an image and the style in which it is presented.
- It also proposes an architecture on Deep Neural Network that can create artistic images of high perceptual quality
- Hence it intuitively provides a way to ponder about how humans create and perceive artistic imagery.

Conclusion

- This work offers how neural representation can independently capture the content of an image and the style in which it is presented.
- It also proposes an architecture on Deep Neural Network that can create artistic images of high perceptual quality
- Hence it intuitively provides a way to ponder about how humans create and perceive artistic imagery.

Conclusion

- This work offers how neural representation can independently capture the content of an image and the style in which it is presented.
- It also proposes an architecture on Deep Neural Network that can create artistic images of high perceptual quality
- Hence it intuitively provides a way to ponder about how humans create and perceive artistic imagery.

Thank You!

Yo! That's All For Valentine's Day :)

Thank You!
Yo! That's All For Valentine's Day :)