Motivation
- Image style transfer is to transfer the style of a style image onto a content image.
- A consistent observation from existing work is that VGG is the best architecture as default feature extractor.
- We aim to 1) explore why VGG performs better and 2) a solution to mitigate the problem of other non-VGG architectures.

Importance of residual connections
- We experiment on pre-trained models (p-) and random-weight ones (r-) and find the quality varies drastically.
- Peaky maximum and small entropy:
  - Outlier sensitivity of L2, partially emphasize 'peaky' positions, overfit on a few style patterns and ignore the remaining.

Ablation study
- We perform an ablation study over many network components and find the poor performance of ResNet is mainly caused by its residual connections.

Why do residual connections degrade performance?
- Peaky maximum and small entropy:
  - Hard to suppress peaky values due to residual connections.

Evaluation
- On different non-VGGs
- On different methods

Why are residual network activations and Gram matrices peaky?
- Hard to suppress peaky values due to residual connections.

Stylization With Activation smoothing (SWAG)?
- Smooth the activation:
  \[ \mathcal{L}_{swag}(x) = \frac{1}{2} \left( \frac{1}{\gamma} \left\| F(x) - F^c(x) \right\|_2^2 - \gamma \mathcal{L}_{cons}(x) \right) \]

Preliminaries
- Given a content and a style image, and a fixed feature extractor, the result is obtained by:
  \[ x^* = \arg \min \alpha \mathcal{L}_{cons}(x^*) + \beta \mathcal{L}_{style}(x^*) \]
  with:
  \[ \mathcal{L}_{cons}(x^*) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \| F^c(x^*_i, j) - F(x^*_i, j) \|_2^2 \]
  \[ \mathcal{L}_{style}(x^*) = \sum_{i=1}^{H} \sum_{j=1}^{W} \| G^c(F^c(x^*_i, j)) - G(F(x^*_i, j)) \|_2^2 \]

Reference