딥러닝 올인원

설명 가능한 AI

26강

딥러닝호형

설명 가능한 AI(eXplainable AI - XAI)

XAI 중요성

- 결과 분석 용이
- 시각적 이해도 향상
- 모델 내부에 대한 이해도 향상

설명 가능한 AI(eXplainable AI - XAI)

이미지 분야의 XAI

- Class Activation Map(CAM)
- Attention
- Activation Maxmization

설명 가능한 AI(eXplainable AI - XAI)

Class Activation Map(CAM)

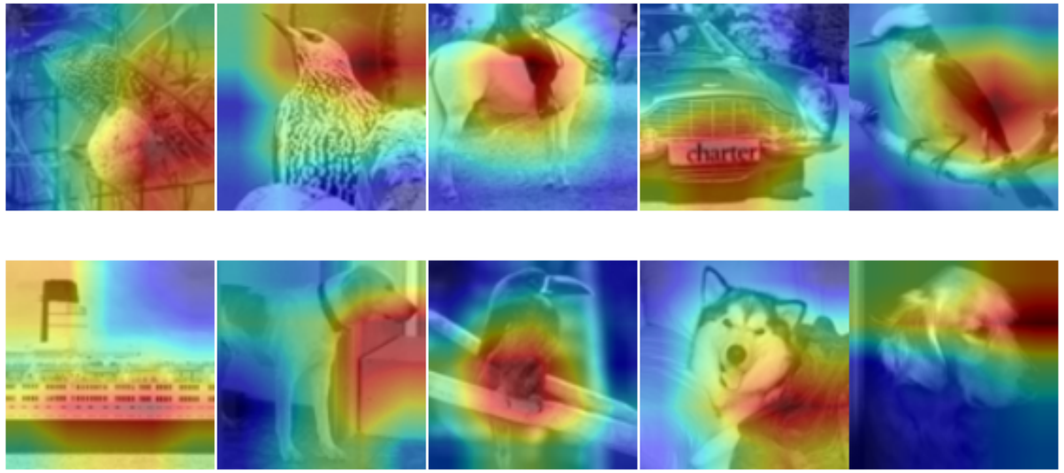
- CAM
- Grad-CAM
- Guided Grad-CAM

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba,
Learning Deep Features for Discriminative Localization, CVPR, 2016

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra,
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, CVPR, 2017

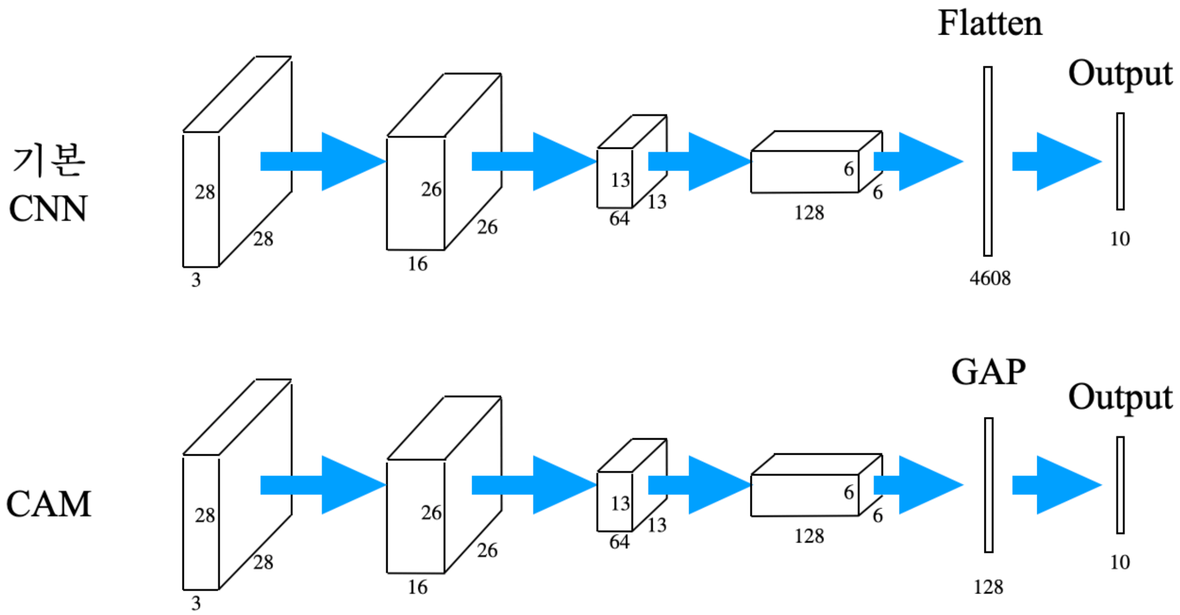
설명 가능한 AI(eXplainable AI - XAI)

Class Activation Map(CAM)

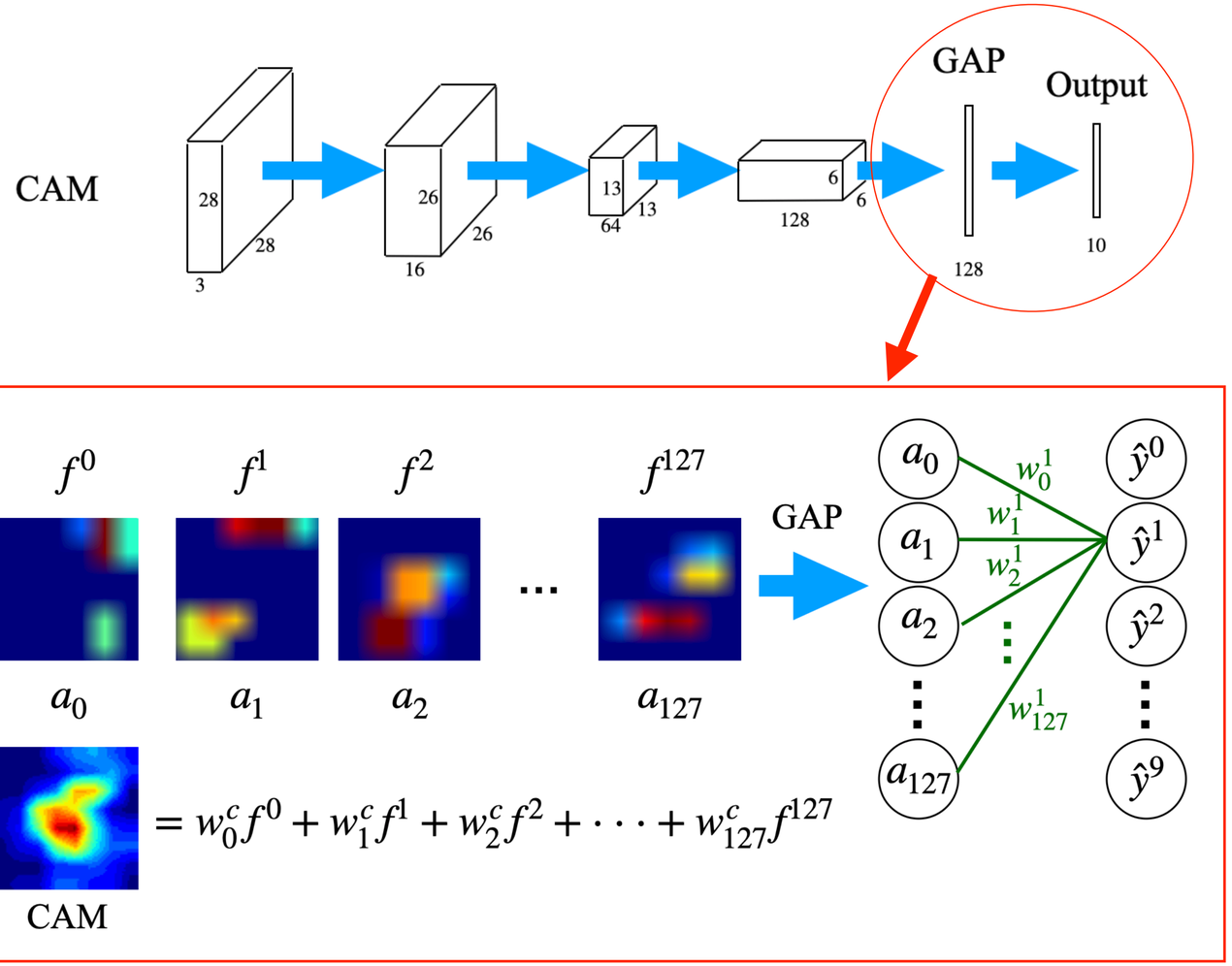


설명 가능한 AI(eXplainable AI - XAI)

Class Activation Map(CAM)

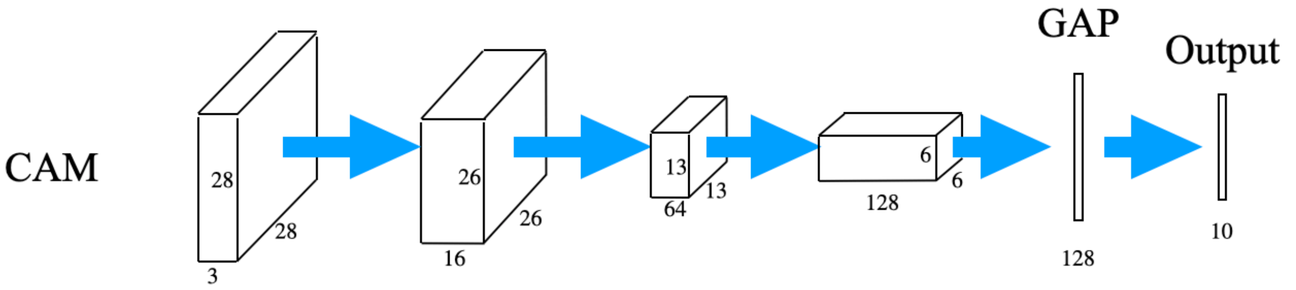


설명 가능한 AI(eXplainable AI - XAI)



설명 가능한 AI(eXplainable AI - XAI)

CAM 수식



$$\hat{y}^c = \sum_k w_k^c \frac{1}{N} \sum_i \sum_j f_{ij}^k = \frac{1}{N} \sum_i \sum_j \sum_k w_k^c f_{ij}^k = \frac{1}{N} \sum_i \sum_j \sum_k w_k^c f^k$$

CAM

설명 가능한 AI(eXplainable AI - XAI)

CAM의 한계

- FC를 GAP로 변경함으로써 모델 구조의 제약을 가짐
- 성능 하락의 가능성
- 마지막 피쳐맵에 대해서만 해석이 가능

설명 가능한 AI(eXplainable AI - XAI)

Grad-CAM의 등장

- 기존 FC를 그대로 사용 가능

설명 가능한 AI(eXplainable AI - XAI)

CAM과 Grad-CAM

$$CAM = \sum_k w_k^c f^k \qquad GradCAM = ReLU(\sum_k \alpha_k^c f^k) \quad (\alpha_k^c = \frac{1}{N}\sum_i \sum_j \frac{\partial \hat{y}^c}{\partial f_{ij}^k})$$

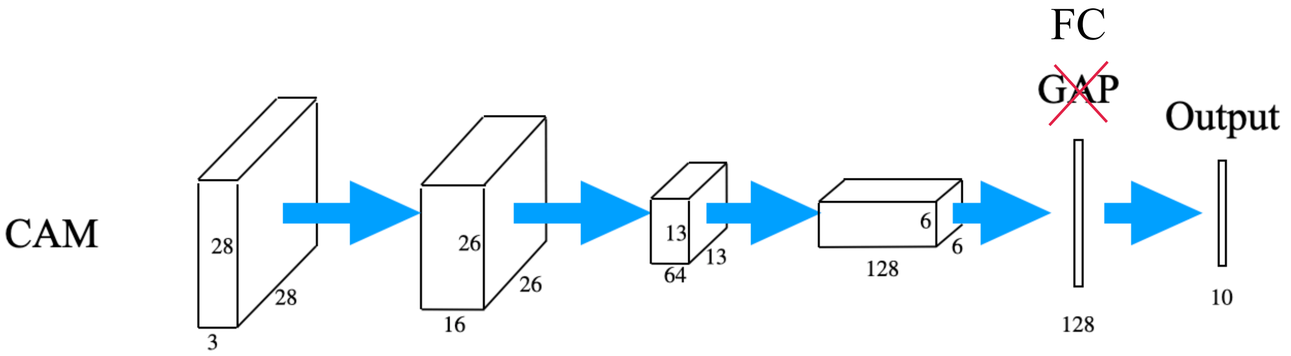
$$\hat{y}^c = \sum_k w_k^c \frac{1}{N} \sum_i \sum_j f_{ij}^k = \sum_k w_k^c a_k \quad (a_k = \frac{1}{N}\sum_i \sum_j f_{ij}^k)$$

$$\frac{\partial \hat{y}^c}{\partial a_k} = \frac{\frac{\partial \hat{y}^c}{\partial f_{ij}^k}}{\frac{\partial a_k}{\partial f_{ij}^k}} = N\frac{\partial \hat{y}^c}{\partial f_{ij}^k} \rightarrow w_k^c = N\frac{\partial \hat{y}^c}{\partial f_{ij}^k} \rightarrow \sum_i \sum_j w_k^c = N\sum_i \sum_j \frac{\partial \hat{y}^c}{\partial f_{ij}^k}$$

$$\rightarrow Nw_k^c = N\sum_i \sum_j \frac{\partial \hat{y}^c}{\partial f_{ij}^k} \rightarrow w_k^c = \sum_i \sum_j \frac{\partial \hat{y}^c}{\partial f_{ij}^k}$$

설명 가능한 AI(eXplainable AI - XAI)

CAM과 Grad-CAM



$$GradCAM = ReLU(\sum_k \alpha_k^c f^k) \quad (\alpha_k^c = \frac{1}{N}\sum_i \sum_j \frac{\partial \hat{y}^c}{\partial f_{ij}^k})$$